From electrons to proteins to data

FROM ELECTRONS TO PROTEINS TO DATA

HOW TO LOCALISE, OBSERVE AND ORGANISE THEM

Layout Typeset by the author with LaTeX
Cover Adaptation of a painting by Nolan van Schayck
ISBN 978-94-6496-023-5
Printing Gildeprint, the Netherlands

Financial support for the printing of this thesis was given by the Dutch Society for Microscopy (Nederlandse Vereniging voor Microscopie, NVvM, microscopy.nl)

FROM ELECTRONS TO PROTEINS TO DATA

HOW TO LOCALISE, OBSERVE AND ORGANISE THEM

DISSERTATION

To obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus, Prof. dr. P. Habibović,
in accordance with the decision of the Board of Deans,
to be defended in public
on Thursday 7th of March 2024, at 13.00 hours

by

Johannes Paul van Schayck

Promotor
Prof. Dr. Raimond B. G. Ravelli [1]
Prof. Dr. Peter J. Peters

Copromotor
Dr. Carmen Lopez-Iglesias

Assessment Committee
Prof. Dr. Ilja C. W. Arts (chair)
Dr. Arjen Jakobi, Delft University of Technology
Prof. Dr. Michel Dumontier
Prof. Dr. Ir. Frank J. W. Verhaegen
Dr. Katy Wolstencroft, Leiden University

This research was part of the M4i research program and received financial support from the Dutch Research Council (NWO) within the framework of the Fund New Chemical Innovationsm project MOL3DEM (project number 731.014.248); from the Netherlands Electron Microscopy Infrastructure (NEMI), project number 184.034.014 of the National Roadmap for Large-Scale Research Infrastructure of the Dutch Research Council (NWO); the PPP Allowance made available by Health-Holland, Top Sector Life Sciences & Health, to stimulate public–private partnerships, project 4DEM, number LSHM21029 and the LINK program from the Province of Limburg, the Netherlands.

[1] 25 March 1968–30 June 2023

De elektronenmicroscopist

Een stroom van negatieve golven
schiet als hagel door de loop
om het ongekende doel te raken
de steekvlam van een leger draken
martelt de materie
en verdampt de tijd tot lege uren
in de jacht naar het onbekende

Dan gebeurt het langverwachte
uit de golven spoelt iets aan
op de oevers van de wetenschap
de jutter blij verrast
houdt de schelp aan zijn oor
en ergens in de verte
klinkt het geruis van iets heel nieuws

Het nieuwe resultaat
smeult en fonkelt haar gelaat
laat de vlinders stiekem vrij
in de hemel van haar buik
ze glimt onzichtbaar
in de leegheid van het lab
vanavond zet zij
een minuscule stap

Ze zucht, als was ze zelf het leger draken
sjokt dan de schemergangen door
aan het eind treft zij een blik
die ook grijze beelden bewaakt
op zoek naar een stip
een beweging
een sprankje poëzie
die deze avond anders maakt
want dan moet hij zich spoeden
door dezelfde schemergangen
van de ongewisse wetenschap

Maar nee, hij draait alleen zijn blik
Kijk, sprak hij, dáár is
het onverwachte onbekende

Pauline van Schayck
Bart van Haastrecht

Contents

Preamble

Life sciences, or biology, is the study of life, such as microorganisms, plants and animals, including human beings. It is one of the two major branches of natural sciences, the other branch being physical science, the study of non-living matter.

In this thesis, I will look at life sciences from two perspectives. The first being structural biology: the study of how the structure and dynamics of biological macromolecules, such as **proteins**, determine and affect their function. There are several techniques for **observing** the structures and dynamics of these macromolecules. One of them is electron microscopy. To record an image of a sample, a transmission electron microscope projects the **electrons** that have passed through the sample onto a recording device, a detector. The challenge for the detector is to faithfully record and **localise** the incoming electrons and form the best possible image from the signal generated by the microscope. In this thesis, I will specifically look at the characterisation, integration and optimisation of such a detector for electron microscopy.

The second perspective I will look at is the management and description of **data** being generated in life sciences research and the **organisation** thereof. The amount of data, complexity, and multimodality of research data being generated has been ever increasing, while the urgency for openly and reusable publishing of said data has become greater as well.

These perspectives are wildly different from each other. However, they are both aspects of the large and complex task of understanding how life functions. I was given the generous opportunity to explore both in the same thesis. But let me first introduce these perspectives separately.

1
Introduction

1.1 Digital detectors for cryo-electron microscopy

1.1.1 The use of cryo-transmission electron microscopy in biology

Transmission electron microscopes (TEM) use electrons to form an image of the sample being observed. Electrons are first accelerated within a high-vacuum column before being shaped into a beam by a number of condensing lenses. This beam of electrons interacts with the sample causing the electrons to scatter. The objective lens forms an image of the scattered and unscattered electrons, which subsequently passes through diffraction and projection lenses. The magnified image can then be recorded onto a recording device, a detector. These basic principles of the electron microscope have not changed since its first description in 1932 by their inventors Max Knoll and Ernst Ruska [1, 2].

It was quickly realised that electron microscopes had great potential to study biological ultrastructure. Physician and biologist Helmut Ruska (Ernst brother) and Bodo von Borries studied bacteria and viruses in the late 1930s using the electron microscope [3, 4]. Compared to the photons used in light microscopy, the wavelength (defined by De Broglie [5]) we can assign to electrons is much smaller. This means that electron microscopes can fundamentally resolve much smaller objects than light microscopes. However, to be able to resolve such small objects, two other factors come into play when studying bio-samples. One is the amount of scattering the electrons will undergo when traversing the sample. For biological samples which mostly consist of carbon, hydrogen, oxygen and nitrogen this is relatively low. The second factor is the radiation damage that the electrons will inflict while passing the sample [6]. Combined, these two factors limit the amount of contrast in the images of bio-samples making it difficult to resolve them. The factors also interplay; the low amount of scattering cannot be countered by using more electrons as this would irreversibly destroy the radiation sensitive bio-sample.

These fundamental limits to electron microscopy on bio-samples were realised early on [7] and have been well described later [8]. In recent years, cryo-electron microscopy (cryo-EM) single-particle analysis (SPA) has become a viable option. SPA is a technique in which isolated and purified macromolecular structures are trapped in a thin vitreous ice layer and transferred into the vacuum of an electron microscope. In addition to fixating the bio-sample in its native aqueous condition, cryo conditions also reduce the radiation damage done to macromolecules [6, 9]. In SPA, hundreds, or even tens of thousands of images are taken from these particles. These so-called micrographs contain images of individual particles and, ideally, each particle would have a different orientation. From these 2D images of the particle in different orientations, it is possible to reconstruct a 3D electrostatic potential map of the average of that particle. Because the quality of the micrograph

depends on the recording of the small number of incident electrons that can be used to form the image, the recording device of the microscope plays an important role. The history of the development of a digital recording device (a detector) capable of supporting the SPA workflow has been a long road that we will briefly explore.

1.1.2 A brief history of recording devices for transmission electron microscopy

Knoll and Ruska used a fluorescent screen to visualise their first images in the early 1930s. Such a screen works by making use of the property of fluorescent atoms that enter an excited state when hit by an electron and emit this energy in the form of visible light. The same principle was used for the display screens of early oscilloscopes and later in consumer cathode ray tube (CRT) televisions. Knoll and Ruska initially used a photographic apparatus placed outside the microscope to record the light coming from their fluorescent glass plate [1]. In 1935, during an early phase of rapid progress, Ladislaus Marton developed an electron microscope with an air lock to directly introduce photographic plates into the vacuum chamber of the microscope [10]. This was a great improvement to the efficiency of electron detection and the practicality of the microscope and led to the first commercial microscope available in 1939 [11].

The introduction of the computer, after the Second World War, meant it became possible to perform a large number of automated calculations. Together with the invention of the Fast Fourier Transform algorithm [12], a large number of digital image processing algorithms became feasible to execute now. This led to the first 3D helical reconstruction from a series of micrographs of the tail of bacteriophage T4 [13]. At this time, photographic negatives must first be developed and digitised before the images are available on the computer. Digitisation was still done using a microdensitometer and each photographic plate or negative had to be manually scanned, which was a tedious process. The use of photographic plates and film also had a number of fundamental limitations in image quality. For example, the effect of the non-linear response of photographic material necessitated complicated calibration and linearisation techniques [14]. These aspects of image quality and the inconvenient and time-consuming practical aspects meant that a digital recording device was needed.

Charge-coupled devices (CCDs) were an invention of the Bell Lab in 1970 [15]. CCDs are made up of rows and columns of semiconductor material that form the pixels. CCD pixels consist of metal oxide-semiconductor (MOS) capacitors which convert the incoming photons into electrical charge, in the form of hole pairs. CCD pixels are connected ('coupled') and, by shifting charge from column

to column, individual pixels can be read out. This forms the digital image. The practical potential of CCDs was quickly realised for (consumer-grade) video cameras and digital still photography, but they also made their way into many scientific applications. One of the first fields to greatly benefit from this was space astronomy, where it became possible for the first time to transfer high quality images wirelessly back to Earth. For scientific applications often a so-called slow-scan version of the CCD sensor was used, which cannot record with continuous exposure, but benefits from much higher efficiency pixel-to-pixel charge rates and lower read-out noise [14].

The first application of a slow-scan CCD in electron microscopy was by Mochel and Mochel in the 1980s [16]. The main challenge of using CCDs for electron microscopy is that a primary incident electron will generate too much charge and the pixel quickly becomes saturated. To overcome this, in the late 1980s Hans Tietz, Paul Mooney and others started to build CCDs optically coupled through a fibre-optical taper to a scintillator layer [17]. The scintillator layer undergoes the same process as the fluorescence screens where atoms in the layer become excited and discharge this energy in the form of light. This reduced the amount of signal received by the CCD and made them viable for use in electron microscopes.

The digital CCD detectors allowed researchers to conveniently apply digital correction methods on the images and quantitatively analyse the results. In the mid-1990s, shortly after the introduction of CCDs for EM, Richard Henderson made predictions of the fundamental limits of cryo-EM [8]. But for a long time, cryo-EM was still in the era of 'blobology'. The disadvantage of CCD is that the scintillator smears the beam across several pixels and has a nonlinear response to the incoming amount of electrons, adding to the noise of the image and thus lowering the signal-to-noise ratio. This is caused by dynamic scattering in both the scintillator layer and inside the fibres used to optically couple the sensor to the scintillator layer. So, while the convenience of digital detectors with CCDs was there, there was still a strong need for improvements.

This advance came in the early 2000s, in the form of the monolithic active-pixel sensor (MAPS) detectors. The active pixel sensor, manufactured using the complementary metal-oxide-semiconductor (CMOS) process, had already gained much popularity in consumer grade electronics, mainly due to its lower cost production process compared to CCDs. For space astronomy applications, active pixels sensors were manufactured as a monolithic layer, reducing the amount of 'dead space' where no detection could take place [18]. Crucial for low-light astronomy applications, but, as would turn out, also for cryo-EM. Furthermore, in an active-pixel sensor, part of the analogue-to-digital logic is integrated directly into the pixel. This additional logic allows the detector to accommodate for the much higher direct charge deposition of the typical primary incident electrons (100 keV to 300 keV) used in cryo-EM. This direct detection of electrons, compared

to the indirect scintillator layer of the CCD, led to the MAPS detector being dubbed direct electron detectors.

The first testing of such a direct electron MAPS detector, originally intended for space astronomy, was performed by Wasi Faruqi and colleagues at the MRC Laboratory of Molecular Biology (LMB) in Cambridge [19]. They also showed that a disadvantage of the MAPS detector is that the additional transistors must withstand the radiation damage caused by the incident electrons. Over the next few years, radiation hard MAPS detectors were developed, leading to the first commercially available direct electron detectors [20].

Considerable improvement of the MAPS detectors was achieved by back-thinning the sensor to, eventually, less than 30 μm thickness. The effect of back-thinning is the reduced chance of an electron scattering back into the sensor layer and depositing its energy in pixels adjacent to the point of impact. This inter-pixel charge sharing will degrade the performance of the point spread function of a detector [21]. Combined with back-thinning, a further enhancement was the introduction of the electron counting mode, in which every electron is localised to a single pixel and given an equal weight. Various algorithms for event-localisation have been published [22, 23]. For the current range of MAPS detectors, these algorithms require a sufficiently low electron flux to prevent the signal from adjacent incident electrons from overlapping with each other within one frame. Details of the specific algorithms used by commercial MAPS detectors have not been released, but accurate event-localisation is still seen as a bottleneck [24]. A final later enhancement was the ability of MAPS detectors to collect multiple frames per second (a movie), to correct for beam-induced motion and sample drift within a single recording [25]. Combined with advances in microscope automation [26, 27] and processing algorithms [28], MAPS detectors have enabled structural biologists to reveal macromolecular structures at near-atomic resolution using cryo-EM single particle analysis (SPA) [29–31].

1.1.3 Hybrid pixel detectors

At the LMB in Cambridge, parallel to and at roughly the same time as the early testing of the MAPS detectors, there was testing of yet another breed of detectors: Hybrid Pixel Detectors (HPDs). HPDs have been in development since the early 1990s at the European Organisation for Nuclear Research (CERN) and have their origins in high-energy particle physics [32]. Conceptually, their origin can even be traced back to the cloud chambers used in the 1920s and later to detect and visualise high-energy ionising particles, such as alpha or beta rays [33]. These energetic charged particles interact with the saturated gaseous mixture inside the cloud chamber by knocking electrons off gas molecules via electrostatic forces during collisions, resulting in a visible trail of an ionised gas particle.

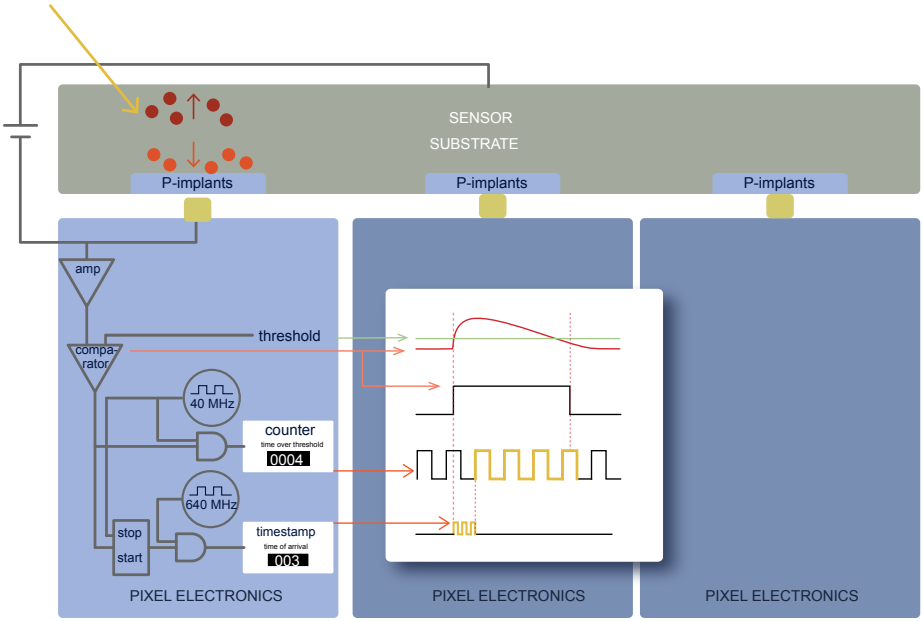


Figure 1.1: Schematic description of the Timepix3 hybrid pixel detector. A separate sensor and electronics layer is a feature shared between all hybrid pixel detectors. In yellow, a primary incident electron enters the sensor layer. This generates electron-hole pairs (red-orange), which will drift under the influence of the bias voltage via the electrodes and bump bonds (ocker) to the individual pixel. Inside the pixel, the signal is amplified, thresholded and digitally counted.

HPDs are characterised by their electronics being separate from the sensor layer (Figure 1.1). The sensor layer consists of a pixelated piece of semiconductor with individual bump bonds that connect the p-implants in the sensor layer to the electronics chip. Analogous to the cloud chambers, high-energy ionising particles will traverse the semiconductor material and leave behind electron-hole pairs. These electron-hole pairs will drift, under the influence of a bias voltage, to the electrodes. The electrodes are connected by bump bonds to the pixels. Inside the pixel, the signal is amplified and thresholded in an analogue circuit, before being digitised and counted in a digital circuit. The digital signal then leaves the pixel and is aggregated before being read-out by specialised hardware.

The multitude of different applications outside of high-energy particle physics HPDs could be used for was soon realised. This led to the founding of the first Medipix consortium in the late 1990s, with the submission of the first Medipix

chip in 1997 [33]. Aside from, for example, X-ray imaging and dosimetry, electron microscopy was also considered early on. The first testing of a hybrid pixel detector was only a couple of years later, at the turn of the century, by Faruqi and others [34]. Their testing of a Medipix2 chip showed it was technical feasible to use an HPD for EM and highlighted the benefits of having per-pixel thresholding of the signal, resulting in a noiseless dark image.

After initial successes with Medipix chips, it was realised that more information could be extracted about the ionising particle traversing the sensor layer [35]. Instead of only counting whenever a charged particle was over the comparator threshold, it was also possible to count how long the signal was over the threshold (Time over Threshold), and the exact moment the signal crossed the threshold (Time of Arrival). This led in 2006 to the submission of the first Timepix chip [36].

In the next iteration of the Timepix chip, which was dubbed Timepix3 because it was part of the Medipix3 consortium, two major new features were added to the chip [37]. First, it became possible to simultaneously measure and read the ToA and ToT values. With the first Timepix chip, a choice had to be made per recorded frame. Second, the readout was no longer purely frame-based, but it also became possible to have data-driven readout. In this mode, events are made available for readout as soon as they come in. This can be a more efficient way to read out in sparse data situations. Finally, an extra fine ToA clock was added, which increased the time resolution to 640 MHz. The Timepix3 chip is shown schematically in Figure 1.1.

1.2 Managing research data in life sciences

The advance of the use of digital detectors in electron microscopy is a showcase of a pattern happening across all domains of modern life science research. As a consequence of the digital processing and increased automation of workflows, researchers are now dealing with vastly increasing volumes of data and metadata. This is not for lack of reason. Modern life sciences studies depend on the collection, management, and analysis of comprehensive datasets. The focus in life sciences is no longer on the collection of a single sample but on arrays of samples analysed and collected in parallel. Researchers have opted for multimodal approaches in their experiments because multiple techniques are required to reveal better insights into the dynamic molecular mechanisms underlying biochemical processes. The limiting factor in research nowadays is the speed at which researchers can analyse their data rather than the amount of samples that can be processed [38].

This challenge appears in all aspects of the research data life cycle, such as

processing, description, sharing, and publishing of (raw) data and metadata. Parallel to this, there has been an increasing urgency to openly publish research data. Initiatives such as FAIR and Open Science, among others, have stimulated and urged researchers to share annotated research data openly and in a structured way to aid in more reproducible and reusable science [39].

In life sciences, as in many other fields, the research is conducted by small teams, and the bulk of data are generated by PhD students with fixed short-term contracts. The consequence of this combination of increasing (meta)data volumes, top-down initiatives, and lack of (long-term) incentives for researchers means that at worst no research data are made available, or at best only in forms which hampers actual reuse. The solutions to this must come from both technical solutions and organisational changes.

1.3 Scope of the thesis

This thesis focusses on the following two aspects. (1) The integration, characterisation, optimisation and application of the Timepix3 hybrid pixel detector to support versatile workflows for observing macromolecular structures using cryo-electron microscopy. (2) The description, storage, and management of life sciences research data and their organisation to help make research data more findable, accessible, interoperable, and reusable.

In **Chapter 2** the use of the Timepix3 hybrid pixel detector is characterised for cryo-EM. We build Monte Carlo simulations of electrons interacting with the Timepix3 and provide this as ground-truth data for the training of a convolutional neural network for the sub-pixel localisation of the incident position of electrons. Here, we make use of the special per-pixel spectroscopic properties of Timepix3. In **Chapter 3** a structure is resolved to a 3.0 Å resolution using the Timepix3 and the developed event localisation algorithm. We describe how we integrate Timepix3 in a cryo-EM workflow and optimise the hardware and software setup for SPA. We compare our results with a commercial MAPS detector on the same sample. In **Chapter 4** the use of a single layer of conductive graphene deposited on a grid is investigated, using Timepix3, as a way to mitigate the effects of electrostatic charging the sample by the electron beam.

In **Chapter 5** a case study is made of the DataHub Maastricht research data management (RDM) support group and the lessons learnt after being in operation for five years. We discuss not only technical solutions, in terms of flexible and powerful software to deal with the volume of data and metadata, but also solutions on an organisational level. Aspects such as the role of data stewards and the incentives for researchers to improve their RDM are discussed. In **Chapter 6** details and an implementation to improve the description and structure of

metadata in the RDM platform iRODS are given.

Finally, both aspects of this thesis and their future prospects are discussed together in the **General Discussion**. We describe how current detector development, and future ones such as those of the Timepix4, will pave the way towards more efficient low-dose cryo-EM workflows. The current developments and our lessons learnt in providing RDM support are placed in context of the broader FAIR and Open Science community, and we discuss how we can ensure that at a minimum the most reuseable data are made available.

1.4 References

1. Knoll, M. & Ruska, E. Das Elektronenmikroskop. *Zeitschrift für Physik* **78,** 318–339. doi:10.1007/bf01342199 (1932).
2. Ruska, E., Binnig, G. & Rohrer, H. Nobel Prize in Physics. *Binnig and Rohrer cited" for their design of the scanning tunneling microscope* (1986).
3. Ruska, H., v. Borries, B. & Ruska, E. Die Bedeutung der Übermikroskopie für die Virusforschung. *Archiv für die gesamte Virusforschung* **1,** 155–169. doi:10.1007/bf01243399 (1939).
4. Ruska, H. Die Sichtbarmachung der bakteriophagen Lyse im Übermikroskop. *Naturwissenschaften* **28,** 45–46. doi:10.1007/bf01486931 (1940).
5. Broglie, L. D. Recherches sur la théorie des Quanta. *Annales de Physique* **10,** 22–128. doi:10.1051/anphys/192510030022 (1925).
6. Taylor, K. A. & Glaeser, R. M. Electron microscopy of frozen hydrated biological specimens. *Journal of Ultrastructure Research* **55,** 448–456. doi:10.1016/s0022-5320(76)80099-8 (1976).
7. Glaeser, R. M. Limitations to significant information in biological electron microscopy as a result of radiation damage. *J Ultrastruct Res* **36,** 466–82. doi:10.1016/s0022-5320(71)80118-1 (1971).
8. Henderson, R. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q Rev Biophys* **28,** 171–93. doi:10.1017/s003358350000305x (1995).
9. Karuppasamy, M., Nejadasl, F. K., Vulovic, M., Koster, A. J. & Ravelli, R. B. G. Radiation damage in single-particle cryo-electron microscopy: effects of dose and dose rate. *Journal of Synchrotron Radiation* **18,** 398–412. doi:10.1107/s090904951100820x (2011).
10. Marton, L. Electron Microscopy of Biological Objects. *Physical Review* **46,** 527–528. doi:10.1103/physrev.46.527 (1934).

11. Mulvey, T. Origins and historical development of the electron microscope. *British Journal of Applied Physics* **13,** 197. doi:10.1088/0508-3443/13/5/303 (1962).

12. Cooley, J. W. & Tukey, J. W. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation* **19,** 297–301. doi:10.1090/s0025-5718-1965-0178586-1 (1965).

13. DeRosier, D. J. & Moore, P. B. Reconstruction of three-dimensional images from electron micrographs of structures with helical symmetry. *Journal of Molecular Biology* **52,** 355–369. doi:10.1016/0022-2836(70)90036-7 (1970).

14. Ruijter, W. J. D. Imaging properties and applications of slow-scan charge-coupled device cameras suitable for electron microscopy. *Micron* **26,** 247–275. doi:10.1016/0968-4328(95)00054-8 (1995).

15. Boyle, W. S. & Smith, G. E. Charge Coupled Semiconductor Devices. *Bell System Technical Journal* **49,** 587–593. doi:10.1002/j.1538-7305.1970.tb01790.x (1970).

16. Mochel, M. E. & Mochel, J. M. A CCD imaging and analysis system for the VG HB5 STEM. *Proceedings, annual meeting, Electron Microscopy Society of America* **44,** 616–617. doi:10.1017/s0424820100144528 (1986).

17. Krivanek, O. L. & Mooney, P. E. Applications of slow-scan CCD cameras in transmission electron microscopy. *Ultramicroscopy* **49,** 95–108. doi:10.1016/0304-3991(93)90216-k (1993).

18. Turchetta, R. *et al.* A monolithic active pixel sensor for charged particle tracking and imaging using standard VLSI CMOS technology. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **458,** 677–689. doi:10.1016/s0168-9002(00)00893-7 (2001).

19. Faruqi, A. R., Henderson, R., Pryddetch, M., Allport, P. & Evans, A. Direct single electron detection with a CMOS detector for electron microscopy. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **546,** 170–175. doi:10.1016/j.nima.2005.03.023 (2005).

20. Faruqi, A. R. & McMullan, G. Electronic detectors for electron microscopy. *Quarterly Reviews of Biophysics* **44,** 357–390. doi:10.1017/s0033583511000035 (2011).

21. McMullan, G. *et al.* Experimental observation of the improvement in MTF from backthinning a CMOS direct electron detector. *Ultramicroscopy* **109,** 1144–1147. doi:10.1016/j.ultramic.2009.05.005 (2009).

22. McMullan, G., Clark, A. T., Turchetta, R. & Faruqi, A. R. Enhanced imaging in low dose electron microscopy using electron counting. *Ultramicroscopy* **109,** 1411–1416. doi:10.1016/j.ultramic.2009.07.004 (2009).

23. Battaglia, M., Contarato, D., Denes, P. & Giubilato, P. Cluster imaging with a direct detection CMOS pixel sensor in Transmission Electron Microscopy. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **608,** 363–365. doi:10.1016/j.nima.2009.07.017 (2009).

24. Faruqi, A. R. & McMullan, G. Direct imaging detectors for electron microscopy. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **878,** 180–190. doi:10.1016/j.nima.2017.07.037 (2018).

25. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat Methods* **10,** 584–90. doi:10.1038/nmeth.2472 (2013).

26. Potter, C. S. *et al.* Leginon: a system for fully automated acquisition of 1000 electron micrographs a day. *Ultramicroscopy* **77,** 153–161. doi:10.1016/s0304-3991(99)00043-1 (1999).

27. Mastronarde, D. N. SerialEM: A Program for Automated Tilt Series Acquisition on Tecnai Microscopes Using Prediction of Specimen Position. *Microscopy and Microanalysis* **9,** 1182–1183. doi:10.1017/s1431927603445911 (2003).

28. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* **180,** 519–30. doi:10.1016/j.jsb.2012.09.006 (2012).

29. Liao, M., Cao, E., Julius, D. & Cheng, Y. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504,** 107–112. doi:10.1038/nature12822 (2013).

30. Amunts, A. *et al.* Structure of the Yeast Mitochondrial Large Ribosomal Subunit. *Science* **343,** 1485–1489. doi:10.1126/science.1249410 (2014).

31. Kühlbrandt, W. The Resolution Revolution. *Science* **343,** 1443–1444. doi:10.1126/science.1251652 (2014).

32. Heijne, E. H. M. *et al.* Development of silicon micropattern pixel detectors. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **348,** 399–408. doi:10.1016/0168-9002(94)90768-4 (1994).

33. Heijne, E. H. Semiconductor micropattern pixel detectors: a review of the beginnings. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **465,** 1–26. doi:10.1016/s0168-9002(01)00340-0 (2001).

34. Faruqi, A. R., Cattermole, D. M., Henderson, R., Mikulec, B. & Raeburn, C. Evaluation of a hybrid pixel detector for electron microscopy. *Ultramicroscopy* **94,** 263–276. doi:10.1016/s0304-3991(02)00336-4 (2003).

35. Ballabriga, R., Campbell, M. & Llopart, X. Asic developments for radiation imaging applications: The medipix and timepix family. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **878,** 10–23. doi:10.1016/j.nima.2017.07.029 (2018).

36. Llopart, X., Ballabriga, R., Campbell, M., Tlustos, L. & Wong, W. Timepix, a 65k programmable pixel readout chip for arrival time, energy and/or photon counting measurements. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **581,** 485–494. doi:10.1016/j.nima.2007.08.079 (2007).

37. Poikela, T. *et al.* Timepix3: a 65K channel hybrid pixel readout chip with simultaneous ToA/ToT and sparse readout. *Journal of Instrumentation* **9,** C05013–C05013. doi:10.1088/1748-0221/9/05/c05013 (2014).

38. Mons, B. *Data Stewardship for Open Science* doi:10.1201/9781315380711 (Chapman and Hall/CRC, 2018).

39. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3,** 160018. doi:10.1038/sdata.2016.18 (2016).

2

Sub-pixel electron detection using a convolutional neural network

Adapted from: **Van Schayck, J. P**, van Genderen, E., Maddox, E., Roussel, L., Boulanger, H., Fröjdh, E., Abrahams, J.-P., Peters, P. J. & Ravelli, R. B. G. Sub-pixel electron detection using a convolutional neural network. *Ultramicroscopy* **218,** 113091. doi:10.1016/j.ultramic.2020.113091 (2020).

Abstract

Modern direct electron detectors (DEDs) provided a giant leap in the use of cryo-electron microscopy (cryo-EM) to study the structures of macromolecules and complexes thereof. However, the currently available commercial DEDs, all based on the monolithic active pixel sensor, still require relative long exposure times and their best results have only been obtained at 300 kV. There is a need for pixelated electron counting detectors that can be operated at a broader range of energies, at higher throughput and higher dynamic range. Hybrid Pixel Detectors (HPDs) of the Medipix family were reported to be unsuitable for cryo-EM at energies above 80 kV as those electrons would affect too many pixels. Here we show that the Timepix3, part of the Medipix family, can be used for cryo-EM applications at higher energies. We tested Timepix3 detectors on a 200 kV FEI Tecnai Arctica microscope and a 300 kV FEI Tecnai G2 Polara microscope. A correction method was developed to correct for per-pixel differences in output. Timepix3 data were simulated for individual electron events using the package Geant4Medipix. Global statistical characteristics of the simulated detector response were in good agreement with experimental results. A convolutional neural network (CNN) was trained using the simulated data to predict the incident position of the electron within a pixel cluster. After training, the CNN predicted, on average, 0.41 pixel and 0.50 pixel from the incident electron position for 200 keV and 300 keV electrons respectively. The CNN improved the MTF of experimental data at half Nyquist from 0.50 to 0.68 at 200 kV, and from 0.06 to 0.65 at 300 kV respectively. We illustrate that the useful dose-lifetime of a protein can be measured within a 1 second exposure using Timepix3.

2.1 Introduction

Cryo-electron microscopy (cryo-EM) of biological samples depends on the recording of a small number of incident electrons that can be used to form an image before the sample is destroyed by radiation damage. Therefore the detector plays a more important role for these samples than for less or non-radiation sensitive samples. The emergence of direct electron detectors in the form of the monolithic active pixel sensors (MAPS) has been a breakthrough in cryo-EM [reviewed in 1]. Combined with advances in processing and software algorithms, the direct electron detector has enabled structural biologists to reveal macromolecular structures at near-atomic resolution using cryo-EM [reviewed in 2, 3].

One of the key performance indicators used for a detector in cryo-EM is its modulation transfer function (MTF) [4, 5]. The MTF measures the transfer of contrast as function of spatial frequency and is thereby a direct measure of the spatial resolution of a detector. An ideal detector has MTF of unity at all frequencies, however, due to a detectors finite pixel size, it decreases with increasing spatial frequency. A second, often used, performance indicator is the detective quantum efficiency (DQE), which gives the ratio between the squared signal to noise of the outgoing signal and the squared signal to noise of the incoming signal [6, 7]. The detectors ability to maintain signal and not add noise is especially important for experiments where the available signal is limited by radiation damage to the sample.

The MAPS technology already existed but considerable improvement for cryo-EM was achieved by back-thinning the sensor to, eventually, less than 30 μm thickness. The effect of back-thinning is the reduced chance of an electron scattering back into the sensor layer and depositing its energy in pixels adjacent to the point of impact. This back scattering will degrade the MTF performance of a detector [8]. With the current generation of MAPS detector (Thermo Fisher Falcon3 [9][1], Gatan K3 and Direct Electron DE-64) the limit of back-thinning has been reached (Guerrini, personal communication).

Combined with back-thinning, a further enhancement was the introduction of electron counting mode in which every electron is localised and given an equal weight [11]. A number of different algorithms for event localisation have been published [12, 13]. For the current range of MAPS detectors, these algorithms require a sufficiently low electron flux to prevent the signal from adjacent incident electrons from overlapping with each other within one frame. Low electron fluxes provide multiple downsides; (a) The lengthy exposure times, up to one minute or more for Falcon3 detectors, compromises the total throughput during a single particle analysis (SPA) study. Sample drift is very significant during long

[1] During the review process of this manuscript the Falcon4 was announced[10].

exposures and needs to be corrected for. (b) These detectors are unsuitable for diffraction experiments, which would wreck the counting algorithm and might even damage the detector by radiation [14]. (c) At the onset of the exposure of a pristine biological sample, one would expect to be able to observe the highest resolution features of such sample. However, in practice, the high resolution information of the early frames within a recording are down weighted or even removed as they are disturbed by factors such as beam induced movements [15–17]. These fast movements cannot be accurately corrected for from the data from the current DEDs. Details of the specific algorithms used by commercial MAPS detectors have not been released, but accurate event localisation is still seen as a bottleneck [1].

An alternative approach to back-thinning of the detector is to fully absorb the electron and to obtain as much information about it as possible, thereby optimising the ability to computationally localise the point of impact. Hybrid pixel detectors (HPD) allow this alternative approach and can have a much higher readout speed. HPDs are characterised by having their application-specific integrated circuit (ASIC) separate from the sensor layer. The sensor layer consists of a pixelated piece of semiconductor with individual bump bonds connecting it to the readout electronics chip. Such detector has recently been used to collect preliminary SPA data at 100 kV [18, 19]. The Medipix family of HPDs have a relatively large pixel size of 55 μm when compared to MAPS detectors, but this allows for more per-pixel electronics. In the Medipix HPDs, each pixel has its own signal threshold and counter and this gives it a noise-free readout and high dynamic range [reviewed in 20]. These properties have made them very appropriate sensors for electron diffraction and STEM [21–24]. Any electron moving through silicon with an energy over 80 kV will spread well beyond the 55 μm pixel pitch. It was shown that by increasing the signal threshold far beyond the noise edge of the Medipix such that each electron will be recorded in one single pixel, a near perfect MTF could be restored [25]. However, in that case some electrons remain undetected and this reduces the detectors DQE. To overcome the problem of the electron affecting multiple pixels, Mir *et al.* employed the Medipix3RX equipped with a charge sharing mode (CSM) that effectively combines the charge of a particle in four adjoining pixels to one pixel [26, 27]. It accomplishes this by having additional logic in the analog front-end of the ASIC. They also showed that the Medipix3RX using CSM for 80 keV and 120 keV can maximise the MTF without leaving electrons undetected [28]. Above 120 kV the charge is spread beyond the four pixels used by the CSM and the MTF still degrades. Therefore, at those energies, more information about the primary electron track is required to localise the point of impact accurately.

The Timepix3 is currently the latest generation of the Medipix HPD family [29]. This ASIC is capable of simultaneously measuring the time that a signal

is over the comparator threshold (time over threshold; ToT) and the timing of when the signal crosses the comparator threshold (time of arrival; ToA). The ToT effectively gives a measure for the amount of energy deposited in a pixel. The ToA accuracy is determined by the ASIC's fast ToA clock which can achieve a time binning of 1.5625 ns (640 MHz). For example, any 200 kV incident electron will travel most of a 300 µm thick silicon sensor layer at a velocity higher than 50% light speed and thus at much shorter time scales than 1.5625 ns. However, in passing, the incident electron will create electron hole pairs (E-H pairs) in the sensor. These charge carriers will drift towards the electrodes under influence of the bias voltage and will have a drift time well over 1.5625 ns. Drift velocity for holes in silicon are in the order of $10^6\,\mathrm{cm\,s^{-1}}$ [30]. Therefore in a fully depleted 300 µm silicon sensor layer drift times for holes can range up to 30 ns. The ToA information thus effectively gives a z-position where the incident electron created a E-H pair in the sensor layer [31]. This principle was used by Bergmann *et al*. for the 3D reconstruction of a 120 GeV high energy pion particles and accompanying 2 to 3 MeV delta rays [32].

Here, we show that the ToT and ToA information can be used to reconstruct the point of impact of the electron in the sensor layer at energies typically used in cryo-EM (200-300 keV). Instead of attempting to directly reconstruct the point of impact using the ToT and ToA information, we present an *ad hoc* prediction model using a convolutional neural network (CNN). Neural networks have become an increasingly popular machine learning method over recent years and CNNs are well suited for extracting 2D feature information from training input. Neural networks have been used in other detector applications for reconstructing the impact position of a particle hitting a sensor layer. At the ATLAS experiment of the Large Hadron Collider (at CERN) neural networks are being employed to separate and determine the direction of high energy particles hitting their detector planes [33]. For positron emission tomography (PET) neural networks have been employed to determine the incident position of a 512 keV X-ray hitting a scintillating crystal block [34, 35]. A notable difference between both the high energy particles (> TeV) detected in the ATLAS experiment and the visible light photons detected from a scintillating crystal in the case of PET, is that electrons will undergo multiple scattering events while travelling the sensor layer. This trajectory of the electron through multiple pixels, while erratic, follows rules and patterns. The neural network is, by training, recognising the pattern and able to deduct the trajectory and thereby the incident position of the electron.

For training a neural network, usually a ground-truth dataset is used; here, this would mean a dataset with known incident positions and their corresponding detector output. Within an electron microscope the electron beam cannot easily be limited to a sub-pixel area of the detector and therefore the incident position is not known with enough accuracy for training purposes. Instead, in here, we

have chosen to generate training data by means of simulation. After training the neural network, the obtained prediction model can be applied to experimental data to improve the MTF of the Timepix3. We implement a method for correcting the non-uniform ToT response of the Timepix3. Finally, we illustrate that we are able to image, in electron counting mode, the useful dose-lifetime of a protein within one second.

2.2 Installation of detector

The Timepix3 detector chip assembly, camera housing, SPIDR readout electronics board and control software were provided by Amsterdam Scientific Instruments (ASI) [23]. The vacuum camera housing provided temperature control through water cooling, X-ray shielding and a mechanical aluminium shutter positioned 5 mm above the sensor layer. The Timepix3 was assembled in a quad configuration giving a total of 512×512 pixels. In this configuration we reached a maximum readout speed of $110\,\mathrm{Mhit\,s^{-1}}$. The SPIDR readout board developed by the Dutch National Institute for Subatomic Physics (Nikhef) provided a $10\,\mathrm{Gbit\,s^{-1}}$ fibre optic connection to the detector PC [36]. The detector control and SPIDR readout software SoPhy were developed by ASI. Using SoPhy, a detector equalisation method was performed where the global threshold and the per pixel thresholds were set close to the electronic noise edge. The detector was mounted under a FEI Tecnai Arctica operating at 200 keV at Maastricht University (Figure 2.1), the Netherlands and under a FEI Tecnai G2 Polara operating at 300 keV at C-CINA, Basel, Switzerland. The sensor consisted of 300 μm and 500 μm of silicon in Maastricht and Basel, respectively. The sensor was connected as a single slab, via bump bonds, to the quad configured ASIC. The thickness was chosen to optimally absorb the respective 200 keV and 300 keV incident electrons within the sensor. Whereas the camera housing is identical in both setups, the flight tubes have different lengths, resulting in a post-magnification of 2.06 and 1.28 relative to the nominal magnification for Maastricht and Basel, respectively. The field of view of the images collected on the Polara was restricted in a circular fashion by the dimensions of the flight tube. Full dose radiation shielding checks were performed successfully on both sites.

2.3 Monte Carlo simulations

Monte Carlo simulations provided us with the training data of an incident electron hitting a known position on the sensor layer and its potential detector responses. The *Geant4* framework has been widely applied for Monte Carlo simulations of

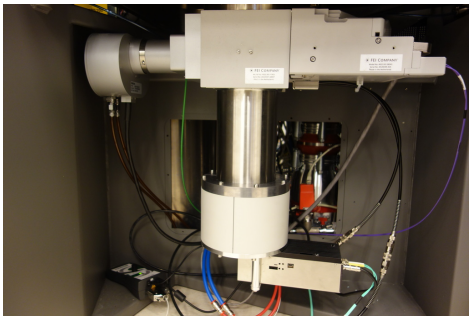


Figure 2.1: Timepix3 camera installed at a FEI Tecnai Arctica (200 kV, cryo-EM) in Maastricht, the Netherlands.

particles passing through matter [37]. Using the *Geant4* framework, the simulation tool *Geant4Medipix* has been developed to simulate a number of chips from the Medipix family, including the Timepix3 [38, 39]. We used it to simulate the passing of electrons through the sensor, the drifting of the E-H pairs to the electrodes as well as the readout electronics response. The *Geant4Medipix* code was adapted to be multi-threaded and provide HDF5 output such that a large number ($n = 50000$) of events can be simulated efficiently. Simulations were performed for a 300 µm or 500 µm thick silicon slab with 20×20 pixels at a 55 µm pitch. The simulated readout electronics were configured to match the properties of the Timepix3. Each simulation event consisted of one electron accelerated at 200 keV or 300 keV hitting a random position within one pixel of the 20×20 matrix. This ensured an even distribution of incident positions within one pixel. For each simulation event multiple pixels are contributing and this led to a variable number of hits in the output. Each hit contained information of both ToT and ToA. The lowest ToA value of a cluster was set to 0 where after the ΔToA of each hit in the cluster was calculated. The cluster output of each event was normalised to a 10×10 matrix with each cluster starting at the $0, 0$ position. This last step ensured that the simulated cluster output could be compared to the output obtained from the detector.

The output of the simulations were globally validated with experimental detector output using a variety of histograms. For each cluster, either experimental or simulated, the sum of ToT values was calculated. The simulations were scaled to the experimental data by adjusting the simulated Krummenacher current such that the most abundant summed ToT values would coincide [40]. The simulator includes a parameter to describe electronic noise contributions from the analogue front end. This parameter required minor tuning to further optimise the

concurrence between simulation and experiment.

Figure 2.2 shows a 2D histogram of the number of hits in a cluster versus the summation of the ToT values of all hits within such a cluster. The 2D histogram for simulated data (Figure 2.2a) is in good agreement with experimental data (Figure 2.2b). The main difference occurs in the lower left corner of the histograms, where the experimental data shows more counts. The histogram bin at (1,1) can be attributed to X-rays generated in the electron microscope which were not simulated. There is a tail visible between the (1,1) bin and the peak of the histogram in both the experimental and simulated data which can be attributed to backscattered electrons from the sensor. These electrons are not fully absorbed and leave smaller clusters behind with less energy deposited. They account for about 10% of the total number of electrons. The experimental data also contained data points such as stray cosmic rays, edge events, edge pixels, masked or unresponsive pixels, and coincident electrons, which were not included in the simulation.

2.4 Correcting for non-uniform pixel response in Time over Threshold

Every pixel of the Timepix3 has its own correlation between the amount of energy deposited and the time the signal is over the threshold (ToT). A transmission electron microscope provides a very monochromatic beam of less than 1 kV spread. This means that every pixel should have a similar distribution of the ToT signal. By using a flat field of the electron beam and sampling a large number of events it is possible to correct for non-uniformities in the silicon sensor layer and the pixel electronics. These inhomogeneities can be attributed to local differences in the lithography process. Using a minimum of 10 Ghit of flat field data, a per-pixel correction was calculated. For this correction, an average response was calculated with roughly 50% of the active pixels. Pixels with a response too far from the average were discarded. This average response was then transformed into a normalised cumulative curve. This curve was used as an ideal reference cumulative distribution curve and was fitted with multiple 3^{th}-order polynomial. For every pixel the cumulative distribution curve was calculated and by means of histogram matching a new ToT distribution was created for each of these pixels by being matched to the conclusive 3^{th}-order polynomial curves (Figure 2.3a and b). This created a per-pixel correction look-up table ($512 \times 512 \times 1024$) of a measured ToT value to a corrected ToT value (Figure 2.3c).

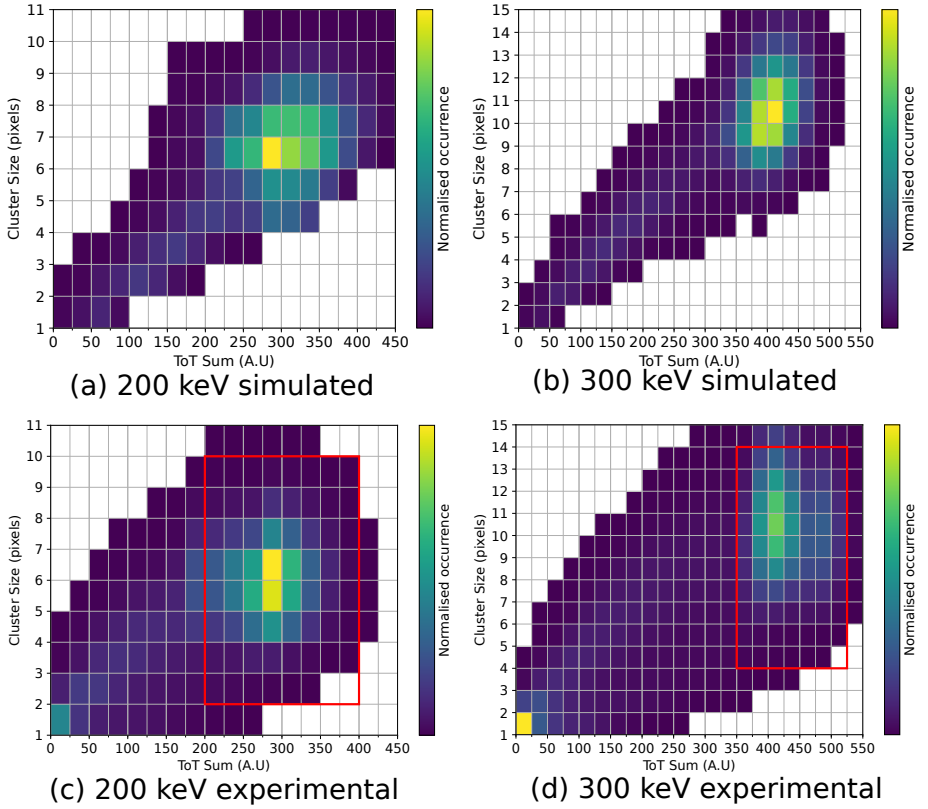


Figure 2.2: 2D histogram of the number of pixels contributing to an event (cluster size) versus the ToT sum (sum of ToT values of a cluster) at 200 kV (**a** and **c**) and 300 kV (**b** and **d**) using simulated (**a** and **b**) or experimental data (**c** and **d**). The red box in (**c**) and (**d**) denotes the boundaries used for filtering electrons from other detected events (Section 2.6). On average, after filtering, 4.9 pixels are contributing to a 200 kV electron event and 8.9 pixels are contributing to a 300 kV electron event.

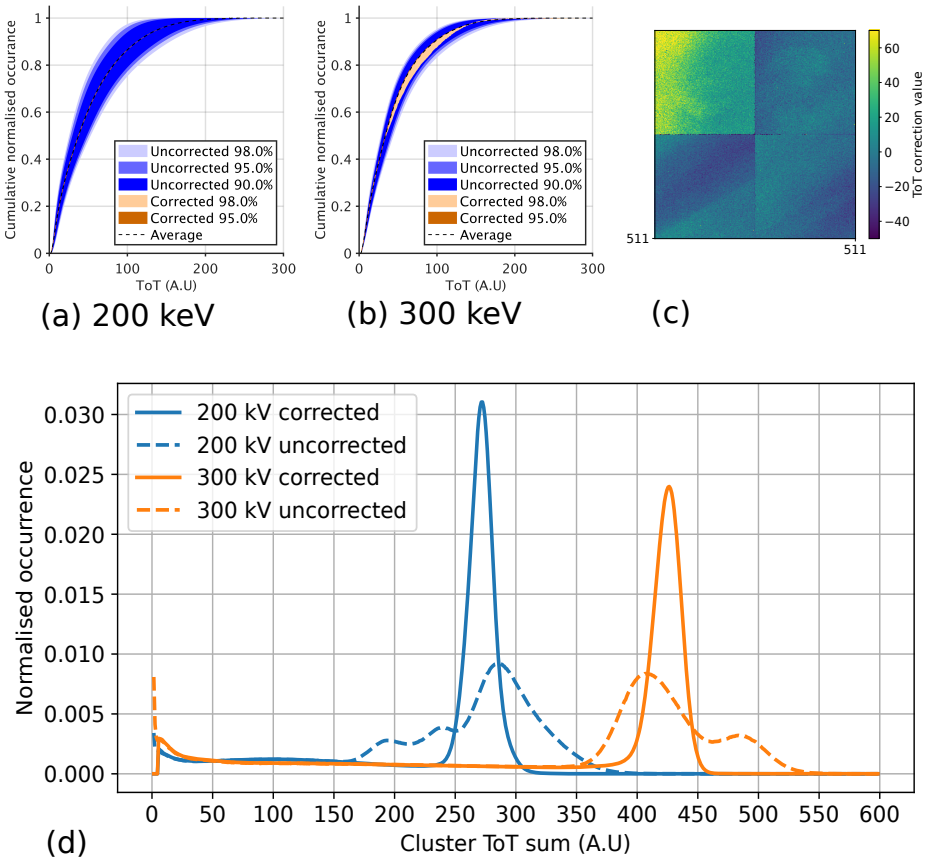


Figure 2.3: The percentile deviation from the average for the uncorrected (blue) and the corrected (orange) cumulative response for 200 kV (**a**) and 300 kV (**b**). In (**c**) an example spatial distribution of ToT correction values is shown (at 200 keV and ToT=100). In (**d**) the normalised occurrence of clusters versus the cluster ToT sum is plotted for both corrected and uncorrected data at 200 kV and 300 kV. From the FWHM of the corrected curves an energy resolution of 13.6 kV and 23.5 kV for 200 kV and 300 kV respectively has been determined.

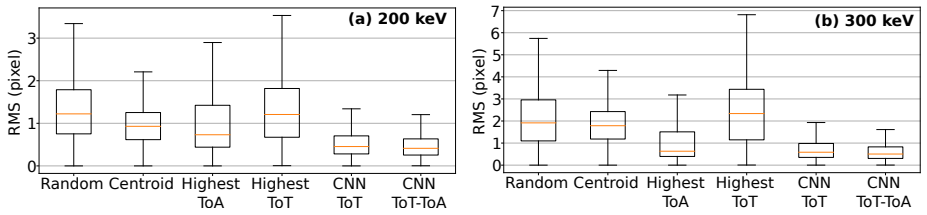


Figure 2.4: Distance (in pixels) between incident position and the determined position of different localisation methods for 200 kV (**a**) and 300 kV (**b**). Boxes represents Q1-Q3. Orange line is median.

2.5 Methods for incident electron event localisation

Six different algorithms for electron localisation were tested: random position (as a control), centroid, highest ToT, highest ToA, CNN trained on ToT (CNN-ToT) and CNN trained on ToT and ToA (CNN-ToT-ToA). The random method selects a random sub-pixel position within the pixels forming a cluster. The centroid method calculates the geometrical center of a cluster where each hit has been weighted according to their ToT values. The highest ToT and highest ToA method selects the centre of the pixel with the highest ToT and ToA value respectively or randomly between them in case of a draw.

Training of the CNN was performed using 50,000 independently simulated events on either just the ToT channel or on both the ToT and ToA channel. The CNN used the simulated 10×10 matrix as input and consists of a separable 2D convolutional layer followed by three times a drop-out/dense layer to gradually reduce to an x and y output. It used the Adam optimiser and ReLU activation method [41, 42]. It was trained for 200 epochs towards convergence. The CNN has been implemented in Tensorflow 1.4 using Python3 and Keras 2.1.

The event localisation methods were tested using 50,000 events for both 200 keV electrons and 300 keV electrons and calculated the mean distance, the root mean square deviation (RMSD) as well as the median between the simulated incident positions and the predicted electron positions. Figure 2.5 shows three example prediction plots and Figure 2.4 shows the prediction distance for each method. The CNN-ToT-ToA method predicts the point of impact of the incident electrons within 0.50 (0.62, 0.41) and 0.68 (0.92, 0.50) pixel on average (RMSD, median) at 200 keV and 300 keV respectively.

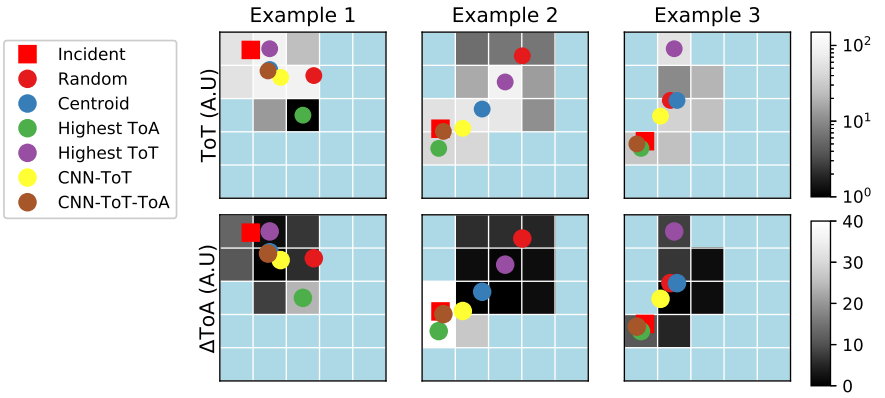


Figure 2.5: Three examples of simulated detector output at 300 kV. The simulated incident position is shown as red square and all event localisation methods are circles in their respective colour. Light blue pixels did not receive a hit.

2.6 Processing pipeline to form an image from raw data

The Timepix3 ASIC delivered, in data driven mode, a stream of hits, each hit containing positional, ToT and ToA information. Several steps were needed to process these raw hits into a final image or movie (Figure 2.6). First, the ToT values within the raw data stream were corrected using the obtained ToT lookup table (Section 2.4). Subsequently, we searched for clusters of hits which are formed by a single incident electron. The DBSCAN clustering algorithm [43] was chosen experimentally for its speed in handling clusters of various sizes. The euclidean distance between all hits was calculated and the DBSCAN parameter ϵ (eps), specifying the radius of a neighbourhood with respect to another hit, was set to 1. Clusters were formed by hits which were directly adjacent to each other in time and space: adjacent hits were selected that occur within a time interval of 50 ToA clock ticks (~78 ns). The 50 ToA clicks were scaled to a value of 1, to match the euclidean distance of 1. Then, clusters were filtered based on their cluster size and cluster ToT sum. These values were between 2 and 10 for the cluster size and between 200 and 400 for cluster ToT sum for 200 kV and 4 and 14 for the cluster size and between 350 and 525 for the cluster ToT sum for 300 kV (Figure 2.2 c and d). The lowest ToA value of a cluster is set to 0 and thereby the ΔToA of each hit in the cluster was calculated. A sub-pixel position was determined for each individual cluster by the selected event localisation algorithms. Finally, these obtained positions were placed within the complete

image frame. To compensate for a skewed sub-pixel distribution the edges of the sub-pixels within the original pixel were adjusted such that the distribution became uniform. These adjustments were at most 5% (Figure 2.7).

The processing pipeline, including event localisation, was written in Python3 making use of the Numpy, Scipy, Keras and Tensorflow libraries [44]. Processing of 70 Mhit (typical 1 second exposure at an electron flux of $40\,e^-\,Å^{-2}\,s^{-1}$ at 200 keV) still lacks significantly behind with the exposure time. Data conversion currently takes about 2 min, the event localisation step about 3 min (Intel Xeon E5-2680, 20 cores, NVIDIA GeForce GTX 1080 Ti). The current cluster finding algorithm needs further optimisation.

Figure 2.6 b and c show the unprocessed hit image and the processed CNN-ToT event localised image of a protein sample recorded in Maastricht. A truncated mutant of the *Mycobacterium tuberculosis* protein EspB was expressed in *Escherichia coli*, affinity purified using a nickel-column on a His6-tag, SEC purified, concentrated to $1.3\,mg\,mL^{-1}$, prepared on R1.2/1.3 Quantifoil grids 300 Au mesh (www.quantifoil.com), using the Vitrobot (Blot force 5, blot time 4). Images were recorded over 2 seconds, using a pixel magnification of 1.24 Å and an electron flux of $40\,e^-\,Å^{-2}\,s^{-1}$. At 200 kV each electron hits on average 4.9 pixels (Figure 2.2c). This electron flux corresponded to $51.3\,Mhit\,s^{-1}$ at the detector and thus well within the achieved maximum of $110\,Mhit\,s^{-1}$.

2.7 Validation of prediction model

A visual way to asses the performance of various event localisation methods is to analyse the image of an EM grid [45]. We obtained low magnification images to ensure sample drift is minimal. The resulting images had high contrast with well-defined spots in their power spectra that extends beyond the Nyquist frequency of the detector. Figure 2.8 shows the image taken of UltrAuFoil 200 mesh grids at 225x magnification. To quantify the differences between an event localisation method and the original image, spots at increasing spatial resolution were chosen. From the ratio of the normalised amplitudes the MTF enhancement could be obtained (Figure 2.8 and supplemental Figure 2.10 and 2.11) [12].

The MTF of the detector with and without event localisation methods applied was measured using the knife edge method. The benefit of this method is that it measures the MTF over all spatial frequencies [8]. Knife edge images were obtained using the aluminium shutter positioned approximately one centimetre above the detector. The shutter was positioned to partially cover the detector. MTFs were calculated from the images formed by hits alone as well as from images reconstructed by the event localisation method (Figure 2.9) [46]. DQE is a less relevant measure here as, among others, the noise spectrum cannot

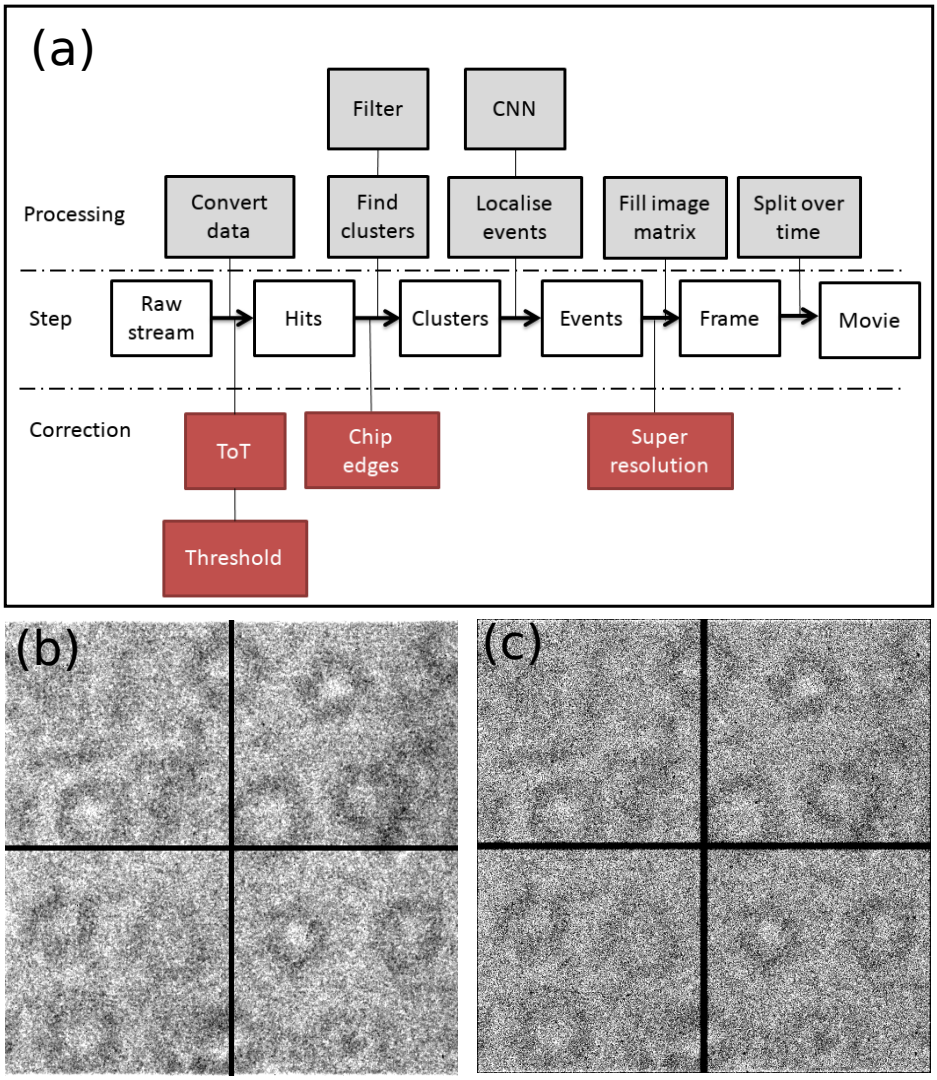


Figure 2.6: Schematic overview of steps taken during processing (**a**). Images of a truncated mutant of the *Mycobacterium tuberculosis* EspB protein without (**b**) and with (**c**) event localisation (CNN-ToT-ToA).

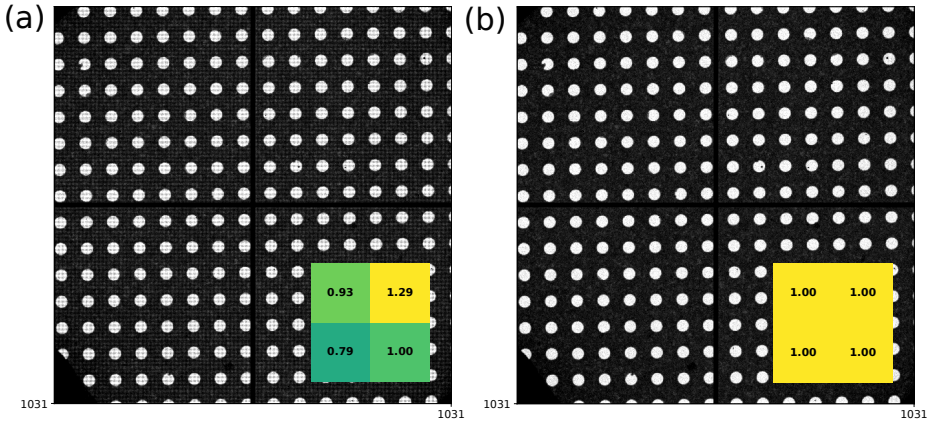


Figure 2.7: Image of UltrAuFoil 200 mesh grids at 225x magnification. Recorded at 200 kV and with event localisation (CNN-ToT). Images are shown at super-resolution (1032 pixels) without (**a**) and with sub-pixel correction (**b**). The insets show a 2D histogram of the normalised quad sub-pixel distribution.

be accurately determined due to replacement of events in counting algorithms [9]. The MTFs shown in Figure 2.9 validates the benefits of the presented event localisation methods described in here.

2.8 Discussion, Conclusion and Outlook

Every electron counts when using cryo-EM for the imaging of radiation sensitive samples. Accurate event localisation is seen as a bottleneck for optimal cryo-EM detector performance[1] . In here, we have shown a new method for event localisation using the Timepix3, with its unique ToT and ToA channels.

Like most pixelated detectors, the Timepix3 detector displays a non-uniform pixel response that needs to be corrected (Figure 2.3). The histogram normalisation described in Section 2.4 corrects for ToT inhomogeneities. Upon correction, we can measure the energy of each incident electron with a resolution of 13.6 and 23.5 kV for 200 kV and 300 kV respectively (Figure 2.3d). This allows us to distinguish primary electrons from other events such as cosmic and X-rays (Figure 2.2). The energy resolving power may have some applications for thick specimens, albeit its resolution remains far above the energy resolutions used in energy-filtered EM or electron energy loss spectroscopy.

Unfortunately, there was still a systematic pattern left after event-localisation

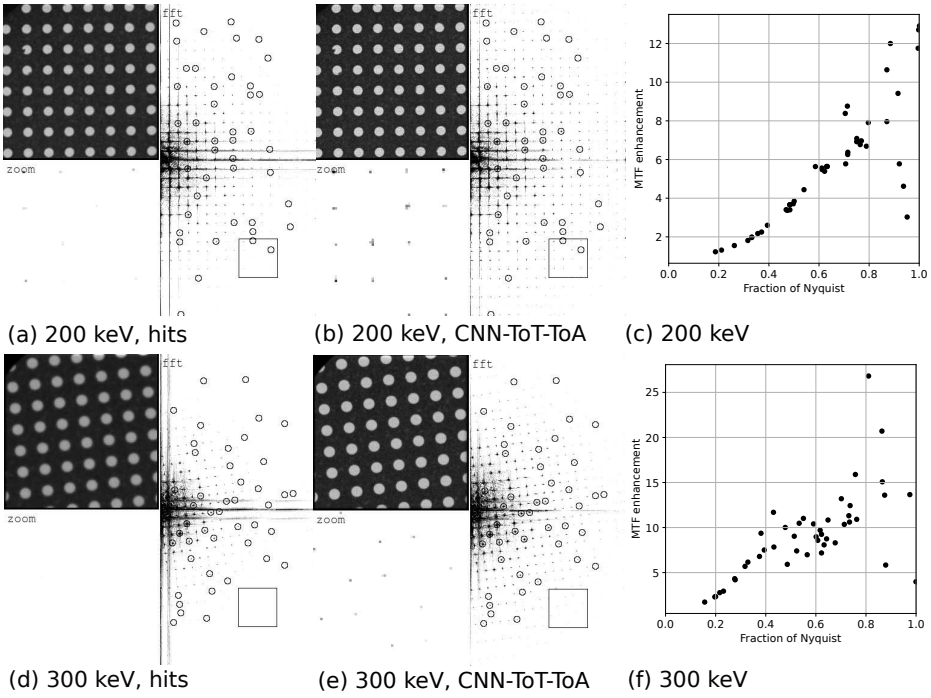


Figure 2.8: Images of UltrAuFoil 200 mesh grids at 225 times magnification at 200 keV (**a**,**b** and **c**) or 300 kV (**d**, **e** and **f**) with their corresponding power spectra. Images are shown with (**b** and **c**) and without (**a** and **d**) event localisation (CNN-ToT-ToA). The rectangle box in the power spectrum is shown zoomed in on the bottom left of the image. The MTF enhancement (**c** and **f**) was obtained from the normalised ratio of amplitudes measured with and without event localisation of the randomly chosen spots which are encircled in the images. In Supplemental Figure 2.10 and 2.11 all other described event localisation methods are included for 200 kV and 300 kV data respectively.

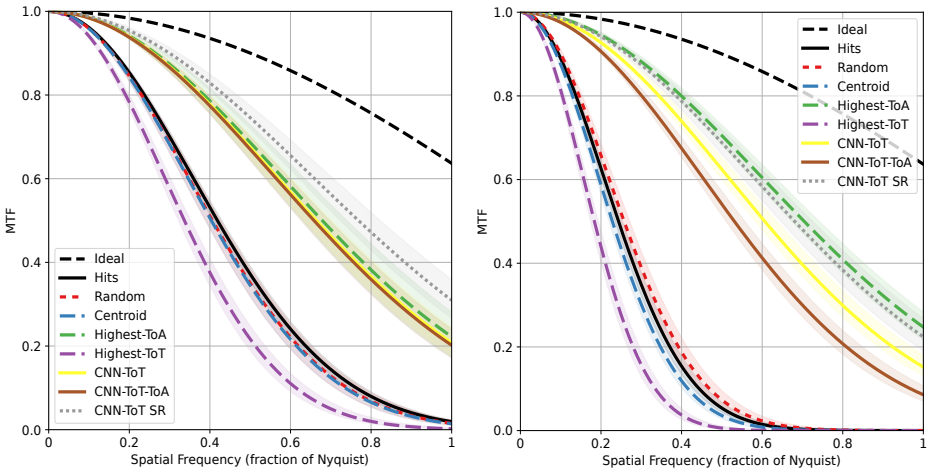


Figure 2.9: The modulation transfer function (MTF) at 200 keV (**a**) and 300 keV (**b**) as obtained from knife edge measurements. The shading represents the 3-sigma confidence interval. The super-resolution (SR) data is shown at the recorded spatial frequency.

using the CNN trained on both ToT and ToA information (Figure 2.12), which can be attributed to inhomogeneities in the ΔToA data. Our data suggests that there is a systematic difference in ToA response between pixels in the ASIC. The observed pattern hints at the source being the super pixel structure of the Timepix3 ASIC. Others have also reported seemingly similar systematic patterns and reported solutions in the form of extra calibration steps [47]. Such calibration steps would have required us to unmount the detector from the microscope. Future alternative calibration steps may overcome that requirement.

We could simulate the response of the Timepix3 detector using Monte Carlo simulations (Section 2.3). This enabled us to numerically evaluate the performance of the detector at different energies, for different sensor layers and thickness. The simulations confirmed that primary electrons would excite multiple pixels of a silicon sensor layer at energies > 80 kV, calling for accurate sub-pixel localisation schemes. The availability of the simulated electron events allowed us to use these as training data for machine learning methods.

We show that selecting the centre of a pixel with the highest ToT value (the pixel which received the most energy), provides the least accurate point-of-impact location, both at 200 kV and 300 kV (Figure 2.5 and 2.9). This may be due to the electron losing most of its energy in the last part of its trajectory [7]. The

highest-ToT method may thus, in many cases, be selecting the last pixel of an electron trajectory. Similarly, the centroid method is therefore selecting a pixel away from the point-of-impact.

The ToA channel provides unique data about the relative timing of the signal that each pixel within a cluster received. The ΔToA value of a pixel proved to be a very effective measure for the z-position of the trajectory of the incident electron (Section 2.1). A high ΔToA value reflects a pixel where the trajectory of the incident electron was high in the sensor layer. Such a high ΔToA value is therefore likely to be close to the incident electron position as the electrons travel mainly downwards.

Machine learning methods can improve on the point-of-impact localisation accuracy compared to traditional methods. We arrived at using a convolutional neural network (CNN) due to its suitability for handling multi-channel 2D features and ease of use in the available software frameworks. A CNN, capable of exploiting both ToA and ToT channels simultaneously, gave the best point-of-impact localisation results. Surprisingly, the CNN trained on only ToT information is, at both 200 kV and 300 kV, performing nearly as well as the model trained on both ToT and ToA information.

Analysing the sub-pixel assignments of the event localisation methods showed a non-uniform distribution (Figure 2.7). For both the experimental and the simulated data, an even distribution of electrons across each pixel was used. However, the different event localisation methods do not provide an even distribution as outcome. While it cannot be excluded that some of the non-uniform distribution originates from the event localisation model itself, it is striking that others have also reported non-uniform sub-pixel distributions using the Timepix3 [48]. This could hint at a possible problem with the ASIC. We applied a pragmatic approach to correct for the observed non-uniform sub-pixel assignments (Section 2.6).

We validated our CNN event localisation scheme by analysing high contrast images of UltrAuFoil grids and determining the MTF using the knife edge method (Section 2.7). Both CNN models show significant improvements. We compared the obtained MTF curves with the theoretical MTF values using Eq. 10 and Eq. 14 of [7]. Using the RMSD values of Figure 2.4 as σ ($=\lambda/\sqrt{2}$) within those equations, theoretical MTF curves were obtained that were slightly better then the observed ones (supplemental Figure 2.13). Future work on providing improved neural network training data and better accounting for residual ΔToA inhomogeneities, could minimise these differences.

Our results indicate that hybrid pixel detectors can be used as a counting direct electron detector for cryo-EM at 200 kV or 300 kV in imaging or diffraction mode. Using the Timepix3 we were able to image the entire useful dose lifetime ($40\,e^{-}Å^{-2}$) of a protein within a single second exposure. This provides great

prospects for single particle applications. Being able to work with a higher flux mitigates sample-stage drift issues and enhances throughput. As the Timepix3 in data driven mode does not record frames, but rather a stream of events, alternative data collection and storage schemes could be envisioned. A synchronised sample-beam movement with a continuous streaming of localised electrons could accelerate cryo-EM by at least an order of magnitude [49]. This will require several improvements. These include the handling of the area in between the adjacent quadrants of the chip, the effect of masked pixels and the systematic pattern in the ΔToA data. The tile-ability of the Timepix4 should address the limited field of view of the Timepix3, whereas the maximum count rate per detector area and ToA time resolution should also be improved for the Timepix4. As of this writing, the Timepix4 is being developed, just like numerous alternative detector developments. These could all make huge impact in getting better data faster, both in imaging and diffraction mode.

We envision that the use of neural networks could also help in improving the point-of-impact localisation procedures implemented for current MAPS detectors. Electrons make erratic tracks through the sensor layer which can be trained to a neural network when an accurate simulator is available. Alternatively, the neural network could be trained with experimental data provided a small sub-pixel beam could be directed to a known position within a pixel of such a detector. The success of a CNN trained on ToT alone makes us optimistic that the gap between actual and ideal detectors can be further narrowed in the years to come.

2.9 Acknowledgements

We would like to thank Martin van Beuzekom (Nikhef) and Yue Zhang for helpful discussions. We thank Abril Gijsbers and Ye Gao for providing protein samples. We are grateful to the M4i Microscopy CORE Lab team of FHML Maastricht University for their support and collaboration, with special thanks to Hans Duimel and kVin Knoops. We thank Amsterdam Scientific Instruments team members for their support in building and operating the detectors. We also acknowledge the following people: Ariane Fecteau-Lefebvre, Kenneth Goldie and Henning Stahlberg for their help on technical support, operation and the use of the Polara at C-CINA, respectively. This research received funding from the Netherlands Organisation for Scientific Research (NWO) in the framework of the Fund New Chemical Innovations, project MOL3DEM, number 731.014.248, as well as from the Province of Limburg, the Netherlands.

2.10 Statement about competing interests

The Maastricht University filed a patent (EP3525229) with some of the authors as inventors regarding event localisation as outlined in this manuscript.

2.11 Data Availability

The dataset [50] containing simulated data, unprocessed experimental data, the ToT-correction matrices, and the CNN models, together with instructions to reproduce results and figures, are publicly available on Zenodo.

2.12 Code Availability

The code for simulation [51], machine learning [52], ToT correction [53], processing experimental Timepix3 data [54], generating images [55] and generating the figures of this paper [52] are publicly available on Zenodo.

2.13 References

1. Faruqi, A. R. & McMullan, G. Direct imaging detectors for electron microscopy. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **878,** 180–190. doi:10.1016/j.nima.2017.07.037 (2018).
2. Fernandez-Leiro, R. & Scheres, S. H. W. Unravelling biological macromolecules with cryo-electron microscopy. *Nature* **537,** 339–346. doi:10.1038/nature19948 (2016).
3. Vinothkumar, K. R. & Henderson, R. Single particle electron cryomicroscopy: trends, issues and future perspective. *Q Rev Biophys* **49,** e13. doi:10.1017/s0033583516000068 (2016).
4. Ruijter, W. J. D. Imaging properties and applications of slow-scan charge-coupled device cameras suitable for electron microscopy. *Micron* **26,** 247–275. doi:10.1016/0968-4328(95)00054-8 (1995).
5. Vulovic, M., Rieger, B., van Vliet, L. J., Koster, A. J. & Ravelli, R. B. G. A toolkit for the characterization of CCD cameras for transmission electron microscopy. *Acta Crystallographica Section D: Biological Crystallography* **66,** 97–109. doi:10.1107/s0907444909031205 (2010).

6. Meyer, R. R. & Kirkland, A. I. Characterisation of the signal and noise transfer of CCD cameras for electron detection. *Microscopy Research and Technique* **49,** 269–280. doi:10.1002/(sici)1097-0029(20000501)49:3<269::aid-jemt5>3.0.co;2-b (2000).
7. McMullan, G., Chen, S., Henderson, R. & Faruqi, A. R. Detective quantum efficiency of electron area detectors in electron microscopy. *Ultramicroscopy* **109,** 1126–1143. doi:10.1016/j.ultramic.2009.04.002 (2009).
8. McMullan, G. *et al.* Experimental observation of the improvement in MTF from backthinning a CMOS direct electron detector. *Ultramicroscopy* **109,** 1144–1147. doi:10.1016/j.ultramic.2009.05.005 (2009).
9. Kuijper, M. *et al.* FEI's direct electron detector developments: Embarking on a revolution in cryo-TEM. *Journal of Structural Biology* **192,** 179–187. doi:10.1016/j.jsb.2015.09.014 (2015).
10. Nakane, T. *et al.* Single-particle cryo-EM at atomic resolution. *Nature* **587,** 152–156. doi:10.1038/s41586-020-2829-0 (2020).
11. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat Methods* **10,** 584–90. doi:10.1038/nmeth.2472 (2013).
12. McMullan, G., Clark, A. T., Turchetta, R. & Faruqi, A. R. Enhanced imaging in low dose electron microscopy using electron counting. *Ultramicroscopy* **109,** 1411–1416. doi:10.1016/j.ultramic.2009.07.004 (2009).
13. Battaglia, M., Contarato, D., Denes, P. & Giubilato, P. Cluster imaging with a direct detection CMOS pixel sensor in Transmission Electron Microscopy. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **608,** 363–365. doi:10.1016/j.nima.2009.07.017 (2009).
14. Rodriguez, J. A. *et al.* Structure of the toxic core of α-synuclein from invisible crystals. *Nature* **525,** 486–490. doi:10.1038/nature15368 (2015).
15. Scheres, S. H. Beam-induced motion correction for sub-megadalton cryo-EM particles. *eLife* **3,** e03665. doi:10.7554/elife.03665 (2014).
16. Russo, C. J. & Passmore, L. A. Progress towards an optimal specimen support for electron cryomicroscopy. *Current Opinion in Structural Biology* **37,** 81–89. doi:10.1016/j.sbi.2015.12.007 (2016).
17. Russo, C. J. & Henderson, R. Charge accumulation in electron cryomicroscopy. *Ultramicroscopy* **187,** 43–49. doi:10.1016/j.ultramic.2018.01.009 (2018).

18. Peet, M. J., Henderson, R. & Russo, C. J. The energy dependence of contrast and damage in electron cryomicroscopy of biological molecules. *Ultramicroscopy* **203,** 125–131. doi:10.1016/j.ultramic.2019.02.007 (2019).

19. Naydenova, K. *et al.* CryoEM at 100 keV: a demonstration and prospects. *IUCrJ* **6,** 1086–1098. doi:10.1107/s2052252519012612 (2019).

20. Ballabriga, R., Campbell, M. & Llopart, X. Asic developments for radiation imaging applications: The medipix and timepix family. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **878,** 10–23. doi:10.1016/j.nima.2017.07.029 (2018).

21. Nederlof, I., van Genderen, E., Li, Y.-W. & Abrahams, J. A Medipix quantum area detector allows rotation electron diffraction data collection from submicrometre three-dimensional protein crystals. *Acta Crystallographica Section D: Biological Crystallography* **69,** 1223–1230. doi:10.1107/s0907444913009700 (2013).

22. Krajnak, M., McGrouther, D., Maneuski, D., Shea, V. O. & McVitie, S. Pixelated detectors and improved efficiency for magnetic imaging in STEM differential phase contrast. *Ultramicroscopy* **165,** 42–50. doi:10.1016/j.ultramic.2016.03.006 (2016).

23. Van Genderen, E. *et al.* Ab initio structure determination of nanocrystals of organic pharmaceutical compounds by electron diffraction at room temperature using a Timepix quantum area direct electron detector. *Acta Crystallographica Section A: Foundations and Advances* **72,** 236–242. doi:10.1107/s2053273315022500 (2016).

24. Heidler, J. *et al.* Design guidelines for an electron diffractometer for structural chemistry and structural biology. *Acta Crystallographica Section D* **75,** 458–466. doi:10.1107/s2059798319003942 (2019).

25. McMullan, G. *et al.* Electron imaging with Medipix2 hybrid pixel detector. *Ultramicroscopy* **107,** 401–413. doi:10.1016/j.ultramic.2006.10.005 (2007).

26. Pennicard, D., Ballabriga, R., Llopart, X., Campbell, M. & Graafsma, H. Simulations of charge summing and threshold dispersion effects in Medipix3. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **636,** 74–81. doi:10.1016/j.nima.2011.01.124 (2011).

27. Ballabriga, R. *et al.* Medipix3: A 64k pixel detector readout chip working in single photon counting mode with improved spectrometric performance. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **633,** S15–S18. doi:10.1016/j.nima.2010.06.108 (2011).

28. Mir, J. A. *et al.* Characterisation of the Medipix3 detector for 60 and 80keV electrons. *Ultramicroscopy* **182,** 44–53. doi:10.1016/j.ultramic.2017.06.010 (2017).

29. Poikela, T. *et al.* Timepix3: a 65K channel hybrid pixel readout chip with simultaneous ToA/ToT and sparse readout. *Journal of Instrumentation* **9,** C05013–C05013. doi:10.1088/1748-0221/9/05/c05013 (2014).

30. Ottaviani, G., Reggiani, L., Canali, C., Nava, F. & Alberigi-Quaranta, A. Hole drift velocity in silicon. *Physical Review B* **12,** 3318–3329. doi:10.1103/physrevb.12.3318 (1975).

31. Filipenko, M., Gleixner, T., Anton, G. & Michel, T. 3D particle track reconstruction in a single layer cadmium-telluride hybrid active pixel detector. *The European Physical Journal C* **74,** 3013. doi:10.1140/epjc/s10052-014-3013-1 (2014).

32. Bergmann, B. *et al.* 3D track reconstruction capability of a silicon hybrid active pixel detector. *The European Physical Journal C* **77,** 421. doi:10.1140/epjc/s10052-017-4993-4 (2017).

33. collaboration, T. A. A neural network clustering algorithm for the ATLAS silicon pixel detector. *Journal of Instrumentation* **9,** P09009. doi:10.1088/1748-0221/9/09/p09009 (2014).

34. Delorme, S., Frei, R., Joseph, C., Loude, J.-F. & Morel, C. Use of a neural network to exploit light division in a triangular scintillating crystal. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **373,** 111–118. doi:10.1016/0168-9002(95)01511-6 (1996).

35. Bruyndonckx, P. *et al.* Neural Network-Based Position Estimators for PET Detectors Using Monolithic LSO Blocks. *IEEE Transactions on Nuclear Science* **51,** 2520–2525. doi:10.1109/tns.2004.835782 (2004).

36. Visser, J. *et al.* SPIDR: a read-out system for Medipix3 & Timepix3. *Journal of Instrumentation* **10,** C12028. doi:10.1088/1748-0221/10/12/c12028 (2015).

37. Allison, J. *et al.* Recent developments in Geant4. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **835,** 186–225. doi:10.1016/j.nima.2016.06.125 (2016).

38. Schübel, A., Krapohl, D., Fröjdh, E., Fröjdh, C. & Thungström, G. A Geant4 based framework for pixel detector simulation. *Journal of Instrumentation* **9,** C12018–C12018. doi:10.1088/1748-0221/9/12/c12018 (2014).

39. Krapohl, D., Schübel, A., Fröjdh, E., Thungström, G. & Fröjdh, C. Validation of Geant4 Pixel Detector Simulation Framework by Measurements With the Medipix Family Detectors. *IEEE Transactions on Nuclear Science* **63,** 1874–1881. doi:10.1109/tns.2016.2555958 (2016).

40. Krummenacher, F. Pixel detectors with local intelligence: an IC designer point of view. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **305,** 527–532. doi:10.1016/0168-9002(91)90152-g (1991).

41. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* (2014).

42. Nair, V. & Hinton, G. E. *Rectified Linear Units Improve Restricted Boltzmann Machines* in (Omnipress, 2010), 807–814.

43. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise* in (1996), 226–231.

44. Van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering* **13,** 22–30. doi:10.1109/mcse.2011.37 (2010).

45. McMullan, G., Turchetta, R. & Faruqi, A. R. Single event imaging for electron microscopy using MAPS detectors. *Journal of Instrumentation* **6,** C04001. doi:10.1088/1748-0221/6/04/c04001 (2011).

46. Tinti, G. *et al.* Electron crystallography with the EIGER detector. *IUCrJ* **5,** 190–199. doi:10.1107/s2052252518000945 (2018).

47. Pitters, F. M. *et al.* Time and Energy Calibration of Timepix3 Assemblies with Thin Silicon Sensors. http://cds.cern.ch/record/2649493 (2018).

48. Khalil, M., Dreier, E. S., Kehres, J., Jakubek, J. & Olsen, U. L. Subpixel resolution in CdTe Timepix3 pixel detectors. *Journal of Synchrotron Radiation* **25,** 1650–1657. doi:10.1107/s1600577518013838 (2018).

49. Chreifi, G., Chen, S., Metskas, L. A., Kaplan, M. & Jensen, G. J. Rapid Tilt-Series Acquisition for Electron Cryotomography. *Journal of Structural Biology* **205,** 163–169. doi:10.1016/j.jsb.2018.12.008 (2019).

50. Van Schayck, J. P. *et al. Sub-pixel accuracy in electron detection using a convolutional neural network* version v1.0.0 (Zenodo, Feb. 2020). doi:10.5281/zenodo.3635923.

51. Van Schayck, J. P., Roussel, L., Fröjdh, E., Schübel, A. & Kraphol, D. *M4I-nanoscopy/geant4medipix* version v1.0.1. Feb. 2020. doi:10.5281/zenodo.3667660.

52. Van Schayck, J. P. *M4I-nanoscopy/tpx3-event-localisation* version v1.1.0. Aug. 2020. doi:10.5281/zenodo.3980701.

53. Van Schayck, J. P. & van Genderen, E. *M4I-nanoscopy/tpx3-tot-correction* version 1.0.0. Feb. 2020. doi:10.5281/zenodo.3641268.

54. Van Schayck, J. P. *M4I-nanoscopy/tpx3HitParser* version 1.0.1. Mar. 2020. doi:10.5281/zenodo.3693995.

55. Van Schayck, J. P. *M4I-nanoscopy/tpx3EventViewer* version v1.0.3. Mar. 2020. doi:10.5281/zenodo.3693990.

Appendix A Supplemental Information

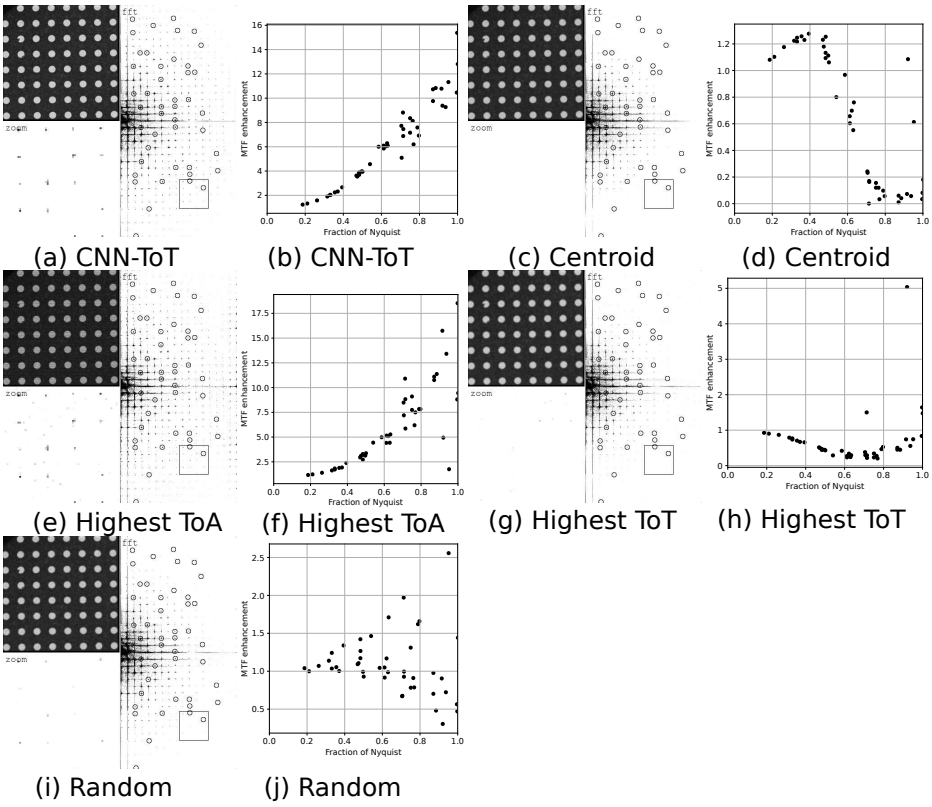


Figure 2.10: Images of UltrAuFoil 200 mesh grids at 225 times magnification at 200 kV with their corresponding power spectra. The rectangle box in the power spectrum is shown zoomed in on the bottom left of the image. The spots which are encircled in the images are used for the calculation of the MTF enhancement graph.

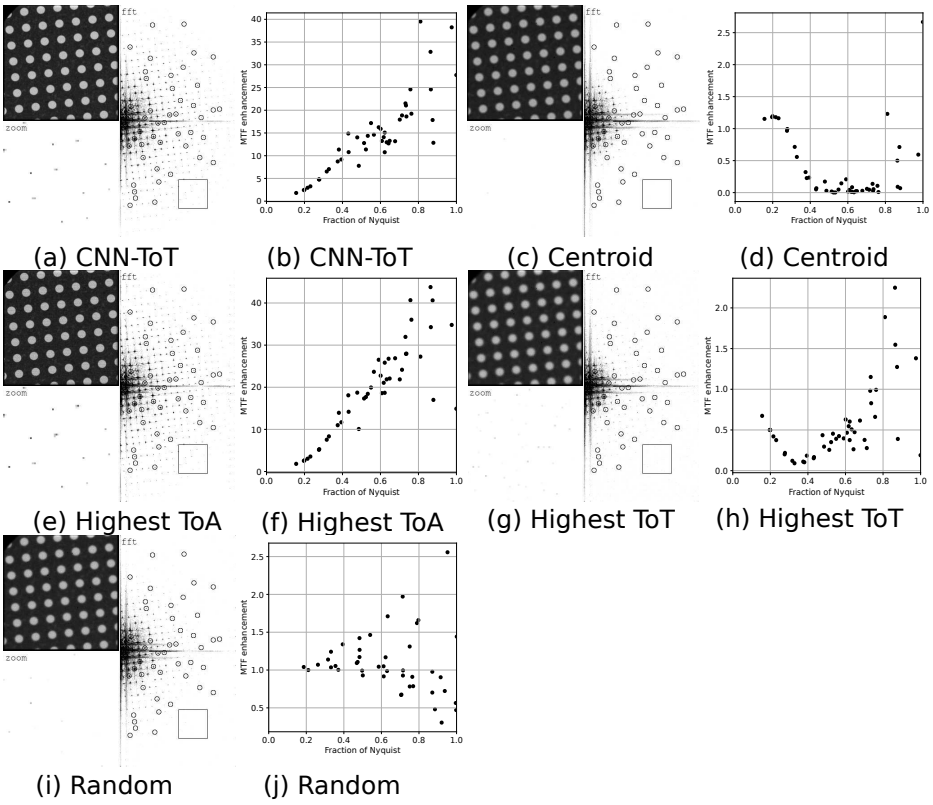


Figure 2.11: Images of UltrAuFoil 200 mesh grids at 225 times magnification at 300 kV with their corresponding power spectra. The rectangle box in the power spectrum is shown zoomed in on the bottom left of the image. The spots which are encircled in the images are used for the calculation of the MTF enhancement graph.

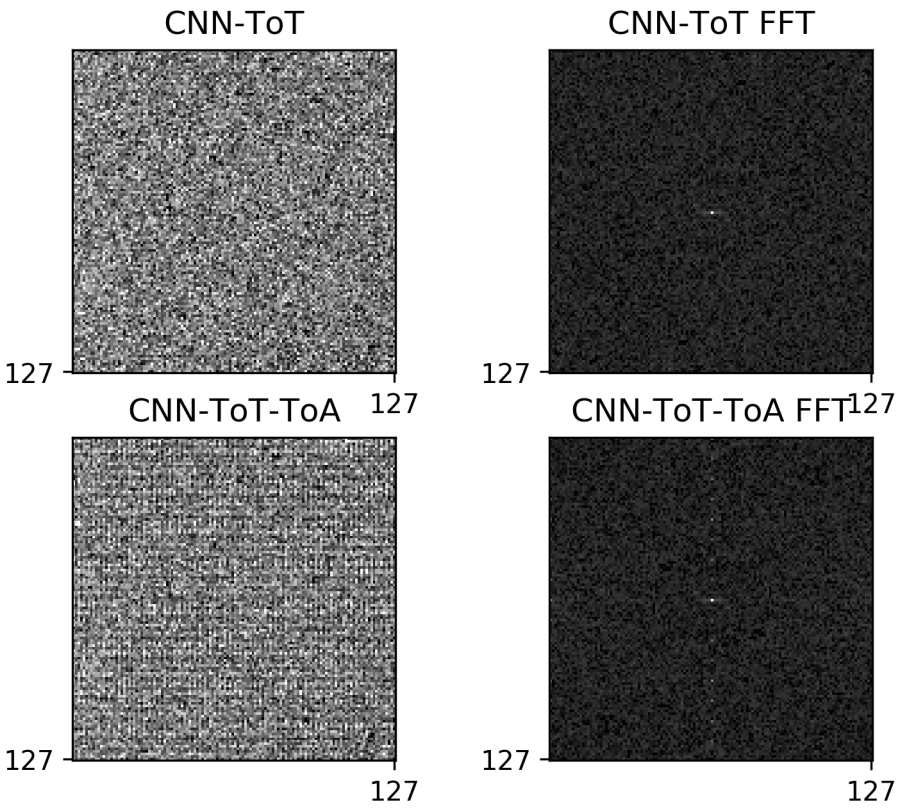
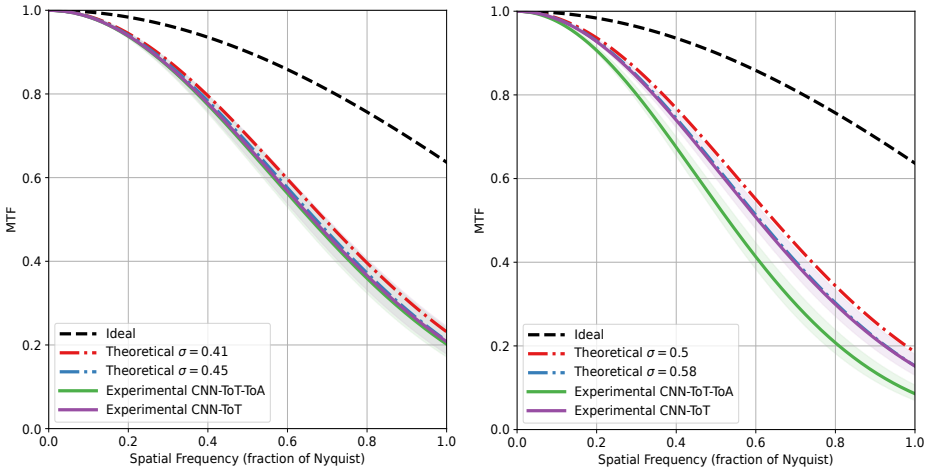


Figure 2.12: Images (**A and C**) and their corresponding power spectra (**B and D**) of a flat beam image after event localisation using the CNN-ToT method (**A and B**) or the CNN-ToT-ToA method (**C and D**).



(a) 200 keV (b) 300 keV

Figure 2.13: The experimentally obtained and simulated modulation transfer function (MTF) at 200 kV (**a**) and 300 kV (**b**) of CNN-ToT and CNN-ToT-ToA. Simulated MTF curves were calculated by fitting the point spread function from the data underlying Figure 2.4 to find λ (Eq. 10 and Eq. 14 of [7]). The shading of the experimental curve represents the 3-sigma confidence interval.

3

Integration of an event-driven Timepix3 hybrid pixel detector into a cryo-EM workflow

Adapted from: **Van Schayck*, J. P**, Zhang*, Y., Knoops, K., Peters, P. J. & Ravelli, R. B. G. Integration of an event-driven Timepix3 hybrid pixel detector into a cryo-EM workflow. *Microscopy and Microanalysis* **29,** 352–363. doi:10.1093/micmic/ozac009 (2022).
* both authors contributed equally

Abstract

The development of direct electron detectors has played a key role in low-dose electron microscopy imaging applications. Monolithic active-pixel sensor (MAPS) detectors are currently widely applied for cryo-electron microscopy (cryo-EM); however, they have best performance at 300 kV, have relatively low read-out speed and only work in imaging mode. Hybrid pixel detectors (HPDs) can operate at any energy, have a higher DQE at lower voltage, have unprecedented high time resolution, and can operate in both imaging and diffraction modes. This could make them well-suited for novel low-dose life-science applications, such as cryo-ptychography, iDPC, and liquid cell imaging. Timepix3 is not frame-based, but truly event-based, and can record individual hits with 1.56 ns time resolution. Here, we present the integration of such a detector into a cryo-EM workflow and demonstrate that it can be used for automated data collection on biological specimens. The performance of the detector in terms of MTF and DQE has been investigated at 200 kV and we studied the effect of deterministic blur. We describe a single-particle analysis structure of 3 Å resolution and compare it with Falcon3 data collected using the same microscope. These studies could pave the way towards more efficient dose-efficient single-particle techniques.

3.1 Introduction

Cryo-electron microscopy (cryo-EM) of biological samples depends on the recording of a small number of incident electrons that can be used before the aqueous sample is destroyed by radiation damage [1]. The useful available signal in the recorded image is low and only barely above the shot and background noise. Therefore, subsequent processing of the data is highly dependent on having the best signal-to-noise ratio (SNR) throughout the process. There are many factors that determine the overall SNR in cryo-EM. These can be related to sample preparation, e.g. sample thickness, or related to the microscope, such as spatial and temporal coherence of the electron source and aberration of lenses [2]. The chosen microscope workflow also plays a very important role. The available signal will propagate differently and will have different noise influences, depending on the chosen imaging or diffraction technique [3]. However, no matter the workflow, microscope, or sample, nearly all cryo-EM workflows will use a pixelated detector to record and digitise the signal generated by the microscope.

Ideally, the pixelated detector faithfully records the entire signal generated by the microscope and does not add any additional noise to the signal. The ratio SNR^2_{out} to SNR^2_{in} is called the detective quantum efficiency (DQE) of the detector [4]. Because pixelated detectors are inherently limited by their finite pixel size, their DQE is limited to the sinc function $sinc^2(\pi\omega/2)$, as a function of spatial resolution ω. Practically, DQE can be calculated by measuring the modulation transfer function (MTF) and the noise power spectrum (NPS or Wiener spectrum) and is scaled by the detection efficiency DQE(0) [5–8]. The normalised NPS (NNPS) is obtained after normalising NPS with NPS(0).

$$DQE(\omega) = DQE(0)\frac{MTF(\omega)^2}{NNPS(\omega)} \tag{3.1}$$

The emergence of direct electron detectors in the form of monolithic active-pixel sensors (MAPS) has been a breakthrough in cryo-EM [9, 10]. Three developments surrounding MAPS detectors were pivotal in this breakthrough. First, the MTF of the detectors was significantly improved by backthinning the sensor layer to reduce the chance that electrons scatter into adjacent pixels [11]. Second, the MTF was further improved by computationally counting single electron events, at the cost of limiting the electron flux below a coincidence level [12, 13]. These two developments made MAPS detectors the first cryo-EM detectors to have significantly better SNR performance compared to film-based solutions. The third development was the ability of MAPS detectors to collect multiple frames per second (a movie), to correct for beam-induced motion and sample drift within a single recording [14]. Combined with advances in microscope automation and

processing algorithms, MAPS detectors have enabled structural biologists to reveal macromolecular structures at near-atomic resolution using cryo-EM single particle analysis (SPA) [15, 16].

However, the use of MAPS detectors is very restricted and is basically limited to this type of cryo-EM workflow. Due to limits in counting algorithms and radiation damage, they cannot be used in experiments with high electron flux or condensed direct beam, such as electron diffraction and ptychography. Although considerable improvements have been made, MAPS detectors have relatively low frame rates (< 1 kHz). Both their radiation fragility and low frame rates can primarily be attributed to their thin backthinned sensor layers with little room for additional electronics or shielding. Furthermore, current MAPS detectors operate optimally at 300 kV; their MTF and DQE(0) deteriorate at energies below 300 kV [17]. Lower (i.e 100 kV) voltage may provide better SNR due to the more favourable elastic/inelastic scattering rations [2, 18]. In addition, lower voltage microscopes might improve access to cryo-EM, due to their lower acquisition and maintenance costs [19]. Therefore, there is interest in developing detectors that can support more versatile workflows and work on a wider range of voltage levels.

Hybrid pixel detectors (HPDs) have many properties that can overcome some of these limitations. By definition, HPDs have separate sensor and electronic layers [20]. This makes them radiation-hard by design and allows additional electronics for much higher read-out speeds [21]. On the other hand, this also comes with two important drawbacks, which so far have limited their practical use in many cryo-EM workflows. (1) Hybrid pixel detectors have a relatively large pixel size compared to MAPS detectors (i.e. 55 μm for Medipix HPD family, compared to less than 14 μm for MAPS detectors). This limits, in practise, the maximum field of view of HPDs in any kind of imaging application. (2) Electrons above 80 kV will scatter beyond a single (55 μm) pixel. In effect, the MTF of the HPD at energies above 80 kV deteriorates rapidly [22]. Overcoming these two important HPD drawbacks would make great strides towards HPD applicability in more cryo-EM workflows.

Two different methods have previously been shown to limit the effect of electron charge sharing between pixels in HPDs. The first method is to use a semiconductor material with a higher atomic number. Such a high-Z material reduces the distance from which the electron can scatter and thereby eliminates charge sharing. This is a promising and elegant method, but it comes with additional challenges, such as sensor cooling and crystal imperfections of the high-Z semiconductor material used as the sensor layer [8]. High-Z materials will also lead to a higher number of undetected electrons, thus damping DQE(0) due to increased backscattering compared to a silicon semiconductor. The second method is to electronically assign the sum of the charge of four surrounding pixels to

the one pixel that has received the highest charge within that group of pixels. This charge sharing mode (CSM) is present in the Medipix3 HPD and has been used to demonstrate near ideal MTF and DQE up to 80 kV electrons [8, 23, 24]. However, above 80 kV the charge spreads beyond the four-pixel group, which limits the benefits of CSM. The scattering of electrons through the sensor layer is erratic. They tend to shed most of their energy far away from the impact pixel, which limits the effectiveness of CSM, even if it could operate beyond a four-pixel cluster. Therefore, any sort of computational effort to reduce the effect of charge sharing needs to be more veracious.

In previous work, we have shown how the properties of the Timepix3 HPD can be used to count and localise individual 200 kV or 300 kV electrons to a sub-pixel location (Chapter 2). The Timepix3 is capable of simultaneously measuring the time a signal is over the comparator threshold (time over threshold; ToT) and the timing of when the signal crosses the comparator threshold (time of arrival; ToA) [25]. Previously, we explored these spectroscopic per-pixel properties of the Timepix3 and trained a neural network on simulated Timepix3 data to localise individual electrons from a pixel cluster. This negates the charge-sharing effect and by doubling the pixel matrix dimensions in a super-resolution mode mitigates the downsides of the relatively large physical pixel size. Here, we show how this detector can be incorporated in an actual cryo-EM workflow for a life sciences sample, and demonstrate its performance at 200 kV, an energy at which we could compare it with the commercial Falcon3 detector.

Cryo-EM workflows require a deep integration of the microscope and detector, both on the hardware and software level. First, the operation of the detector should not negatively affect the microscope itself. A detector comes with cooling and electronics: possible vibrations or electromagnetic effects generated by these should not compromise the quality of the data the microscope could bring. There is also integration at the software level. Often cryo-EM workflows run for hours, up to multiple days, in an automated fashion, where data collection decisions are made based on the output of the detector. These can be adjustments to the lenses to set focus, correct for aberrations, or verify the movement of the sample stage to the next acquisition location. Due to the radiation sensitivity of cryo-EM samples, the blanking and unblanking of the electron beam also needs to be tightly synchronised with the detector. Turning on the beam, before the detector is enabled, would lead to the loss of precious signal. Different methods can be thought of to reach this level of hard and software integration [26]. Here, we present and explain the choices we made to obtain optimal performance of our Timepix3 setup for a cryo-EM workflow. We have chosen SPA as the cryo-EM workflow for this, being the most widely used by the community right now.

Beyond software and hardware integration, detector data need to be prepared in such a way that downstream processing software can handle it. The Timepix3

is an event-driven detector and outputs a sparse stream of data with, per electron, a cluster of hits containing position, time and energy. After event localisation, this sparse stream is reduced to a single event containing position and time for each electron. In the case of an SPA workflow, the processing software is not yet equipped to deal with this sort of sparse data. With Falcon4 and the EER format, the first steps towards event-driven readout for MAPS detectors have been made [27, 28]. However, in its current form, EER is still based on frame-driven detectors, and it is not able to take advantage of the precise nanosecond timing of Timepix3. Therefore, we still had to convert our event stream to a classical movie of fractions (frames). These fractions must be gain-corrected and free from large defects, such as those caused by the edges of the chip. We describe how we have optimised the placement of these events in the image frames and how we have corrected for large defects.

This work describes how we have integrated a Timepix3 HPD into an automated cryo-EM SPA workflow. We measured MTF, NNPS, and DQE, and compared the results obtained with the commercial Falcon3 detector from the same protein and under the same microscope. We discuss how Timepix3 and future HPDs such as Timepix4 could enable the development of new versatile cryo-EM workflows, which could allow one to obtain better and complementary structural information within the limited dose lifetime of aqueous life sciences samples.

3.2 Experimental setup

3.2.1 Hardware integration

The detector pod containing the Timepix3 HPD was mounted on a 200 kV Tecnai Arctica (Thermo Fischer Scientific) as previously described (Chapter 2). The detector pod (now commercially available from Amsterdam Scientific Instruments as CheeTah T3) provides the vacuum housing, radiation shielding, cooling, mechanical shutter, and houses the readout hardware. The SPIDR readout board developed by the Dutch National Institute for Subatomic Physics (Nikhef) is connected to the detector PC via a $10\,\mathrm{Gbit\,s^{-1}}$ fibre optic connection [29].

The detector pod cooling lines were initially connected to a dedicated chiller system (SMC HECR002-A5-FP), which cools the refrigerant to 18 °C. As refrigerant, water mixed with ThermoClean DC (Bioanalytic GmbH) was used to prevent algae growth in the refrigerant. The chiller was located a few metres from the microscope and had a reported, non-adjustable, flow rate of $120\,\mathrm{L\,h^{-1}}$. Both the proximity and the relatively high flow rates of the dedicated chiller were a source of concern, as they may interfere with the operation of the microscope. Fortunately, we also had the possibility of connecting the Timepix3 to the existing

central chiller used by the rest of the microscope, enabling us to measure possible interferences of different cooling regimes on the data quality.

We collected a series of high-resolution micrographs of a Crossed Line Grating Replica (EMS diasum, #80051). The micrographs were recorded using Falcon3 in electron counting mode, at a nominal magnification of 215 000 times, 40 s exposure, with a total fluence of 40 e^{-} $Å^{-2}$. Movie stacks were corrected for drift (rigid body) using Relion motion correction [30]. Power spectra were calculated from the resulting sum. An area of the 2.35 Å gold diffraction ring was selected and the mean pixel value in this area was calculated. For each condition, a series of 30 power spectra was recorded with 5-minute intervals. The conditions tested were, in order: dedicated chiller, dedicated chiller when not connected to the detector pod, central chiller at 5 L h^{-1}, and central chiller at 25 L h^{-1}. As a control, between different conditions, all Timepix3 equipment was turned off. The results are shown in Figure 3.1. They indicate that there was interference from both the refrigerant that passed through the detector and the chiller itself. This interference could have been electromagnetic or due to vibrations. From these results we concluded that we should cool our Timepix3 detector using the central microscope chiller at 5 L h^{-1}. During such an operation, the temperature of Timepix3 would only increased marginally from 42 °C to 47 °C.

3.2.2 Software integration

The Timepix3 was integrated with SerialEM [31] using a custom camera plugin. This plugin provided communication between SerialEM and Serval, a network service developed by Amsterdam Scientific Instruments (ASI) which facilitates remote procedure calls between the detector PC and the Timepix3 SPIDR readout board. Figure 3.2A shows the flow diagram schematically. Through the plugin and Serval, SerialEM is able to set the exposure time and can toggle the storage of raw, unprocessed Timepix3 data. In return, the plugin receives unprocessed preview images. These preview images are gain-corrected by SerialEM and used for montage, navigation, or autofocus procedures. Raw unprocessed Timepix3 data can be stored separately on the detector PC and transferred to network storage to be processed on individual workstations.

The synchronisation of the beam blanker and the camera was investigated by unmasking a noisy Timepix3 pixel. This particular noisy pixel blinked at 100 kHz regardless of any actual events. The ToA stamps of this pixel could be used to time the start and end of acquisition. Initially, control of the blanking and unblanking of the beam was performed by software within SerialEM. SerialEM would attempt, after calibration, to time the unblanking of the beam to coincide with the detector starting the acquisition. Because PCs are, without special adaptation, unreliable

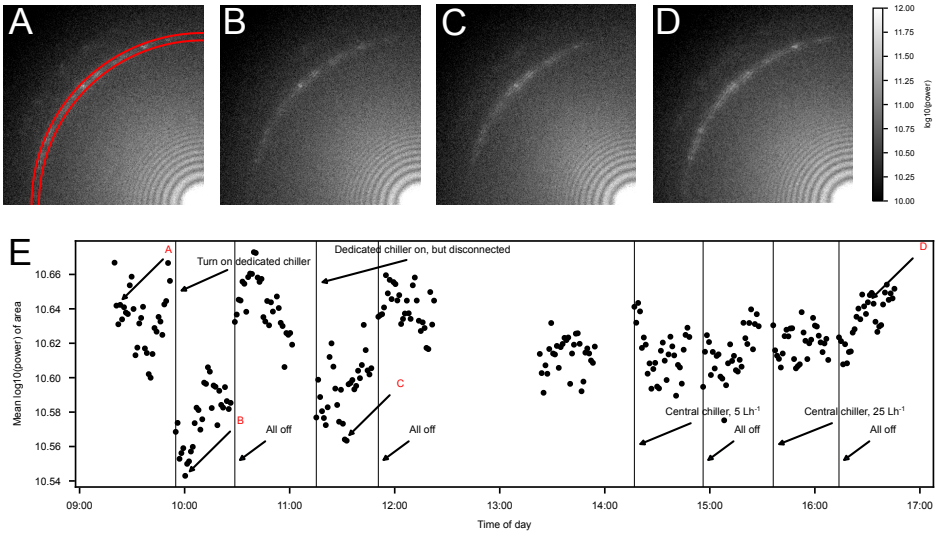


Figure 3.1: Effect of different cooling and controlling regimes of Timepix3 on the intensity of a 2.35 Å gold diffraction ring (marked red in **A**). (**A-D**) show corners of the power spectrum of a motion corrected Falcon3 image recorded in electron counting mode, measured with dedicated chiller off (**A**), dedicated chiller on (**B**), dedicated chiller on but disconnected to the microscope (**B**), and dedicated chiller off (**D**) like in (**A**) but 8 hours later. (**E**) shows the mean intensity of the region marked in red in (**A**). This demonstrates that the dedicated Timepix3 chiller had a detrimental effect on data quality.

to do microsecond timings this was later changed to a direct electronic trigger for more accurate control. To do this, the SPIDR readout board had to be configured to output a 3.3 V LVTTL signal on its HDMI channel whenever the Timepix3 electronic shutter is open. This signal was inverted (low when the Timepix3 shutter was open, high when it was closed) and connected to the microscope beam-blanking switchboard. Thermo Fisher Scientific had to allow a standalone camera to control the beam-blanker, which a service engineer could do by ticking an option within the Tecnai control software setup (Figure 3.2C shows the now-enabled option). This direct electronic trigger of the beam blanker reduced the delay between detector acquisition start and beam unblanking to a very reliable 10 µs to 20 µs.

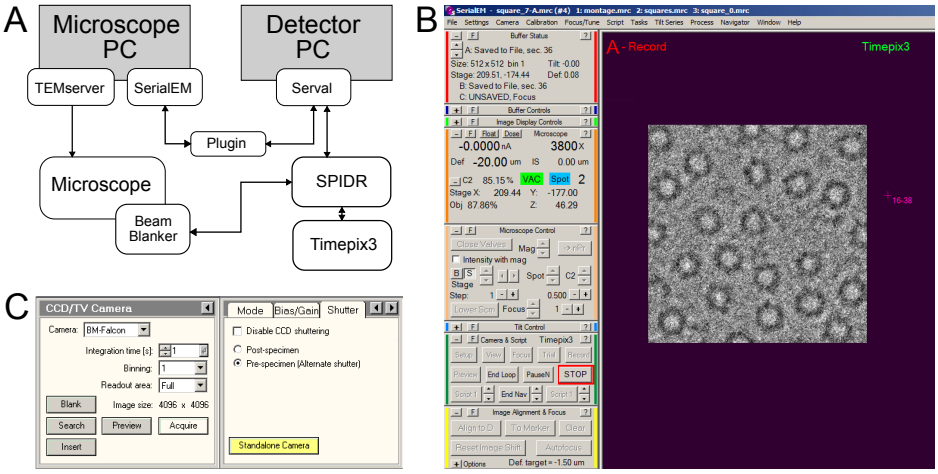


Figure 3.2: (**A**) Flow diagram of how the Timepix3 is integrated in the microscope. (**B**) Computer screen shot of SerialEM controlling the Timepix3 through the custom plugin. (**C**) Computer screen shot of the Tecnai control software showing the standalone camera control button. Note that the Falcon3 MAPS detector is retracted at the same time.

3.2.3 Electron event localisation

The processing of Timepix3 data into micrographs is largely similar to what we described previously (Chapter 2). Briefly, raw data are read from its binary Timepix3 format and ToA information is recalculated to one uniform clock. Some known defect pixel columns experience a phase-shifted ToA clock, which is corrected for. Next, non-uniformities in the ToT response are corrected for using previously obtained calibration data. Individual hits are clustered on the basis of being direct neighbours within a 50 ns ToA interval. Clusters are filtered according to their total sum of ToT values and the number of pixels affected. The filtered clusters are then passed to the convolutional neural network, which calculates a sub-pixel incident position based on the shape of the cluster and the ToT values. An overview of all individual steps is shown in green blocks in Figure 3.3A.

To improve processing performance, two improvements have been made over the previous work. First, a new custom clustering algorithm was written. This algorithm works by first sorting hits in time, and then searching recursively for neighbouring pixels within a time limit. For extra performance, this algorithm was written in Rust and compiled as a standalone library to be called from Python. Second, a rewrite of the processing code was made to perform all processing steps

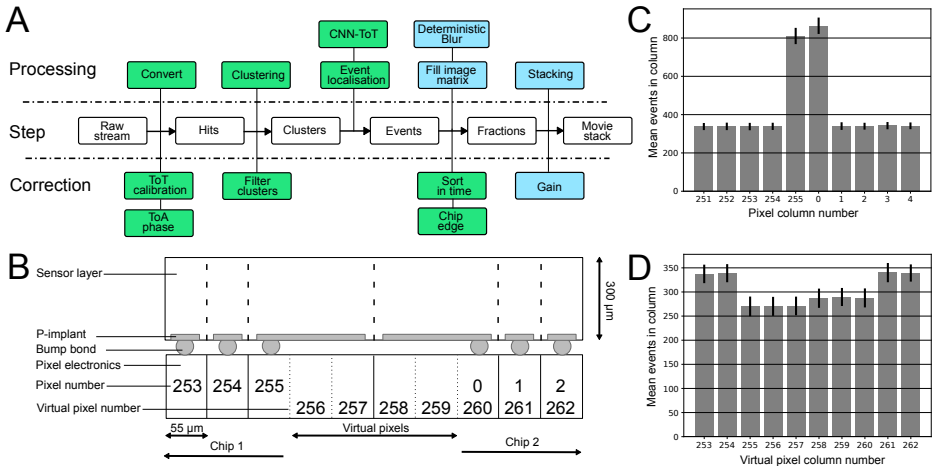


Figure 3.3: (**A**) Flow diagram describing how Timepix3 data is processed from raw data to a movie stack. The green blocks show processing steps performed in tpx3HitParser [32]. The blue blocks represent processing steps performed in tpx3EventViewer [33]. (**B**) Schematic representation of edge pixels. (**C** and **D**) Mean values for the number of events in pixels columns between two chip edges, before (**C**) and after edge correction (**D**).

on chunks of data in parallel. Together, this improved processing performance by more than 10-fold from 80 khit s^{-1} to 1 Mhit s^{-1} on the same hardware. A typical SPA exposure (120 Mhit, 50 e^{-} $Å^{-2}$, 1.25 Å pixel size, 700 MB) is processed in 2 min from raw data to sorted events.

Finally, we adjusted the cluster filter window. In our previous work, we had limited the filtering of the clusters (Figure 2.2) to the main peak, thus excluding events such as those hitting edge pixels. The improved algorithms can make better use of these edge pixels. Therefore, the adjusted clustering window will now accept them, resulting in more events being used.

3.2.4 Generation and correction of image frames

From the sparse, event-localised data stream consisting of the position and time of each electron, individual image fractions (frames) are reconstructed over several steps. The complete processing workflow, after event localisation, is shown in blue blocks in Figure 3.3A.

Our Timepix3 setup is in a quad configuration with four individual chips tiled together. Each chip has 256 pixels in each dimension, giving a total of 512 × 512 pixels. The chips are tiled together with a gap of 220 µm, the width of four pixels,

between them. This means that the pixels at the inner edges of each chip represent an area that is triple the width (165 μm) of a normal pixel (Figure 3.3B). Due to their larger size, they also receive more events (Figure 3.3C), a bit less than three times more (Figure 3.3D). To correct for this effect, events in the edge pixels are randomly redistributed to two virtual pixels created next to the original edge pixel. This results in an image frame four pixels larger in both dimensions (516x516). The remaining difference in pixel response between the three edge pixels and the rest of the chip can be compensated for via gain correction.

We split the event stream on the basis of their arrival time to generate fractions of arbitrary exposure time. During raw data processing, the fine ToA clock, the ToA, and the SPIDR global timer are combined to produce one uniform clock. All events are sorted on the basis of this uniform clock, and chunks of events are split out on the basis of the desired per-fraction exposure time.

To form fractions and movies, localised events are placed in an image matrix in several different ways. The convolutional neural network returns, during event localisation, floating precision numbers which can be placed at, practically, any level of sub-pixel accuracy. The first and most naive method is to take the integer of the floating number and place the event in the corresponding pixel (Figure 3.4B_1). Similarly, by upscaling the image matrix by a factor, it was possible to create super-resolution images at this factor. The second method is to apply a deterministic blur to the event, by upscaling the image matrix N times and then applying a Gaussian convolution of the upscaled image with variance $N\sigma^2$. Next, the image matrix is Fourier-cropped to the desired final resolution (Figure 3.4C_1 and D_1). We used $N = 10$ and determined that the optimal σ is 0.5 by calculating all the DQE curves for $\sigma = \{0.1, 0.2, ..., 1.0\}$ (Figure 3.4F).

For comparison, the raw Timepix3 data are also processed without clustering and event localisation. Instead, all individual hits are integrated per pixel (Figure 3.4A_1, integrated hits) without further considering their ToT or ToA values. From there, these data are processed in a similar fashion as described above to correct for edge pixels and are sorted and split on the basis of their arrival time.

Finally, the fractions are gain corrected by multiplying with a normalised average image (gain) calculated from a series of flat-field exposures. Gain images are clipped to a maximum of five-fold difference to the median response. The gain was calculated for each method separately. The fractions were individually gain-corrected and placed together in one MRC movie file stack.

3.2.5 Measuring MTF, NNPS and DQE

The MTF was measured using the so-called knife-edge method as previously described [6, 8]. Briefly, from an image of the knife edge, the angle and intercept of the edge were precisely determined using thresholding and a Sobel filter. The

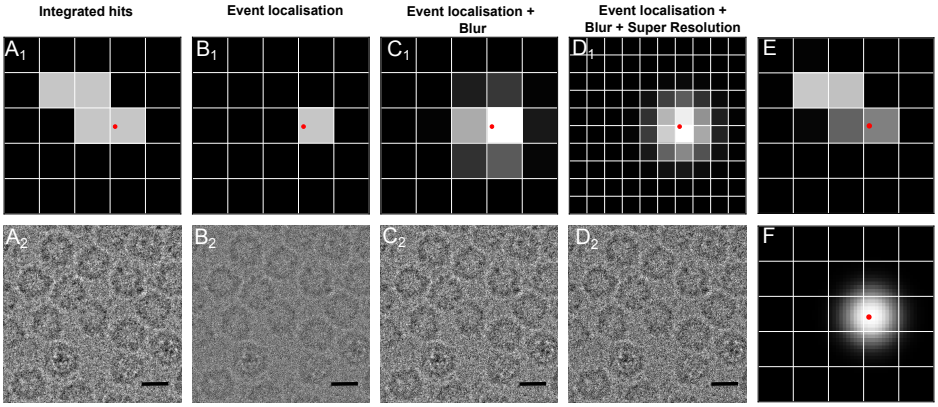


Figure 3.4: (**A_1-D_1**) Different methods used to generate a pixel cluster representing the same single 200 kV electron event. The red dot denotes the predicted incident location of the electron (CNN-ToT). A_1-C_1 are 5x5 pixel matrices. D_1 is a 10x10 super resolution pixel matrix. Brighter pixels are assigned more arbitrary counts. (**A_2-D_2**) Different methods used to form an example micrograph of the BfrB protein from the same raw Timepix3 data. Scale bar represents 10 nm. Micrographs have been gain and motion corrected. (**E**) The input pixel cluster used for the CNN-ToT prediction of the incident location. Brighter pixels received more energy (Timepix3 ToT counter). (**F**) The applied deterministic blur (Gaussian, $10\sigma^2$, $\sigma = 0.5$) of the same single electron event. The Gaussian is calculated in a 100x100 pixel matrix. This is used as precursor step to generate C and D, where the cluster has been Fourier cropped to the desired resolution.

distance to the edge was then calculated for each pixel. From this an edge spread function could be fitted and the MTF calculated [6, Eqs. 12 and 13]. In our setup, the edge image was made using a mechanical aluminium shutter placed directly above the Timepix3 detector. The shutter was placed at a 7° angle from the pixel matrix. For Falcon3, an edge image was made using the microscope beam stopper, which was at a 14° angle from the pixel matrix. For both detectors, a sufficiently straight edge was chosen without defects of the shutter or beam stopper.

NNPS was measured as previously described [6, 8, 34]. Briefly, a series of flat-field images were recorded. For Timepix3, this was done by taking a single exposure and splitting it into multiple fractions. For Falcon3, this was done by taking a single exposure in multiple fractions (a movie). The mean image of the series was calculated and subtracted from each frame in the series. The power spectrum then gives the 2D NPS, and its radial average is the 1D NPS [8, Eq 2.]. NPS(0) was estimated by progressively Fourier-cropping the series of images by larger factors b. The variance of the images σ^2, normalised by the square of the

binning factor, was evaluated. As b increases, $\sigma^2 b^{-2}$ reaches a plateau that was taken as NPS(0). The plateau was estimated by fitting with a logistic function.

From MTF and normalised NPS (NNPS), DQE can be calculated if the scaling factor DQE(0) is known (Eq. 3.1). DQE(0) can only be calculated from a known gain factor [8, Eq. 4] which, in turn, requires precise knowledge of the beam current. For low-beam currents, this is typically done using a Faraday cup. Our experimental setup did not allow for the positioning of a Faraday cup, and hence we were unable to directly measure DQE(0). Figure 11 from [8] shows the DQE(0) for a Medipix3 HPD at 200 kV in single-pixel mode (0.8) and in charge-sharing mode (0.9). Taking into account that the minimum operating threshold for Medipix3 is 40% higher than Timepix3 (700 e^- vs 500 e^-), there is no indication Timepix3 DQE(0) at 200 kV would be lower than 0.9. Therefore, we made the safe assumption that DQE(0) at 200 kV for Timepix3 is at least 0.9.

We are unaware of a published DQE(0) for a Falcon3 operated at 200 kV in electron counting mode. In [35] Falcon3 DQE(0) in integration mode at 200 kV and 300 kV is reported to be, respectively, 0.35 and 0.50. In electron counting mode, DQE(0) at 300 kV is reported to be 0.85. Therefore, we made the safe assumption that DQE(0) at 200 kV Falcon3 in electron counting mode is at most 0.85.

3.2.6 Single-particle analysis workflow

Protein samples of *Mycobacterium tuberculosis* apoferritin (bacterioferritin B, BfrB) were prepared as previously described [36]. Purified BfrB was used at a concentration of 40 mg mL^{-1}. A volume of 2.5 µL was applied onto glow-discharged UltrAuFoil Au300 R1.2/1.3 grids (Quantifoil). The excess liquid was removed by blotting for 3 s (blot force 5) using filter paper followed by plunge freezing in liquid ethane using a FEI Vitrobot Mark IV operated at 100% humidity at 4 °C.

Timepix3 single-particle data were collected at 200 kV using SerialEM. SerialEM was directed using scripts adapted from single-particle scripts from the SerialEM Script Repository[1]. The Timepix3 required some specific settings, in particular to account for the relatively small field of view of the camera. The tilting angle of the beam used during autofocus had to be tuned and a relatively low magnification had to be used for the hole-centring alignment acquisition such that the entire hole of the grid could be obtained. SerialEM-directed gain correction allowed for correction of the chip edges without accounting for the physical size difference of these pixels. One image per hole was acquired using stage movements to navigate between holes. We waited a minimum of 10 s to allow the beam and stage to settle before each data acquisition. Data could be collected autonomously, for

[1] `https://serialemscripts.nexperion.net/`

up to 72 h in succession, only interrupted by autofilling of the microscope liquid nitrogen dewars.

Table 3.1 shows the statistics of the datasets. The same Timepix3 data were processed in four different datasets. Dataset one serves as a control: Timepix3 hits data were integrated without clustering or event localisation. For dataset two, the Timepix3 data were event-localised using CNN-TOT. The third Timepix3 dataset was based on event-localisation using CNN-ToT and deterministic blurring. Finally, the fourth dataset is based on event localisation using CNN-ToT, deterministic blurring, at two times super resolution. All four data sets were treated equally from then on. Data were processed using the RELION pipeline [37]. The movie stacks were corrected for drift (single patch) and dose weighted using MotionCor2 [38]. The Timepix3 sample pixel size was measured using a cross-grating grid at low magnification, and initially extrapolated by SerialEM to the recording magnification. Finer pixel size calibration was achieved by aligning the obtained map with an existing model (PDB:7O6E).

Falcon3 single-particle data were collected from a different sample grid under the same microscope. Data collection was done using EPU. The detector was used in electron counting mode at a nominal magnification of 155,000 times. Table 3.1 shows the statistics of this dataset. Data were processed using the RELION pipeline [37]. The movie stacks were corrected for drift (5×5 patches) and dose weighted using MotionCor2 [38].

From this point on, the Timepix3 and Falcon3 datasets were treated equally. The contrast transfer function (CTF) parameters were determined for drift-corrected micrographs using Gctf [39]. A first set of 2D references was generated from manually picked particles in RELION [37] and these were then used for subsequent automatic particle picking. Table 3.1 lists the number of particles in the final dataset initially picked as well as the number of particles that contributed to the final reconstruction. An initial map was generated by RELION with octahedral symmetry (O) and used for 3D refinement. The beam tilt parameters, anisotropic magnification and local CTF parameters were refined and the particles were polished [30]. The final map was calculated and sharpened using the polished particles (Table 3.1).

For model building, the PDB entry 7O6E [36] was used as a starting model in Coot [40]. The final model was refined against a sharpened cryo-EM map obtained by LocSpiral [41]. The model was refined iteratively through rounds of manual adjustment in Coot [42], real-space refinement in Phenix [43], redone using PDB-REDO [44], and validated using MolProbity [45] (Supplementary Table 3.2).

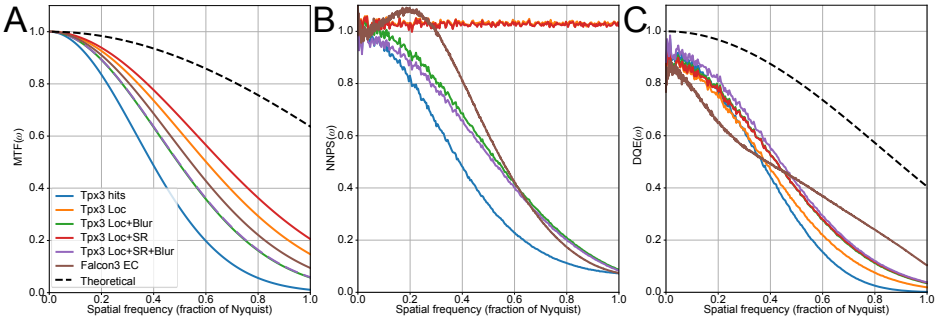


Figure 3.5: MTF (**A**), NNPS (**B**) and DQE (**C**) as function of spatial resolution (as fraction of the Nyquist frequency) for Timepix3 integrated hits data (Tpx3 hits, blue), event-localised Timepix3 (Tpx3 Loc, orange), event-localised Timepix3 with deterministic blur (Tpx3 Loc+Blur, green), event-localised Timepix3 at two times super-resolution (Tpx3 Loc+SR, red), event-localised Timepix3 with deterministic blur at two times super-resolution (Tpx3 Loc+SR+Blur, purple) and Falcon3 in electron counting mode (Falcon3 EC, brown). DQE(0) for Timepix3 is assumed to be 0.9. DQE(0) for Falcon3 electron counting is assumed to be 0.85. To aid in comparison, the spatial frequency of super resolution data has been rescaled to the recorded spatial frequency. The theoretical MTF curve is $sinc(\pi\omega/2)$ and the theoretical DQE curve is $sinc^2(\pi\omega/2)$.

3.3 Results

3.3.1 MTF, NNPS and DQE

Figure 3.5 shows the MTF, NNPS, and DQE obtained at 200 kV using Timepix3 integrated hits data (Tpx3 hits, blue), event-localised Timepix3 (Tpx3 Loc, orange), event-localised Timepix3 with deterministic blur (Tpx3 Loc+Blur, green), event-localised Timepix3 at two times super-resolution (Tpx3 Loc+SR, red), event-localised Timepix3 with deterministic blur at two times super-resolution (Tpx3 Loc+SR+Blur, purple) and Falcon3 in electron counting mode (Falcon3 EC, brown). Integrating just hits results in an inferior performance of the Timepix3 detector, both in terms of MTF and NNPS, leading to the lowest DQE. Event-localised Timepix3 and event-localised Timepix3 super-resolution have the highest MTF. Their flat NNPS make their DQE appearance intuitive: the improved MTF for event-localisation when comparing original with super-resolution, directly results in a better DQE for event-localisation super-resolution compared to original resolution. The effect of deterministic blur is somewhat less intuitive, as it will dampen both the MTF and NNPS. Overall, the best DQE results for the Timepix3 data were obtained for event-localisation and deterministic blur at super-resolution,

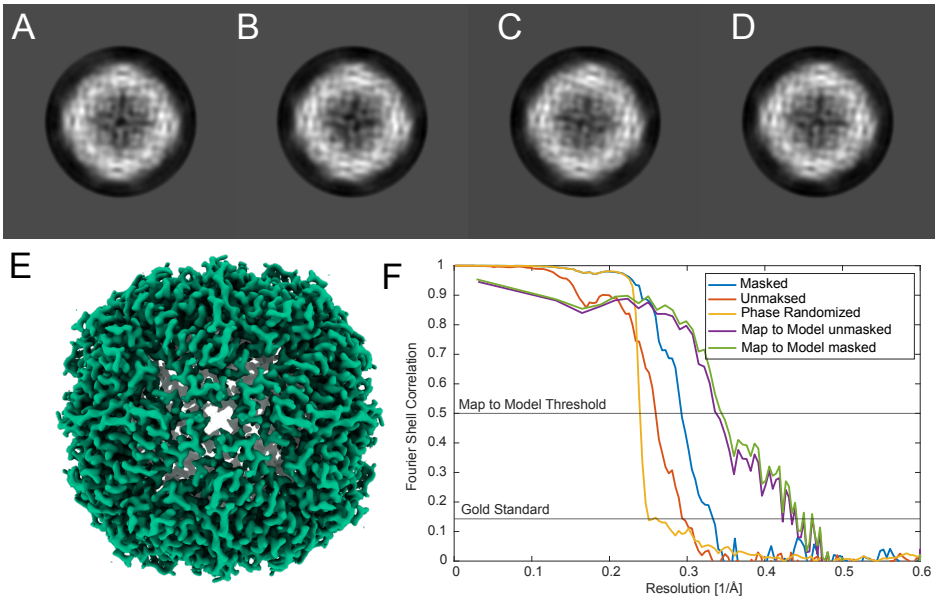


Figure 3.6: Single-particle analysis of BfrB. (**A-D**) 2D class averages; the size of the shown box is 150 Å. (**E**) 3D reconstruction from 11,422 particles at 3.01 Å resolution collected at 200 kV using the Timepix3 with event localisation, deterministic blur and two times super resolution. (**F**) Gold-standard Fourier shell correlation (FSC) before (orange) and after (blue) masking, the phase-randomized FSC (red), and the masked (purple) and unmasked (green) map to model FSC [46].

closely followed by event-localisation and super-resolution. Falcon3 in electron counting mode performed worst until half Nyquist, due to its lower DQE(0), however, hereafter it performed best.

3.3.2 Single particle analysis

Table 3.1 lists the statistics and resolutions obtained for the single-particle analysis data. The resolution of the final map of the event localised and deterministic blurred at two-times super resolution Timepix3 data was 3.01 Å using the gold-standard $FSC = 0.143$ criterion (Figure 3.6). Its sharpening B-factor is -119 Å^2. The Timepix3 dataset was recorded over a 72 h period and ~11,000 usable particles were picked from 2,777 micrographs. The Falcon3 dataset was recorded over a 48 h period and from 979 micrographs 162,975 particles (~14 times more than Timepix3 datesets) were picked. The Falcon3 electron counting dataset resulted

Table 3.1: Cryo-EM SPA data collection, refinement, and validation statistics of Timepix3 and Falcon3 datasets of BfrB protein.

	Timepix3 Integrated Hits	Timepix3 Localised	Timepix3 Localised + Blur[c]	Timepix3 Localised + SR[a]+ Blur[c]	Falcon3 Electron Counting
Magnification	215k	215k	215k	215k	155k
Voltage (kV)	200	200	200	200	200
Flux ($e^- Å^{-2} s^{-1}$)	29	29	29	29	1.76
Exposure time (s)	1.5	1.5	1.5	1.5	30
Frames (no.)	50	50	50	50	48
Defocus range (μm)	0.5 - 1.4	0.5 - 1.4	0.5 - 1.4	0.5 - 1.4	0.5 - 1.8
Pixel size (Å)	1.2	1.2	1.2	1.2/0.6[b]	0.635
Micrographs (no.)	2970	2970	2970	2970	979
Initial particles (no.)	14658	14535	14878	14911	162975
Final particles (no.)	10420	10525	10513	11422	124577
Map resolution (Å)	3.73	3.56	3.17	3.01	2.54
B-factor ($Å^2$)	-183	-132	-132	-119	-93

[a] Two times super resolution.
[b] First value in physical pixel size, second value in pixel size at two times super resolution.
[c] Deterministic blur.

in a resolution of 2.54 Å with a sharpening B-factor of $-93\,Å^2$. The Rosenthal-Henderson B-factor plot is shown in supplementary Figure 3.7.

3.4 Discussion

Our results show a successful integration of Timepix3 as a pixelated detector in a cryo-EM workflow. We were able to integrate this detector both at the hardware and software level. We identified and overcame compromising effects from its original dedicated chiller setup. On the software level, we established a two-way software control of the detector using SerialEM and the remote procedure server Serval from ASI. We arrived at a tightly controlled synchronisation of the electron beam and the Timepix3 shutter event by using SPIDR and a customised trigger box. The successful integration was demonstrated for an SPA workflow, resulting in a 3.01 Å BfrB structure.

Integrating a detector into a TEM can be, as with many engineering endeavours, a project full of idiosyncrasies. Some of which were already mentioned, such as the possible interference from the dedicated chiller, but we also experienced others, such as faulty hardware of the detector PC leading to random crashes

during workflow execution. The Thermo Fisher Scientific TEMs are rather closed platforms: there is very little open documentation available to make hardware and software integration, which led to a considerable amount of trial-and-error work. TEMs can produce vast amounts of data, and making these findable, accessible, interoperable, and reusable is hot topic in the field [47, 48]. We urge microscope vendors to look beyond the data that TEMs produce and to also practise some of these concepts for their hardware and software to allow for better access and interoperability of their platforms such that advances in the field could accelerate.

Our MTF, NNPS and DQE results (Figure 3.5) confirm our previously published observations (Chapter 2): convolutional neural networks can be used to increase the accuracy at which we can predict the impact location of incident electrons. They also show that placing the event singularly in the predicted pixel is not the best way to use this information. Applying a deterministic blur will dampen both MTF and NNPS and can result in an increase in DQE, particularly when the NNPS is aliased [8, 13]. The deterministic blur effect has been modelled for dose-limited detectors used in medical X-ray devices [49–51]. In the field of cryo-EM, the Falcon3 and Falcon4 implement a deterministic blur; however, we are not aware of work describing its fine details [28]. The NNPS of our Falcon3 in electron counting mode (Figure 3.5) has a remarkable shape, significantly deviating from the Gaussian blur functions that we have used. Figure 3.4 shows the comparison of integrated Timepix3 hits data, event localised, and deterministic blurring. Visually, it is hard to see the difference between the integrated hits and the deterministic blurred image: both boost low spatial information, which makes both images visually pleasing. Nonetheless, high spatial-resolution information is much better preserved after event localisation and deterministic blurring (Table 3.1).

Measuring and calculating the MTF, NNPS and DQE to compare cryo-EM detector performance has clear limitations. The measurement and calculation of the MTF is very sensitive to the way the edge spread function is determined. The estimation of NPS(0) is prone to errors, and the measurement of DQE(0) is practically challenging. This echoes the experiences of [52], who reported irreproducible results using the FindDQE programme [7]. Instead of measuring DQE, they examined the quality of their micrographs by using gold-standard Fourier shell coefficient (FSC) curves of their reconstructions. We publish both DQE and FSCs. We have not only openly published our raw data but also the software used to calculate Figure 3.5, and spur other researchers and detector vendors to do the same whenever reporting an MTF, NNPS and DQE [53, 54].

The single-particle analysis of BfrB showed the first cryo-EM workflow integration of a Timepix3 based detector in action. The best Timepix3 results, at 3.01 Å, were obtained using event localisation with deterministic blur at two times super

resolution. The quality of the EM map is illustrated by the clear density of side chains, holes in aromatic rings, and the resolution estimate of 2.9 Å obtained by comparing the map with the refined model. For the most part, the SPA results follow the trend seen in the DQE measurements.

Surprisingly, the effect of super-resolution on the final SPA resolution is relatively minor. There is only a minor improvement in map resolution, and the DQE curves are almost the same. This may indicate that, for these data, the event localisation was not much more precise than 1 pixel. Our previously published results showed that, for simulated data, the point of impact of individual electrons can be found, on average for CNN trained on ToT data (CNN-ToT) at 200 kV, 0.50 pixels away from the incident pixel (Chapter 2). In this work, we have chosen to use CNN-ToT rather than the CNN trained on ToT and ToA data (CNN-ToT-ToA). The latter was only marginally better compared to CNN-ToT, but gave systematic patterns that needed extra (gain) correction. It could also be possible to adapt our event localisation method for the 80-120 kV energy range. The neural network currently benefits from the higher (>= 200 kV) energy electrons generating larger clusters. However, we believe that it should be possible to further improve the event localisation method, also for the 80-120 kV range, for example by correcting for more systematic errors in the ToA data or by training directly on experimental data. The latter could be done with a dedicated setup capable of generating a sub-pixel electron beam at the desired energy.

Timepix3 is the first fully event-driven detector to be used in a cryo-EM workflow. For each electron event a separate timestamp is known, with a maximum accuracy of 1.56 ns. Recently, the electron event representation (EER) format was introduced in conjunction with the Falcon4 MAPS detector [27]. Although the EER representation is an improvement to the scheme in which a set of frames is combined into a fraction of e.g. $1\,e^-\,Å^{-2}$, it is still frame-based. For each time block, events are grouped together. This makes the format impractical for Timepix3 data, where each event has its own timestamp. We could foresee the development of a new, truly event-based data representation and processing pipeline for event-driven detectors such as Timepix3.

The high maximum hit rate ($120\,Mhit\,s^{-1}$) of the Timepix3 (quad) detector allowed us to record the full useful lifetime of a protein in the electron beam ($40\,e^-\,Å^{-2}$) as a stream of individual electron events in just 1.5 seconds. To the best of our knowledge, this is the fastest electron counting camera used for SPA so far. Although the data stream has the promise to allow for very fine corrections that occur during the exposure, such as stage drift, lens drift, and beam-induced motions, best results so far were still obtained while allowing for some settling time between hole selection, autofocus and data acquisition. Data throughput was also still limited by processing speed, although we gained a 10-fold increase

compared to Chapter 2 by improving the clustering algorithm and applying more efficient parallel processing. Further improvements could come from, for example, GPU implementations and the use of software platforms such as LiberTEM [55].

HPDs can be used at much higher electron fluxes in imaging mode or even with direct beams in diffraction mode, e.g. for electron diffraction and ptychography [56, 57]. Such workflows require full integration of the detector in the TEM as well as automation. In this paper, we demonstrate how we could arrive at such integration and automation for SPA. Next, we will explore diffractive imaging techniques, such as ptychography, starting with near-field electron ptychography with a diffuser [58, 59]. Speed is an important practical limitation in cryo-electron ptychography. Our Timepix3 detector should already be much faster than the latest commercially available EMPAD detectors (MAPS detectors cannot be used for ptychography). We are currently building the next generation Timepix4 detector [60], which allows for unprecedented hit-rates ($5\,\mathrm{Ghit\,s^{-1}}$) and time resolution ($200\,\mathrm{ps}$) and can be tiled on 4-sides without extra gaps. The work presented here provides a solid basis for the use of these innovative detectors for new diffraction and imaging techniques, which should allow one to obtain more and novel structural information on biological samples that have a limited lifetime within the electron beam.

3.5 Conclusion

A Timepix3 hybrid pixel detector has been integrated in an automated cryo-EM workflow. Its performance at 200 kV has been demonstrated for different event localisation schemes, both in terms of MTF, NNPS, DQE as well as gold-standard FSCs. High-quality single-particle analysis reconstructions could be obtained with this event-driven HPD, at 200 kV, in electron counting mode, with high hit rates and CNN-based event localisation. HPDs could greatly expand the possibilities of (cryo-)EM for structural biology, since they would allow for, at a wide range of energies, both imaging and diffraction-based experiments.

3.6 Data availability

The refined model has been deposited in the Protein Data Bank as PDB entry 8AEY and the map has been deposited in the Electron Microscopy Data Bank (EMDB) as entry EMD-15389. Raw Timepix3 data, processed MRC movie stacks and extracted particles as used for the EMDB map have been deposited in EMPIAR as entry EMPIAR-11113. Raw Timepix3 data and Falcon3 micrographs for calculating MTF, NNPS and DQE have been deposited in Zenodo [53].

3.7 Software availability

The following software has been deposited at Zenodo: (1) tpx3HitParser to parse raw Timepix3 data in a sparse localised event stream [32], (2) tpx3EventViewer to transform the sparse localised event stream into frames [33], and (3) the scripts to measure MTF, NNPS and calculate DQE [54]. The SerialEM plugin to control Timepix3 through the Serval Remote Procedure Service is available from Amsterdam Scientific Instruments.

3.8 Statement about competing interests

Maastricht University owns a patent with authors Ravelli and Van Schayck as inventors (EP3525229) regarding event localisation. Other authors declare no competing interests.

3.9 Acknowledgements

We thank Ye Gao and Eve Timlin for providing protein samples. We are grateful to the M4i Microscopy CORE Lab team of FHML Maastricht University for their support and collaboration. We thank the members of the Amsterdam Scientific Instruments team for their support in building and operating the detector. We thank David Mastronarde for his help and advice for integrating Timepix3 into SerialEM. This research is funded by the Netherlands Organisation for Scientific Research (NWO) within the framework of the Fund New Chemical Innovations, project MOL3DEM, number 731.014.248; European Union Horizon 2020 Research and Innovation Programme, project Q-SORT, number 766970; the PPP Allowance made available by Health~Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships, project 4DEM, number LSHM21029; as well as by the LINK programme from the Province of Limburg, the Netherlands.

3.10 References

1. Taylor, K. A. & Glaeser, R. M. Electron microscopy of frozen hydrated biological specimens. *Journal of Ultrastructure Research* **55,** 448–456. doi:10.1016/s0022-5320(76)80099-8 (1976).
2. Henderson, R. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q Rev Biophys* **28,** 171–93. doi:10.1017/s003358350000305x (1995).

3. Zhang, Y. *et al.* Single-particle cryo-EM: alternative schemes to improve dose efficiency. *Journal of Synchrotron Radiation* **28,** 1343–1356. doi:10.1107/s1600577521007931 (2021).
4. Dainty, J. C. & Shaw, R. *Image Science: Principles, Analysis and Evaluation of Photographic-type Imaging Processes* https://books.google.nl/books?id=WjsOAQAAIAAJ (1974).
5. Ruijter, W. J. D. Imaging properties and applications of slow-scan charge-coupled device cameras suitable for electron microscopy. *Micron* **26,** 247–275. doi:10.1016/0968-4328(95)00054-8 (1995).
6. McMullan, G., Chen, S., Henderson, R. & Faruqi, A. R. Detective quantum efficiency of electron area detectors in electron microscopy. *Ultramicroscopy* **109,** 1126–1143. doi:10.1016/j.ultramic.2009.04.002 (2009).
7. Ruskin, R. S., Yu, Z. & Grigorieff, N. Quantitative characterization of electron detectors for transmission electron microscopy. *Journal of Structural Biology* **184,** 385–393. doi:10.1016/j.jsb.2013.10.016 (2013).
8. Paton, K. A. *et al.* Quantifying the performance of a hybrid pixel detector with GaAs:Cr sensor for transmission electron microscopy. *Ultramicroscopy* **227,** 113298. doi:10.1016/j.ultramic.2021.113298 (2021).
9. Bammes, B. E., Rochat, R. H., Jakana, J., Chen, D.-H. & Chiu, W. Direct electron detection yields cryo-EM reconstructions at resolutions beyond 3/4 Nyquist frequency. *Journal of Structural Biology* **177,** 589–601. doi:10.1016/j.jsb.2012.01.008 (2012).
10. Kühlbrandt, W. The Resolution Revolution. *Science* **343,** 1443–1444. doi:10.1126/science.1251652 (2014).
11. McMullan, G. *et al.* Experimental observation of the improvement in MTF from backthinning a CMOS direct electron detector. *Ultramicroscopy* **109,** 1144–1147. doi:10.1016/j.ultramic.2009.05.005 (2009).
12. Battaglia, M., Contarato, D., Denes, P. & Giubilato, P. Cluster imaging with a direct detection CMOS pixel sensor in Transmission Electron Microscopy. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **608,** 363–365. doi:10.1016/j.nima.2009.07.017 (2009).
13. McMullan, G., Clark, A. T., Turchetta, R. & Faruqi, A. R. Enhanced imaging in low dose electron microscopy using electron counting. *Ultramicroscopy* **109,** 1411–1416. doi:10.1016/j.ultramic.2009.07.004 (2009).
14. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat Methods* **10,** 584–90. doi:10.1038/nmeth.2472 (2013).

15. Bai, X. C., Fernandez, I. S., McMullan, G. & Scheres, S. H. Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. *Elife* **2,** e00461. doi:10.7554/elife.00461 (2013).

16. Vinothkumar, K. R. & Henderson, R. Single particle electron cryomicroscopy: trends, issues and future perspective. *Q Rev Biophys* **49,** e13. doi:10.1017/s0033583516000068 (2016).

17. Battaglia, M. *et al.* Characterisation of a CMOS active pixel sensor for use in the TEAM microscope. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **622,** 669–677. doi:10.1016/j.nima.2010.07.066 (2010).

18. Naydenova, K. *et al.* CryoEM at 100 keV: a demonstration and prospects. *IUCrJ* **6,** 1086–1098. doi:10.1107/s2052252519012612 (2019).

19. Peet, M. J., Henderson, R. & Russo, C. J. The energy dependence of contrast and damage in electron cryomicroscopy of biological molecules. *Ultramicroscopy* **203,** 125–131. doi:10.1016/j.ultramic.2019.02.007 (2019).

20. Heijne, E. H. Semiconductor micropattern pixel detectors: a review of the beginnings. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **465,** 1–26. doi:10.1016/s0168-9002(01)00340-0 (2001).

21. Ballabriga, R., Campbell, M. & Llopart, X. Asic developments for radiation imaging applications: The medipix and timepix family. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **878,** 10–23. doi:10.1016/j.nima.2017.07.029 (2018).

22. McMullan, G. *et al.* Electron imaging with Medipix2 hybrid pixel detector. *Ultramicroscopy* **107,** 401–413. doi:10.1016/j.ultramic.2006.10.005 (2007).

23. Ballabriga, R. *et al.* Medipix3: A 64k pixel detector readout chip working in single photon counting mode with improved spectrometric performance. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **633,** S15–S18. doi:10.1016/j.nima.2010.06.108 (2011).

24. Mir, J. A. *et al.* Characterisation of the Medipix3 detector for 60 and 80keV electrons. *Ultramicroscopy* **182,** 44–53. doi:10.1016/j.ultramic.2017.06.010 (2017).

25. Poikela, T. *et al.* Timepix3: a 65K channel hybrid pixel readout chip with simultaneous ToA/ToT and sparse readout. *Journal of Instrumentation* **9,** C05013–C05013. doi:10.1088/1748-0221/9/05/c05013 (2014).

26. Wang, B., Zou, X. & Smeets, S. Automated serial rotation electron diffraction combined with cluster analysis: an efficient multi-crystal workflow for structure determination. *IUCrJ* **6,** 854–867. doi:10.1107/s2052252519007681 (2019).
27. Guo, H. *et al.* Electron-event representation data enable efficient cryoEM file storage with full preservation of spatial and temporal resolution. *IUCrJ* **7,** 860–869. doi:10.1107/s205225252000929x (2020).
28. Nakane, T. *et al.* Single-particle cryo-EM at atomic resolution. *Nature* **587,** 152–156. doi:10.1038/s41586-020-2829-0 (2020).
29. Visser, J. *et al.* SPIDR: a read-out system for Medipix3 & Timepix3. *Journal of Instrumentation* **10,** C12028. doi:10.1088/1748-0221/10/12/c12028 (2015).
30. Zivanov, J. *et al.* New tools for automated high-resolution cryo-EM structure determination in RELION-3. *Elife* **7,** e42166. doi:10.7554/elife.42166 (2018).
31. Mastronarde, D. N. SerialEM: A Program for Automated Tilt Series Acquisition on Tecnai Microscopes Using Prediction of Specimen Position. *Microscopy and Microanalysis* **9,** 1182–1183. doi:10.1017/s1431927603445911 (2003).
32. Van Schayck, J. P. & Ravelli, R. B. G. *M4I-nanoscopy/tpx3HitParser* version v2.2.0. July 2022. doi:10.5281/zenodo.6874070.
33. Van Schayck, J. P. & Ravelli, R. B. G. *M4I-nanoscopy/tpx3EventViewer* version v2.0.0. July 2022. doi:10.5281/zenodo.6873946.
34. Vulovic, M., Rieger, B., van Vliet, L. J., Koster, A. J. & Ravelli, R. B. G. A toolkit for the characterization of CCD cameras for transmission electron microscopy. *Acta Crystallographica Section D: Biological Crystallography* **66,** 97–109. doi:10.1107/s0907444909031205 (2010).
35. Kuijper, M. *et al.* FEI's direct electron detector developments: Embarking on a revolution in cryo-TEM. *Journal of Structural Biology* **192,** 179–187. doi:10.1016/j.jsb.2015.09.014 (2015).
36. Gijsbers, A., Zhang, Y., Gao, Y., Peters, P. J. & Ravelli, R. B. G. Mycobacterium tuberculosis ferritin: a suitable workhorse protein for cryo-EM development. *Acta Crystallogr D Struct Biol* **77,** 1077–1083. doi:10.1107/s2059798321007233 (2021).
37. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* **180,** 519–30. doi:10.1016/j.jsb.2012.09.006 (2012).

38. Zheng, S. Q. *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat Methods* **14,** 331–332. doi:10.1038/nmeth.4193 (2017).

39. Zhang, K. Gctf: Real-time CTF determination and correction. *J Struct Biol* **193,** 1–12. doi:10.1016/j.jsb.2015.11.003 (2016).

40. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D: Biological Crystallography* **60,** 2126–2132. doi:10.1107/s0907444904019158 (2004).

41. Kaur, S. *et al.* Local computational methods to improve the interpretability and analysis of cryo-EM maps. *Nature Communications* **12,** 1240. doi:10.1038/s41467-021-21509-5 (2021).

42. Emsley, P., Lohkamp, B., Scott, W. & Cowtan, K. Features and development of Coot. *Acta Crystallographica Section D: Biological Crystallography* **66,** 486–501. doi:10.1107/s0907444910007493 (2010).

43. Afonine, P. V. *et al.* New tools for the analysis and validation of cryo-EM maps and atomic models. *Acta Crystallographica Section D* **74,** 814–840. doi:10.1107/s2059798318009324 (2018).

44. Joosten, R. P., Joosten, K., Cohen, S. X., Vriend, G. & Perrakis, A. Automatic rebuilding and optimization of crystallographic structures in the Protein Data Bank. *Bioinformatics* **27,** 3392–3398. doi:10.1093/bioinformatics/btr590 (2011).

45. Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science* **27,** 293–315. doi:10.1002/pro.3330 (2018).

46. Rosenthal, P. B. & Henderson, R. Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy. *Journal of Molecular Biology* **333,** 721–745. doi:10.1016/j.jmb.2003.07.013 (2003).

47. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3,** 160018. doi:10.1038/sdata.2016.18 (2016).

48. Sarkans, U. *et al.* REMBI: Recommended Metadata for Biological Images—enabling reuse of microscopy data in biology. *Nature Methods* **18,** 1418–1422. doi:10.1038/s41592-021-01166-8 (2021).

49. Rabbani, M., Metter, R. V. & Shaw, R. Detective quantum efficiency of imaging systems with amplifying and scattering mechanisms. *Journal of the Optical Society of America A* **4,** 895. doi:10.1364/josaa.4.000895 (1987).

50. Cunningham, I. A., Westmore, M. S. & Fenster, A. Effect of finite detector-element width on the spatial-frequency-dependent detective quantum efficiency. *Medical Imaging 1995: Physics of Medical Imaging*, 143–151. doi:10.1117/12.208331 (1995).

51. Monnin, P., Bosmans, H., Verdun, F. R. & Marshall, N. W. A comprehensive model for quantum noise characterization in digital mammography. *Physics in Medicine & Biology* **61,** 2083. doi:10.1088/0031-9155/61/5/2083 (2016).

52. Zhu, D., Shi, H., Wu, C. & Zhang, X. An electron counting algorithm improves imaging of proteins with low-acceleration-voltage cryo-electron microscope. *Communications Biology* **5,** 321. doi:10.1038/s42003-022-03284-1 (2022).

53. Van Schayck, J. P., Zhang, Y. & Ravelli, R. B. G. *Integration of an event-driven Timepix3 hybrid pixel detector into a cryo-EM workflow* version v1.0.3 (Zenodo, Mar. 2022). doi:10.5281/zenodo.6851220.

54. Van Schayck, J. P., Zhang, Y. & Ravelli, R. B. G. *M4I-nanoscopy/mtf-nps-dqe* version v1.0.0. July 2022. doi:10.5281/zenodo.6867808.

55. Clausen, A. *et al.* LiberTEM: Software platform for scalable multidimensional data processing in transmission electron microscopy. *Journal of Open Source Software* **5,** 2006. doi:10.21105/joss.02006 (2020).

56. Nederlof, I., van Genderen, E., Li, Y.-W. & Abrahams, J. A Medipix quantum area detector allows rotation electron diffraction data collection from submicrometre three-dimensional protein crystals. *Acta Crystallographica Section D: Biological Crystallography* **69,** 1223–1230. doi:10.1107/s0907444913009700 (2013).

57. Jannis, D. *et al.* Event driven 4D STEM acquisition with a Timepix3 detector: Microsecond dwell time and faster scans for high precision and low dose applications. *Ultramicroscopy* **233,** 113423. doi:10.1016/j.ultramic.2021.113423 (2022).

58. Rodenburg, J. Ptychography and Related Diffractive Imaging Methods. *Advances in Imaging and Electron Physics* **150,** 87–184. doi:10.1016/s1076-5670(07)00003-1 (2008).

59. Allars, F. *et al.* Efficient large field of view electron phase imaging using near-field electron ptychography with a diffuser. *Ultramicroscopy* **231,** 113257. doi:10.1016/j.ultramic.2021.113257 (2020).

60. Llopart, X. *et al.* Timepix4, a large area pixel detector readout chip which can be tiled on 4 sides providing sub-200 ps timestamp binning. *Journal of Instrumentation* **17,** C01044. doi:10.1088/1748-0221/17/01/c01044 (2022).

Appendix A Supplemental Information

Table 3.2: Refinement statistics for the BfrB protein model.

Refinement	
Initial model used (PDB entry)	7O6E
Model resolution against LocSpiral map (Å)	2.9
FSC threshold	0.5
Model composition of monomer	
Atoms	1471
H atoms	0
Protein residues	177
Waters	10
B factors ($Å^2$)	
Protein	80.3
Water	73.5
R.m.s deviations	
Bond length (Å)	0.008
Bond angle (°)	1.108
Correlation coefficients	
Mask	0.89
Box	0.83
Validation	
MolProbity score	1.15
Clash score	3.6
Poor rotamers (%)	0
Ramachandran plot	
Favoured (%)	98.86
Allowed (%)	1.14
Disallowed (%)	0

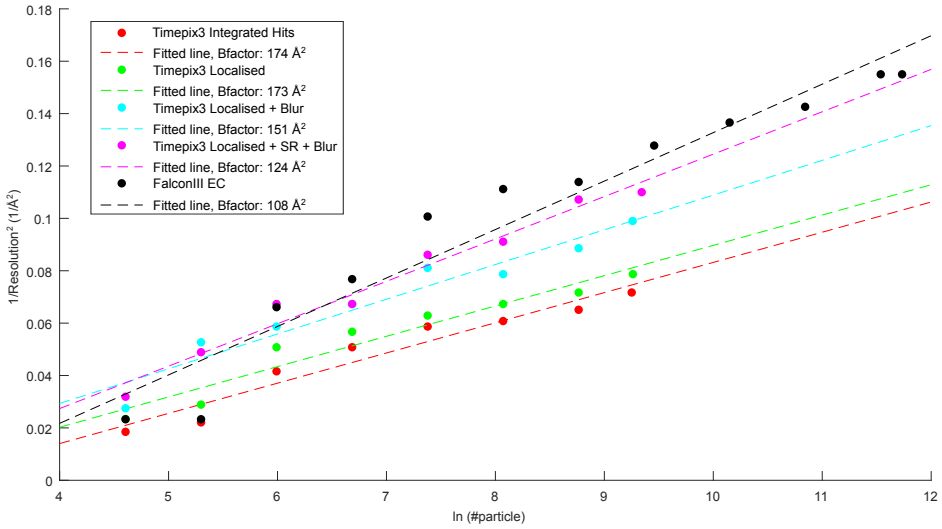


Figure 3.7: The Rosenthal-Henderson B-factor plot showing resolution versus number of particles. B-factors were estimated by fitting a straight line through the inverse of the resolution squared versus the natural logarithm of the number of particles for a range random subsets of full particle list. [46].

4

Charging of vitreous samples in cryo-electron microscopy mitigated by graphene

Adapted from: **Van Schayck*, J. P.**, Zhang*, Y., Pedrazo-Tardajos, A., Claes, N., Noteborn, W. E. M., Lu, P.-H., Duimel, H., Dunin-Borkowski, R. E., Bals, S., Peters, P. J. & Ravelli, R. B. G. Charging of Vitreous Samples in Cryogenic Electron Microscopy Mitigated by Graphene. *ACS Nano* **17,** 15836–15846. doi:10.1021/acsnano.3c03722 (2023).
* both authors contributed equally

Abstract

Cryo-electron microscopy (cryo-EM) can provide high-resolution reconstructions of macromolecules embedded in a thin layer of ice, from which atomic models can be built *de novo*. However, the interaction between the ionising electron beam and the sample results in beam-induced motions and image distortions, which limit the attainable resolutions. Sample charging is one contributing factors of beam-induced motion and image distortion, which is normally alleviated by including part of the supporting conducting film within the beam-exposed region. However, routine data collection schemes avoid strategies whereby the beam is not in contact with the supporting film, whose rationale is not fully understood. Here we characterise electrostatic charging of vitreous samples, both in imaging and diffraction mode. We provide a way to mitigate sample charging by depositing a single layer of conductive graphene on top of regular EM grids. We demonstrate the ability to achieve high-resolution single-particle analysis (SPA) reconstructions at 2 Å when the electron beam only irradiates the middle of the hole on graphene-coated grids, while using data collection schemes that previously failed to produce sub-3 Å reconstructions without the graphene layer. We also observe that the SPA data obtained with the graphene-coated grids exhibit a higher b-factor and reduced particle movement compared to data obtained without the graphene layer. This mitigation of charging could have broad implications for various EM techniques, including SPA and cryo-tomography as well as for the study of the radiation damage and the development of future sample carriers. Furthermore, it may facilitate the exploration of more dose-efficient, scanning transmission electron microscopy–based SPA techniques.

4.1 Introduction

Stimulated by the Resolution Revolution in cryo-electron microscopy (cryo-EM) [1], scientists made spectacular progress in pushing the limits of single particle analysis (SPA) to resolve the structure of biological macro-molecules, both in terms of resolution and particle size. SPA has yielded atomic resolution reconstructions [2, 3], as well as reconstructions from particles smaller than the 38 kDa theoretical size limit [4, 5]. SPA data collection has become much faster over the last years with the availability of faster detectors, advanced microscope automation, fringe-free imaging, aberration-free image shift and hole clustering [6]. This higher throughput has been combined with schemes to improve signal-to-noise ratios (SNR), such as detective quantum efficiency (DQE) improvement of detectors, energy filters, lower energy spread of the electron source, enhanced phase contrast, and reduced sample movements.

Practically, it is well-known that upon irradiation by electrons, biomolecules embedded in a thin hole-spanning vitreous ice layer are observed to move [7]. However, the physical mechanisms behind sample movement are not fully understood. Several hypotheses have been proposed to explain this beam-induced motion [8]. It has been argued that the sample is placed under compressive stress upon rapid cryo-cooling. Electron radiation can induce creep in the presence of this stress that results in doming of the sample in the foil openings [9]. Another hypothesis implicates new mechanical stress from the breakage of chemical bonds and generation of hydrogen gas. Alternatively, electrostatic charging generates an attractive force which cause bending and warping of the thin flexible sample layer. Furthermore, it has been shown biomolecules can also appear to shift without physically moving themselves [10]. Non-ideal and dynamically changing lens conditions would result in image distortions in which molecules appear to move [11]. No matter the cause, movement or distortion during imaging leads to a reduced image quality by dampening high-resolution signals. Several schemes have been proposed to reduce beam-induced motions, such as devitrification [12], vitrification at low cooling rate and elevated temperatures [9, 13], or the use of grids with small holes to have low ice thickness/hole diameter ratio [14].

Charging of biological samples within the TEM has been discussed for decades [10, 15–20]. The ionising electron beam produces secondary electrons that escape from the sample, thus leaving a positive charge on non-conductive specimens [17, 20–24]. Charging can result in unwanted contrast changes, known as the 'bee swarm effect', characterised by fluctuating granularity caused by random surface charging at low magnification and high-defocus conditions [21, 25, 26]. Russo and Henderson described that sample charging is a dynamic process that results from the poor conductivity of the specimen at low electron flux conditions [27]. The positively charged 'footprint' from electron irradiation forms a microlens on

the sample which deflects incoming electron beams, causing a change in phase contrast. This effect, known as the 'Berriman effect', fades when the beam scans nearby regions [22]. The microlens can already be formed within a fluence range of 10^{-3} to $1\,e^{-}\,Å^{-2}$, and may contribute to the defocus change observed in the early frames of micrographs [19, 20, 28].

Sample charging is normally alleviated by including part of the supporting conducting film within the region exposed to the electron beam [8, 10, 22]. Curtis and Ferrier noticed that the 'bee swarm effect' does not happen when part of the beam hits the metal grid bar even when the carbon film and the grid bar within the field of view are not connected [21]. Berriman and Rosenthal designed a special seven-hole C2 aperture, and demonstrated secondary electrons emitted from adjacent areas can reduce the charge on the area of interest [10]. Objective apertures can also reduce specimen charge: secondary electrons emitted by the objective aperture can compensate some of the positive charge of the surface of the specimen [16].

One successful technique to reduce beam-induced motion is 'spot-scan imaging', which focuses an electron beam to a diameter of ~1000 Å and scans it over the specimen to capture multiple images [29–31]. However, this technique would cause charging on thin hole-spanning vitreous ice layer. This technique was only successful with specimens supported by a continuous conductive carbon film, which is undesirable for SPA due to the SNR reduction.

A suitable alternative to amorphous carbon is graphene, which has orders of magnitude higher conductivity than a carbon layer [32, 33], gives minimum background noise and can be overlaid on top of a support layer. Graphene is used in materials science, and increasingly applied in the life sciences [32, 34–40]. Ultra-flat graphene can result in uniform thin ice layers, allowing for high-resolution structure determination of sub-100 kDa proteins [40]. Others described that graphene can be functionalised to improve particle density and orientation [41–43].

In this chapter, we investigate charging effects on vitreous biological specimens and demonstrate that charging can be mitigated by depositing a graphene layer on regular EM grids. We present high-resolution, conventional-TEM SPA reconstructions obtained with such grids, and discuss the importance of understanding charging for future conventional and non-conventional SPA schemes. Being able to mitigate charging by deploying graphene could help to further push the boundaries of resolving high-resolution structures of biomolecules via EM.

4.2 Results

4.2.1 Evaluating the effect of charging in defocused diffraction mode

We used regular SPA samples and grids: *Mycobacterium tuberculosis* ferritin (BfrB) [44] applied to glow-discharged R1.2/1.3 Quantifoil and UltrAuFoil grids. The vitreous ice within the holes of this perforated film acts as an insulator, and a beam size smaller than these holes was used. Similar to Brink *et al.* [22], we used defocused diffraction image (DIFF image) to observe the effect of charging (Figure 4.1a). A hybrid pixel Timepix3 detector ([45], Chapter 2, Chapter 3) was used to record such images. No objective aperture was used in DIFF images/-movies collection. As a function of fluence, the size of the defocused diffracted beam increased in the overfocused condition (Figure 4.1b and c; Supplementary Movies 1 and 2) and decreased in the underfocused condition (Figure 4.1e and f; Supplementary Movies 3 and 4). The normalised DIFF beam size in both overfocus (Figure 4.1d) and underfocus (Figure 4.1g) became stable at a fluence around $1.5\,e^{-}\,Å^{-2}$. The change of the DIFF beam size seems to be insensitive to the type of foil material (carbon or Au) given that the beam is inside the hole. By comparison, the normalised DIFF beam sizes from a conductive crossline grating replica sample (Supplementary Movie 5 and 6) remained constant throughout irradiation.

To provide a more in-depth observation of how samples are charged and discharged, we collected DIFF movies with a moving stage, staying at each location for 5 s (fluence at $1.9\,e^{-}\,Å^{-2}$). At the beginning of irradiation, the diffraction lens was set in overfocus condition, and the beam was confined within a hole on vitreous sample (Figure 4.2a) until the DIFF image beam size fully expanded and became stable (Figure 4.2b). After that, the stage was moved so that the carbon foil came close to the beam. The DIFF image beam size started to decrease when the beam and foil were in close proximity (100 to 150 nm), yet not overlapping (Figure 4.2c). The DIFF image beam size decreased to its minimum when the beam was partially on the supporting foil (Figure 4.2d) and remained at this minimum even when the beam was completely on the foil (Figure 4.2e). The normalised DIFF beam size as a function of time (Figure 4.2f, Supplementary Movie 7) highlights the variation in beam size at various beam locations. There is a noticeable bump at around 50 seconds, reflecting a rapid charge-discharge process as the beam is scanned from one side of the foil to the other. DIFF movies collected in the same manner at a diffraction lens in underfocus condition (Figure 4.2g–l, Supplementary Movie 8) showed a similar yet inverse trend, wherein the DIFF image beam reached its maximum (instead of minimum) and remained stable

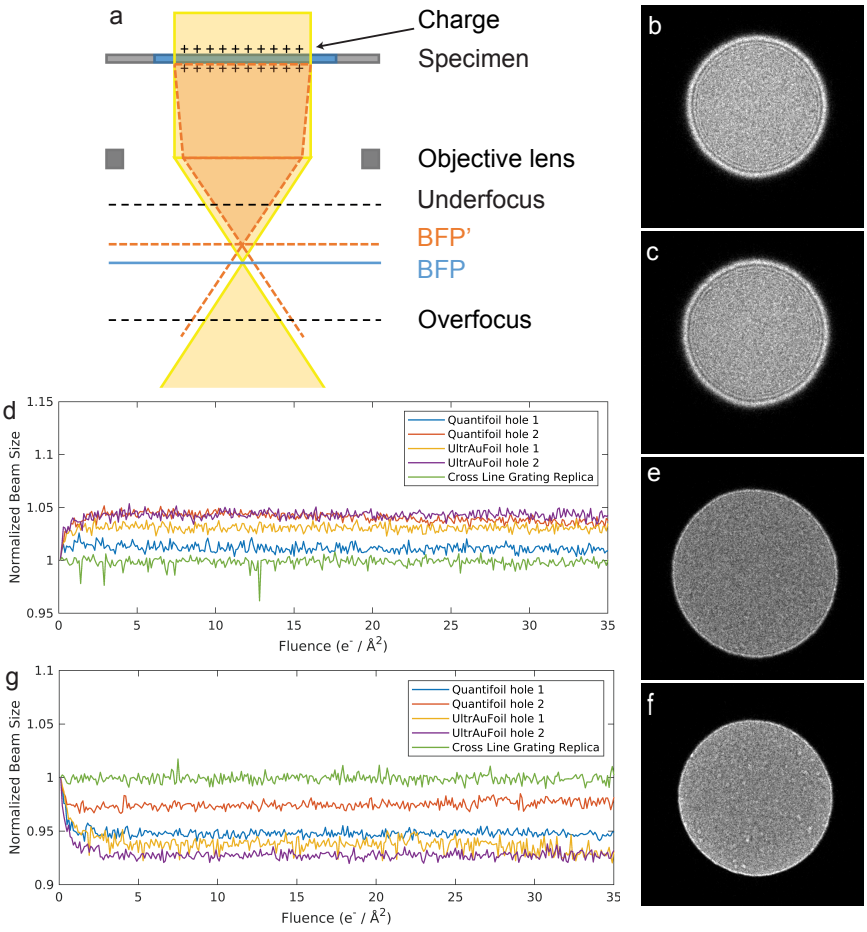


Figure 4.1: Effect of specimen charging on both Quantifoil and UltrAuFoil shown by DIFF images. (**a**) Ray diagram of electron-optical effect of charge on the specimen. The electron beam (solid lines), which irradiates a specimen, is focused by the objective lens at the back focal plane (BFP). The charge on the sample acts as a lens that converges the beam (dash lines) and induces the expansion of the overfocus pattern (**b, c**) and the shrinkage of the underfocus pattern (**e, f**). The DIFF images before (**b, e**) and after (**c, f**) 2 e^- Å^{-2} irradiation of the specimen. The normalised beam radius is plotted as a function of accumulated dose in overfocus (**d**) and underfocus (**g**) conditions for both Quantifoil (blue, red curves) and UltrAuFoil (yellow, purple curves). For comparison, the normalised beam radius as a function of accumulated dose from cross line grating replica (Au) samples is shown (green curve). The electron beam flux on the sample was at 0.38 e^- Å^{-2} s^{-1}.

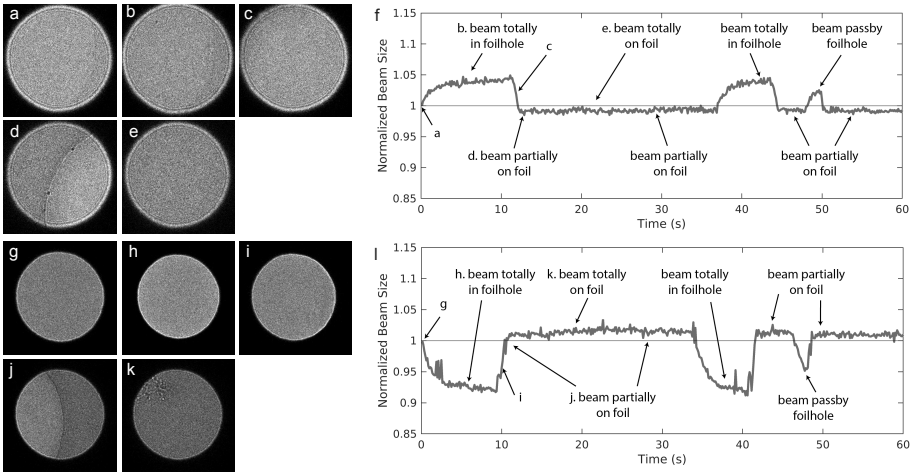


Figure 4.2: Change of DIFF image beam size upon stage move on Quantifoil grids in overfocus (**a–f**) and underfocus (**g–l**) conditions. (**f, l**) The normalised beam radius is plotted as a function of time, with letters corresponding to the panel images shown to the left. In overfocus condition (**a–f**), the beam was completely within the foil hole on ice before irradiation (**a**), increased after a few seconds of irradiation (**b**), then decreased when the carbon foil moves close but not on the beam edge (**c**). The size of DIFF image decreased to its minimum when beam partially hits the carbon foil (**d**), and stayed the same when the beam is completely on the carbon foil (**e**). (**g–h**) In the underfocus condition, a similar but inverse trend was observed. The DIFF image size reduced a few seconds after irradiation (**h**) and increased when the carbon foil moved close to the beam edge (**i, j**), finally reaching its maximum when the beam hit the carbon foil (**j, k**). The electron beam flux on the sample was at 0.38 $e^- Å^{-2} s^{-1}$.

near/on the foil then decreased (instead of increased) in holes. We repeated these experiments with UltrAuFoils and found the same trends (Supplementary Movies 9 and 10).

4.2.2 The use of graphene-coated grids

Next, we repeated the experiments above with grids (both Quantifoil and UltrAuFoil) with an extra graphene layer applied on top of them, to test whether the conductivity from graphene could alleviate charging. We verified the presence of graphene by collecting electron diffraction patterns from samples with amorphous ice on graphene (Figure 4.3a). These show the hexagonal diffraction pattern of graphene, demonstrating that it withstood the 10 s glow discharge, sample appli-

cation, vitrification and grid handling. The DIFF beam size remained unchanged from the beginning of irradiation (Figure 4.3b) up to a fluence of 35 e^- Å^{-2} (Figure 4.3c), in both overfocus (Figure 4.3d) and underfocus (Figure 4.3e) conditions. The DIFF movies with graphene grids are shown as Supplementary Movies 11–14. The DIFF beam size also remained stable for the DIFF image with a moving stage. The normalised DIFF beam size (Supplementary Movies 15–18) as a function of time at both overfocus (Figure 4.3i) and underfocus (Figure 4.3j) conditions showed that it remained unchanged regardless of beam location, whether inside the foil hole (Figure 4.3f), partially on carbon foil (Figure 4.3g), or totally on carbon foil (Figure 4.3h). Notably, the location of the graphene layer (on top of/underneath the vitreous ice in microscope) did not affect the results (data not shown).

4.2.3 Charging in imaging mode

While the DIFF experiments showed clear charging effects up to a fluence of 1.5 e^- Å^{-2}, after which the beam size remained constant, typical SPA data collection schemes conducted in imaging mode uses fluences of tens of e^- Å^{-2}, which may be more prone to aberrations that are not readily apparent in diffraction mode. Thus, we used imaging mode with parallel illumination (Figure 4.1a) to further investigate charging in these regimes. Images were recorded at the flux of 30 e^- Å^{-2} s^{-1} for 1 s with Falcon III detector and no objective aperture at a nominal magnification of 78,000 times and averaged without patch-track motion correction. We first used 20 μm C2 aperture to ensure a beam diameter (800 nm) smaller than the hole size (1.2 μm): the beam did not touch the perforated support film. This setup resulted in severely distorted images, as if the particles were moving outwards from a centre (Figure 4.4a). After 10 s, another image was taken at the same spot with 50 μm C2 aperture (beam size 1.9 μm) so that the beam hit the foil. The resulting image was sharp and similar to the initial state (Figure 4.4b), arguing that the sample had not undergone a plastic deformation, but a reversible process only that affected the image. This image distortion (blurring) and restoration (deblurring) is clearly shown in Supplementary Movie 19. To confirm that this effect is not related to the pre-exposure, we collected movies with apertures in the reverse order, first with 50 μm C2 aperture (Figure 4.4c) and then with 20 μm C2 aperture (Figure 4.4d). We found similar results, a sharp image with the larger aperture and a distorted image with the smaller aperture (Supplementary Movie 20). Importantly, when we performed the same experiments using Quantifoil grids with a graphene layer we consistently obtained sharp images, independent of the size of the beam relative to the foil hole size (Figure 4.4e and f, Supplementary Movie 21; Figure 4.4g and h,

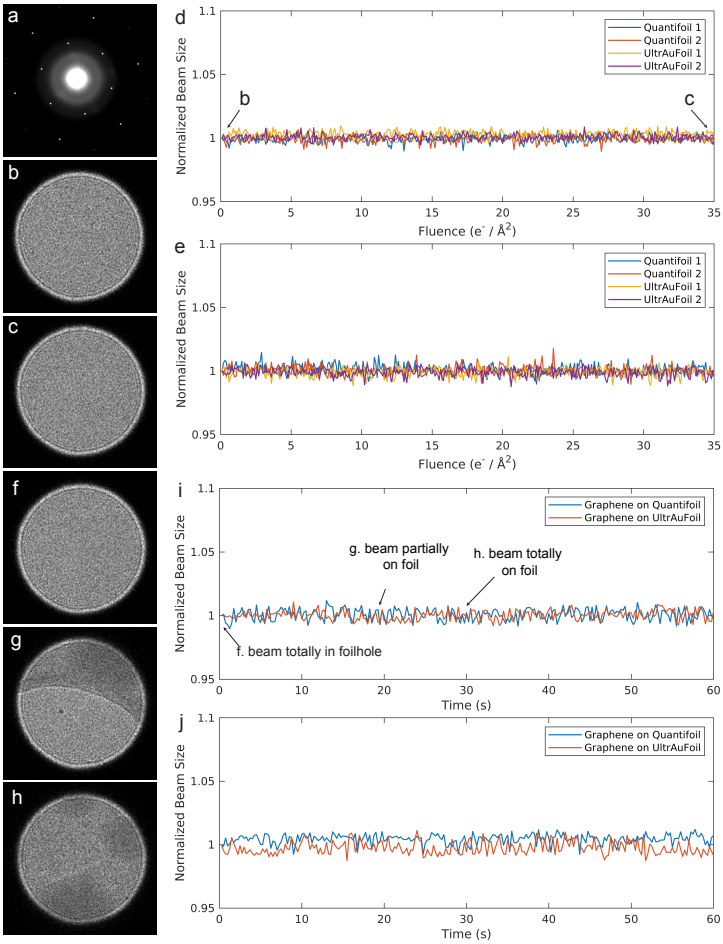


Figure 4.3: DIFF images of vitreous specimen on Quantifoil with a graphene layer, and the similar results were obtained for UltrAuFoil with a graphene layer as well. (**a**) The diffraction pattern of graphene with vitreous ice on it. (**b, c**) The DIFF images at overfocus conditions are similar before (**b**) and after (**c**) 35 e^- Å^{-2} irradiation. (d, e) The normalised beam radius is plotted as a function of accumulated dose in both overfocus (**d**) and underfocus (**e**) conditions for graphene on both Quantifoil (blue, red curves) and UltrAuFoil (yellow, purple curves). When the stage moves, the beam moves from (**f**) foil hole to (**g**) partially on the carbon foil to (**h**) completely on the carbon foil. The DIFF image size was stable as the function of time in (**i**) overfocus and (**j**) underfocus conditions for graphene on both Quantifoil (blue curve) and UltrAuFoil (red curve). The electron beam flux on the sample was at 0.38 e^- Å^{-2} s^{-1}.

Supplementary Movie 22).

4.2.4 Single particle analysis

Next, we determined whether the use of graphene improved image quality for SPA structure determination. We collected SPA datasets of BfrB samples on Quantifoil grids with and without graphene (Table 4.1). Samples on graphene grids exhibited substantially reduced absolute and collective motion compared to samples on grids without graphene throughout the entire SPA fluence period (Figure 4.5a and b). We observed overlapping particles when using the graphene-coated grids and could use a ten-fold dilution of the protein sample to arrive at similar number of particles per micrograph when compared to the grids without graphene. Under conditions where the beam is smaller than the hole and not exposing the supporting foil, we were unable to get a sub-3 Å reconstruction of BfrB using grids without graphene (Supplementary Figure 4.6). However, we could obtain a 2.01 Å reconstruction with graphene grids (Supplementary Figure 2). When the beam size was larger than the hole size, and touching the conductive support, we achieved reconstruction of maps at 2.12 Å resolution with Quantifoil grids without graphene (Supplementary Figure 4.8) and 1.90 Å resolution with graphene-coated grids using a similar number of particles (Supplementary Figure 4.8, Table 4.1).

4.3 Discussion

In this chapter, we examined the effect of charging on cryo-EM SPA samples and demonstrated that the addition of a graphene layer could mitigate this effect, resulting in higher resolution reconstructions allowing for improved SPA data collection schemes.

4.3.1 Charging of sample forms a non-ideal microlens

We set out to observe the effect of charging on cryo-EM SPA sample grids. From Brink *et al.* [22], it is known that the effect of charging is particularly noticeable in defocus diffraction mode (DIFF), where it can be observed as a change in the size of the beam (Figure 4.1a). We found that beam size in DIFF mode became stable after a fluence of $\sim 1.5\,e^{-}\,Å^{-2}$ (Figure 4.1d and g), indicating that the charge saturates at this fluence, in accordance with Schreiber's findings [20]. The beam size change appears to be insignificantly affected by the foil material used, here carbon and gold. However, the results shown in Figure 4.2 indicate they do relate

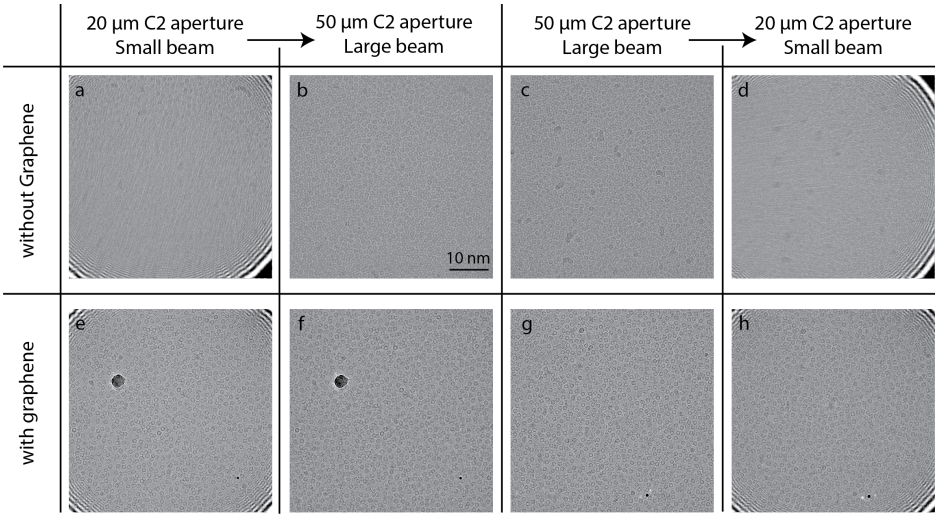


Figure 4.4: Micrographs of BfrB sample collected at the magnification of 78,000 times on a Falcon III at 200 kV. All micrographs were collected at the flux of 30 e^{-} $Å^{-2}$ s^{-1} for 1 s, fractions averaged without motion correction. No objective aperture was used. Micrographs (**a-d**) have samples at the concentration of 50 mg mL^{-1} on normal Quantifoil grids, and micrographs (**e-h**) show samples at the concentration of 5 mg mL^{-1} on Quantifoil grids with graphene layer. The grids have the foilhole size of 1.2 µm in diameter. (**a, b**) Two successive micrographs at the same position on Quantifoil grid, collected with 20 µm C2 aperture, beam size of ~800 nm for 1 s irradiation (**a**), then with a 50 µm C2 aperture, beam size of 1.9 µm), for 1 s irradiation (**b**). (**c, d**) Two successive micrograph pairs at the same position on Quantifoil grids, but first collected with 50 µm C2 aperture for 1 s irradiation, then with 20 µm) C2 aperture for 1 s irradiation. (**e, f**) Two successive micrographs at the same position on Quantifoil grid with a graphene layer, first collected with 20 µm) C2 aperture, beam size of ~800 nm for 1 s irradiation, then with 50 µm C2 aperture, beam size of ~1.9 µm, for 1 s irradiation. (**g, h**) Two successive micrographs at the same position on Quantifoil grids with a graphene layer, but first with 50 µm C2 aperture for 1 s, then with 20 µm C2 aperture for 1 s irradiation.

Table 4.1: Dataset statistics

Dataset	**1**	**2**	**3**	**4**
Grid type	Quantifoil 300 mesh R1.2/1.3	Quantifoil 300 mesh R1.2/1.3 with graphene	Quantifoil 300 mesh R1.2/1.3	Quantifoil 300 mesh R1.2/1.3 with graphene
Microscope	Krios (300 kV)			
Objective aperture (μm)	100			
Nominal magnification	105 000 x			
Pixel size (Å)	0.834			
Camera	K3 (counting)			
Focus range (μm)	-0.8 to -2.0		-0.6 to -1.6	
Exposure time (s)	1.7			
Flux (e^{-} Å^{-2} s^{-1})	23.5			
Fractions	122			
Beam size (μm)	0.9		1.8	
Micrographs (no.)	633	621	2226	1808
Particles (no.)	85707	146626	596238	494154
Symmetry imposed	O			
FSC threshold	0.143			
Map resolution (Å)	3.50	2.01	2.12	1.90
EMDB entry	EMD-18029	EMD-18028	EMD-18030	EMD-18010

to the distance between the beam edge and foil. In the experiment of Figure 4.2, the sample stage is moved while recording and the beam size started to change when foil moved close to the beam edge (100 to 150 nm), but before the beam hits the foil (Figure 4.2c and i). We speculate that the sample starts to discharge when the beam edge and foil are in close proximity. The electron irradiation induces a positive charge on the non-conductive sample surface with its area broader than the beam size. This charge produces a three-dimensional potential distribution that extends further in all directions [17, 20], with the electric field strength of more than a few MV m^{-1} [22, 23]. This potential distribution can deflect incoming electrons and cause a drift of the beam (Supplementary Movies 7-10). These observations confirm the conclusions of [8], where already was indicated that charging leads to a deflection of the incident beam and results in the creation of undesirable lenses. Then the sample discharges when the beam hits the foil (Figure 4.2d and j). Close inspection of the normalised beam size indicates that a non-conducting SPA sample can become charged at an extremely low initial fluence: at the very first frame recorded with 0.047 e^{-} Å^{-2} (Figure 4.2f and l). Overall, the results shown in Figures 4.1 and 4.2 demonstrate that the positive charge induced on the surface of the sample by the electron beam forms a non-ideal microlens that causes the beam size to change in defocused diffraction

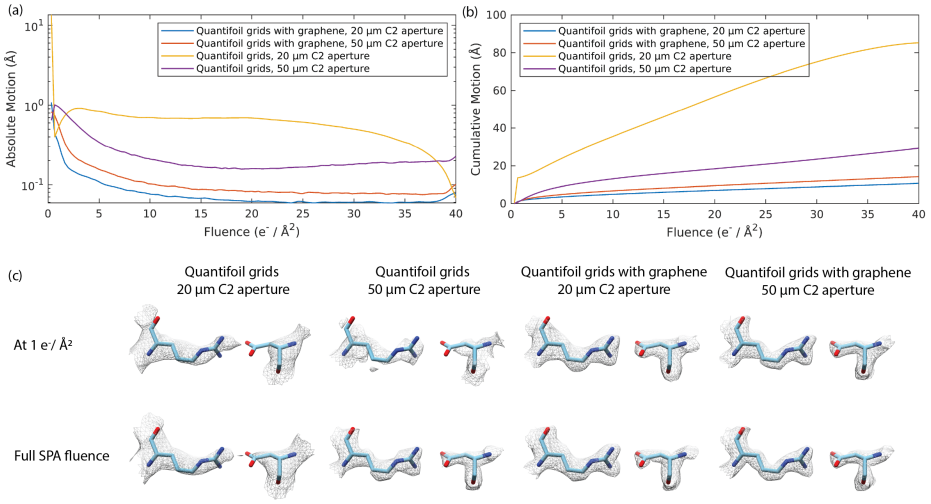


Figure 4.5: Averaged absolute per-frame motion (**a**) and averaged accumulated motion (**b**) as a function of fluence determined by Relion Bayesian polishing of four datasets listed in Table 4.1 (**c**) Density maps reconstructed from four data sets each with a fitted BfrB model (PDB: 7O6E). Reconstructions from $1\,e^{-}\,Å^{-2}$ (top row) and the full SPA fluence up to $40\,e^{-}\,Å^{-2}$ (bottom row). Density maps are drawn at 1.5 RMSD.

mode.

4.3.2 Charging of the sample can severely hinder SPA

Surprisingly, while the change in beam size observed in DIFF mode saturated around $1.5\,e^{-}\,Å^{-2}$, we reproducibly observed continuing distortions in imaging mode, for the full range of fluences normally used in SPA (Figure 4.4a). While distortions were observed in DIFF image as well—features (e.g. ice contamination) were moving outwards even though the DIFF image beam size became stable after $1.5\,e^{-}\,Å^{-2}$ (Supplementary Movie 3), these distortions were reversible, and restored as soon as the beam touched the conductive support film (Figure 4.4b, Supplementary Movie 19), which we attribute to sample discharging. Performing control experiments in reverse order (Figure 4.4c and d) showed that the image distortion was not due to pre-exposure. We speculate that, as sample charging is a dynamic process [27], the continuing distortion of the image in imaging mode might be attributed to aberration effects, which occur due to charge redistribution on the sample surface, that further distort the image even when the absolute

charge of the sample already reached a maximum. Such distortions have hindered routine SPA data collections at the centre of the hole using beam sizes smaller than the hole size. The SPA dataset we collected in this way displayed severe image distortions and selecting particles was challenging (Supplementary Figure 4.6). Efforts to obtain a proper initial map and reconstruction from this dataset were unsuccessful. Despite this, we could eventually obtain a correct BfrB reconstruction utilising modern motion-correction techniques combined with Bayesian polishing, albeit only at 3.5 Å and a low b-factor (Supplementary Figure 4.10). This result not only highlights the power of modern image data processing tools, but also provides a warning as the success of these programs might blind the user to the underlying physical phenomena that prevented them from getting better data.

4.3.3 Graphene mitigates effects of charging

Both DIFF image and imaging experiments illustrate that the conductive graphene layer can alleviate charging. DIFF image beam size was unchanged from the beginning of the irradiation to the full dose typically used in SPA (Figure 4.3d and e). The averaged images presented no blurring independent of the beam location, whether a small beam illuminated middle of a hole (Figure 4.4e) or touched the foil (Figure 4.4g). With a graphene layer, we were able to obtain a good SPA reconstruction of BfrB at 2.01 Å when the beam hits the middle of the hole (Supplementary Figure 4.7b). Additionally, for the BfrB protein that we used to perform the experiment, the graphene helped to concentrate the sample within the holes [13, 41], as we could use a 10 times diluted sample compared to the grids without graphene, and still obtained a similar number of particles per micrograph (Table 4.1). The b-factor remained relatively unchanged regardless of the beam size for sample on graphene-coated grids ($-82\,Å^2$ and $-84\,Å^2$), however, it was significantly higher compared to that of Quantifoil grids ($-178\,Å^2$) (Supplementary Figure 4.10). Despite the b-factor of the dataset without graphene on Quantifoil grids with a large beam ($-83\,Å^2$) being comparable to datasets with graphene, the resolution values on the y-axis at the same x-axis values (number of particles) were lower than the other two datasets, and the standard deviation was high (Supplementary Figure 4.10). While our results demonstrate that graphene mitigated charging and indeed resulted in improvements in movement and resolution (Figure 4.5a and b, Table 4.1), we cannot attribute this improvement to a specific mechanism, whether by suppression of the doming or suppression of the microlens effect. However, it cannot be excluded that graphene may also reduce motion by improving sample stiffness. To address this question, future experiments involving data collection with tilted samples on graphene-coated grids

could provide valuable insight, as it would allow us to observe the movement along the tilt axis and analyse its impact.

When using graphene-coated grids, particle motions were smaller when a small beam irradiated only the middle of the hole compared to when a large beam irradiated both the vitrified ice and the foil, which might relate to the fact that smaller beams (with same flux) deposit less energy in the sample. The findings are consistent with the 'spot-scan imaging' approach described by Downing [31], where a small beam was utilised to minimize the beam-induced motion. This technique expands the options for reducing motion beyond solely using small hole grids [14].

Nevertheless, early-stage, rapid sample motion [8, 46] was observed in Quantifoil grids with and without graphene (Figure 4.5a), indicating that graphene does not prevent stress release at early exposure. To our surprise, reconstructions of maps obtained from the first fractions of the data were very similar to the final maps for the graphene-coated grids, whereas these early-exposure reconstructions show a compromised quality compared to the full exposure reconstructions for grids without graphene (Figure 4.5c). Several methods have been suggested to improve on the quality of the maps that can be obtained from the initial frames, including devitrification [12] and vitrification at low cooling rates and elevated temperatures [13]. However, these techniques may result in the formation of crystalline ice. Although the early-stage motion remains high, our data suggest that the use of graphene can still improve the quality of data obtained from these first frames. However, further investigation is necessary to fully characterise this effect.

4.4 Conclusion

This study focused on investigating the charging effects on cryo-EM SPA samples, and showed that the incorporation of a single graphene layer could effectively alleviate the charging phenomenon with minimal noise. Whether a thicker graphene coating would yield different results would be interesting to explore in future studies. Fabrication of grids with high-quality graphene has been cumbersome due to poor reproducibility, low coverage rate and contamination [47]. We, and others [40], were able to overcome these limitations by using a high quality graphene obtained by an improved transfer method where the reproducibility and coverage rate of graphene grids are increased while minimizing contamination [48]. As a result, these clean graphene grids were able to provide the conductive layer needed to alleviate charging and improve current SPA schemes. While charging is fundamental phenomenon in EM, leading to reduced image contrast and resolution for SPA samples, it also applies for tomography lamellae and non-biological,

non-conductive samples. A conductive layer is essential to minimise charging while imaging non-conductive samples, such as biomolecules in vitreous ice, in particular when the beam is not in contact with the conductive supporting layer. A graphene layer provides good conductivity needed to alleviate charging with minimum background noise. Graphene could also help to reduce radiation damage and provide pristine structures for both materials science and life science samples [49, 50]. Further investigations could be conducted to study the effect of graphene-coated grids on radiation damage in SPA samples, in order to gain a deeper understanding of the electron-induced radiation damage and obtain high-quality structural information of biological molecules. Graphene-coated grids with larger holes could be a promising avenue for future research, allowing for the collection of multiple micrographs with a smaller beam to reduce beam-induced motion and increase data collection throughput. The use of graphene-coated grids could also potentially improve resolution for other cryo-EM applications, for instance, in the detailed structural analysis of lithium batteries, where better insight into materials, interfaces, and degradation mechanisms [51–53] could improve the design and optimization of advanced lithium battery systems.

In summary, the implementation of techniques to mitigate charging is a crucial step in improving current SPA schemes, as well as provide pristine atomic structures of non-biological insulating samples. Mitigation of charging could enable low-dose imaging or scanning transmission EM (STEM) techniques such as ptychography and iDPC [54, 55]. It is worth noting that the effect of charging may differ in TEM and STEM modes, with local charging and discharging occurring in STEM mode depending on the beam's position relative to the sample [56, 57]. Further research is necessary to explore the impact of charging in STEM mode.

4.5 Methods

4.5.1 Production of graphene-coated grids

Graphene grids were prepared starting from a commercially available graphene-Cu foil, produced by chemical vapour deposition (CVD). First, a thin layer of cellulose acetate butyrate (CAB) was applied via spin coating in a CAB-ethyl acetate solution. The CAB-graphene-Cu stack was placed on top of an ammonium persulfate etching solution, etching away the Cu. The etching solvent was gradually diluted using deionised water to neutralise the solution and remove the Cu residues from the etched foil. Filter paper was placed at the bottom of the petri dish and the grids were placed on top using tweezers. The graphene was transferred onto the grids by removing the water using a micropipette. The filter paper containing the graphene-covered grids was placed on a heating plate for

30 min at 35 °C to dry. Finally, the graphene grids were heated in activated carbon for 15 h at 300 °C, which removed the CAB layer (whose melting temperature is between 170 °C and 240 °C).

4.5.2 Sample preparation for cryo-EM

M. tuberculosis BfrB was prepared according to Gijsbers *et al.* [44]and used at a concentration of 50 mg mL^{-1} (as calculated with the Pierce BCA Protein Assay Kit) for Quantifoil and UltrAuFoil (Quantifoil Micro Tools, Germany) experiments. We used 300 mesh grids with 1.2 μm diameter holes. A volume of 2.5 mL was applied onto grids which were glow-discharged in vacuum at the current of 7 mA for 30 s. For grids with graphene layer, a volume of 2.5 μL diluted BfrB (1:10, 5 mg mL^{-1}) was applied onto mildly glow-discharged grids (current of 7 mA for 10 s). Excess liquid was removed by blotting for 3 s using filter paper followed by plunge freezing in liquid ethane using an FEI Vitrobot Mark IV operated under 95% humidity at 4 °C.

4.5.3 DIFF image collection

Diffraction images were collected with Timepix3 hybrid pixel detector ([45], Chapter 2, Chapter 3), on a 200 kV Tecnai Arctica (Thermo Fisher Scientific). The beam was blocked with the pre-specimen beam shutter before exposing the sample for recording. The beam was set to be parallel at a flux of 0.38 e^{-} $Å^{-2}$ s^{-1} passing through amorphous ice films. The DIFF image was obtained by defocusing the diffraction lens. No objective aperture was used for DIFF image data collection.

4.5.4 Single particle data acquisition and image processing

Cryo-EM single-particle data were collected on a Titan Krios at 300 kV with a BioQuantum K3 Imaging Filter with a 20 eV post-column energy filter. The detector was utilised in normal counting mode at a nominal magnification of 105 000 x. Table 4.1 shows the statistics of the data set. Data were processed using the RELION pipeline [58]. Movie stacks were corrected for drift (7 × 7 patches) and dose-weighted using MotionCor2 [59]. The local contrast transfer function (CTF) parameters were determined for the drift-corrected micrographs using Gctf [60]. A first set of 2D references were generated from manually picked particles in RELION [58] and these were then used for subsequent automatic particle picking. Table 1 lists the number of particles in the final data set after particle picking, 2D classification and 3D classification with O symmetry. Beam-tilt parameters, anisotropic magnification and local CTF parameters were refined

and the particles were polished [61]. The resolution of the best final map was 2 Å using the gold-standard $FSC = 0.143$ criterion [62]. The maps have been deposited in the Electron Microscopy Data Bank as entry EMD-18010, EMD-18028, EMD-18029 and EMD-18030.

4.6 Acknowledgements

We thank H. Nguyen for editing the manuscript. We warmly thank the M4i Microscopy CORE Lab team of FHML Maastricht University (MU) for their support and collaboration, and Eve Timlin and Ye Gao (MU) for providing protein samples. Members of the Amsterdam Scientific Instruments team are acknowledged for their Timepix3 detector support. This work benefited from access to the Netherlands Centre for Electron Nanoscopy (NeCEN) with assistance from Ludovic Renault and Meindert Lamers. The authors acknowledge financial support Netherlands Electron Microscopy Infrastructure (NEMI), project number 184.034.014 of the National Roadmap for Large-Scale Research Infrastructure of the Dutch Research Council (NWO); the PPP Allowance made available by Health-Holland, Top Sector Life Sciences & Health, to stimulate public–private partnerships, project 4DEM, number LSHM21029; the LINK program from the Province of Limburg, the Netherlands; as well as financial support from the European Commission under the Horizon 2020 Programme by grant no. 815128 (REALNANO).

4.7 References

1. Kühlbrandt, W. The Resolution Revolution. *Science* **343,** 1443–1444. doi:10.1126/science.1251652 (2014).
2. Nakane, T. *et al.* Single-particle cryo-EM at atomic resolution. *Nature* **587,** 152–156. doi:10.1038/s41586-020-2829-0 (2020).
3. Yip, K. M., Fischer, N., Paknia, E., Chari, A. & Stark, H. Atomic-resolution protein structure determination by cryo-EM. *Nature* **587,** 157–161. doi:10.1038/s41586-020-2833-4 (2020).
4. Henderson, R. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q Rev Biophys* **28,** 171–93. doi:10.1017/s003358350000305x (1995).
5. Zhang, K. *et al.* Cryo-EM and antisense targeting of the 28-kDa frameshift stimulation element from the SARS-CoV-2 RNA genome. *Nature Structural & Molecular Biology* **28,** 747–754. doi:10.1038/s41594-021-00653-y (2021).

6. Konings, S. *et al.* Advances in Single Particle Analysis Data Acquisition. *Microscopy and Microanalysis* **25,** 1012–1013. doi:10.1017/s1431927619005798 (2019).
7. Grubb, D. T. Radiation damage and electron microscopy of organic polymers. *Journal of Materials Science* **9,** 1715–1736. doi:10.1007/bf00540772 (1974).
8. Glaeser, R. M. Chapter Two Specimen Behavior in the Electron Beam. *Methods in Enzymology* **579,** 19–50. doi:10.1016/bs.mie.2016.04.010 (2016).
9. Thorne, R. E. Hypothesis for a mechanism of beam-induced motion in cryo-electron microscopy. *IUCrJ* **7,** 416–421. doi:10.1107/s2052252520002560 (2020).
10. Berriman, J. A. & Rosenthal, P. B. Paraxial charge compensator for electron cryomicroscopy. *Ultramicroscopy* **116,** 106–114. doi:10.1016/j.ultramic.2012.03.006 (2012).
11. Egerton, R. F. Choice of operating voltage for a transmission electron microscope. *Ultramicroscopy* **145,** 85–93. doi:10.1016/j.ultramic.2013.10.019 (2014).
12. Wieferig, J.-P., Mills, D. J. & Kühlbrandt, W. Devitrification reduces beam-induced movement in cryo-EM. *IUCrJ* **8,** 186–194. doi:10.1107/s2052252520016243 (2021).
13. Wu, C., Shi, H., Zhu, D., Fan, K. & Zhang, X. Low-cooling-rate freezing in biomolecular cryo-electron microscopy for recovery of initial frames. *QRB Discovery* **2,** 213–221. doi:10.1017/qrd.2021.8 (2021).
14. Naydenova, K., Jia, P. & Russo, C. J. Cryo-EM with sub–1 Å specimen movement. *Science* **370,** 223–226. doi:10.1126/science.abb7927 (2020).
15. Henderson, R. & Glaeser, R. M. Quantitative analysis of image contrast in electron micrographs of beam-sensitive crystals. *Ultramicroscopy* **16,** 139–150. doi:10.1016/0304-3991(85)90069-5 (1985).
16. Henderson, R. Image contrast in high-resolution electron microscopy of biological macromolecules: TMV in ice. *Ultramicroscopy* **46,** 1–18. doi:10.1016/0304-3991(92)90003-3 (1992).
17. Cazaux, J. Correlations between ionization radiation damage and charging effects in transmission electron microscopy. *Ultramicroscopy* **60,** 411–425. doi:10.1016/0304-3991(95)00077-1 (1995).
18. Vinothkumar, K. R. & Henderson, R. Single particle electron cryomicroscopy: trends, issues and future perspective. *Q Rev Biophys* **49,** e13. doi:10.1017/s0033583516000068 (2016).

19. Russo, C. J. & Henderson, R. Charge accumulation in electron cryomicroscopy. *Ultramicroscopy* **187,** 43–49. doi:10.1016/j.ultramic.2018.01.009 (2018).

20. Schreiber, M. T., Maigné, A., Beleggia, M., Shibata, S. & Wolf, M. Temporal dynamics of charge buildup in cryo-electron microscopy. *Journal of Structural Biology: X* **7,** 100081. doi:10.1016/j.yjsbx.2022.100081 (2023).

21. Curtis, G. H. & Ferrier, R. P. The electric charging of electron-microscope specimens. *Journal of Physics D: Applied Physics* **2,** 1035. doi:10.1088/0022-3727/2/7/312 (1969).

22. Brink, J., Sherman, M. B., Berriman, J. & Chiu, W. Evaluation of charging on macromolecules in electron cryomicroscopy. *Ultramicroscopy* **72,** 41–52. doi:10.1016/s0304-3991(97)00126-5 (1998).

23. Downing, K. H., McCartney, M. R. & Glaeser, R. M. Experimental Characterization and Mitigation of Specimen Charging on Thin Films with One Conducting Layer. *Microscopy and Microanalysis* **10,** 783–789. doi:10.1017/s143192760404067x (2004).

24. Egerton, R. F. Radiation damage to organic and inorganic specimens in the TEM. *Micron* **119,** 72–87. doi:10.1016/j.micron.2019.01.005 (2019).

25. Mahl, H. & Weitsch, W. Nachweis von fluktuierenden Ladungen in dünnen Lackfilmen bei Elektronendurchstrahlung. *Naturwissenschaften* **46,** 487–488. doi:10.1007/bf00626730 (1959).

26. Dove, D. B. Image Contrasts in Thin Carbon Films Observed by Shadow Electron Microscopy. *Journal of Applied Physics* **35,** 1652–1653. doi:10.1063/1.1713709 (1964).

27. Russo, C. J. & Henderson, R. Microscopic charge fluctuations cause minimal contrast loss in cryoEM. *Ultramicroscopy* **187,** 56–63. doi:10.1016/j.ultramic.2018.01.011 (2018).

28. Wang, L. *et al.* Dynamics of the charging-induced imaging instability in transmission electron microscopy. *Nanoscale Advances* **3,** 3035–3040. doi:10.1039/d1na00140j (2021).

29. Downing, K. H. & Glaeser, R. M. Improvement in high resolution image quality of radiation-sensitive specimens achieved with reduced spot size of the electron beam. *Ultramicroscopy* **20,** 269–278. doi:10.1016/0304-3991(86)90191-9 (1986).

30. Bullough, P. & Henderson, R. Use of spot-scan procedure for recording low-dose micrographs of beam-sensitive specimens. *Ultramicroscopy* **21,** 223–230. doi:10.1016/0304-3991(87)90147-1 (1987).

31. Downing, K. H. Spot-Scan Imaging in Transmission Electron Microscopy. *Science* **251,** 53–59. doi:10.1126/science.1846047 (1991).

32. Naydenova, K., Peet, M. J. & Russo, C. J. Multifunctional graphene supports for electron cryomicroscopy. *Proceedings of the National Academy of Sciences* **116,** 11718–11724. doi:10.1073/pnas.1904766116 (2019).

33. Geim, A. K. Graphene: Status and Prospects. *Science* **324,** 1530–1534. doi:10.1126/science.1158877 (2009).

34. Pantelic, R. S., Meyer, J. C., Kaiser, U., Baumeister, W. & Plitzko, J. M. Graphene oxide: A substrate for optimizing preparations of frozen-hydrated samples. *Journal of Structural Biology* **170,** 152–156. doi:10.1016/j.jsb.2009.12.020 (2010).

35. Mohanty, N., Fahrenholtz, M., Nagaraja, A., Boyle, D. & Berry, V. Impermeable Graphenic Encasement of Bacteria. *Nano Letters* **11,** 1270–1275. doi:10.1021/nl104292k (2011).

36. Pantelic, R. S. *et al.* Graphene: Substrate preparation and introduction. *Journal of Structural Biology* **174,** 234–238. doi:10.1016/j.jsb.2010.10.002 (2011).

37. Russo, C. J. & Passmore, L. A. Controlling protein adsorption on graphene for cryo-EM using low-energy hydrogen plasmas. *Nature Methods* **11,** 649–652. doi:10.1038/nmeth.2931 (2014).

38. Deursen, P. M. G. *et al.* Graphene Liquid Cells Assembled through Loop-Assisted Transfer Method and Located with Correlated Light-Electron Microscopy. *Advanced Functional Materials* **30,** 1904468. doi:10.1002/adfm.201904468 (2020).

39. Nickl, P. *et al.* A New Support Film for Cryo Electron Microscopy Protein Structure Analysis Based on Covalently Functionalized Graphene. *Small* **19,** 2205932. doi:10.1002/smll.202205932 (2023).

40. Zheng, L. *et al.* Uniform thin ice on ultraflat graphene for high-resolution cryo-EM. *Nature Methods* **20,** 123–130. doi:10.1038/s41592-022-01693-y (2023).

41. Han, Y. *et al.* High-yield monolayer graphene grids for near-atomic resolution cryoelectron microscopy. *Proceedings of the National Academy of Sciences* **117,** 1009–1014. doi:10.1073/pnas.1919114117 (2020).

42. Xu, J., Cui, X., Liu, N., Chen, Y. & Wang, H.-W. Structural engineering of graphene for high-resolution cryo-electron microscopy. *SmartMat* **2,** 202–212. doi:10.1002/smm2.1045 (2021).

43. Fujita, J. *et al.* Epoxidized graphene grid for highly efficient high-resolution cryoEM structural analysis. *Scientific Reports* **13,** 2279. doi:10.1038/s41598-023-29396-0 (2023).

44. Gijsbers, A., Zhang, Y., Gao, Y., Peters, P. J. & Ravelli, R. B. G. Mycobacterium tuberculosis ferritin: a suitable workhorse protein for cryo-EM development. *Acta Crystallogr D Struct Biol* **77,** 1077–1083. doi:10.1107/s2059798321007233 (2021).

45. Poikela, T. *et al.* Timepix3: a 65K channel hybrid pixel readout chip with simultaneous ToA/ToT and sparse readout. *Journal of Instrumentation* **9,** C05013–C05013. doi:10.1088/1748-0221/9/05/c05013 (2014).

46. Ripstein, Z. A. & Rubinstein, J. L. Processing of Cryo-EM Movie Data. *Methods Enzymol* **579,** 103–24. doi:10.1016/bs.mie.2016.04.009 (2016).

47. Fan, H. & Sun, F. Developing Graphene Grids for Cryoelectron Microscopy. *Frontiers in Molecular Biosciences* **9,** 937253. doi:10.3389/fmolb.2022.937253 (2022).

48. Tardajos, A. P. & Bals, S. https://data.epo.org/gpi/EP4011828A1(2022).

49. Algara-Siller, G., Kurasch, S., Sedighi, M., Lehtinen, O. & Kaiser, U. The pristine atomic structure of MoS2 monolayer protected from electron radiation damage by graphene. *Applied Physics Letters* **103,** 203107. doi:10.1063/1.4830036 (2013).

50. Keskin, S. & de Jonge, N. Reduced Radiation Damage in Transmission Electron Microscopy of Proteins in Graphene Liquid Cells. *Nano Letters* **18,** 7435–7440. doi:10.1021/acs.nanolett.8b02490 (2018).

51. Li, Y. *et al.* Atomic structure of sensitive battery materials and interfaces revealed by cryo–electron microscopy. *Science* **358,** 506–510. doi:10.1126/science.aam6014 (2017).

52. Weng, S., Li, Y. & Wang, X. Cryo-EM for battery materials and interfaces: Workflow, achievements, and perspectives. *iScience* **24,** 103402. doi:10.1016/j.isci.2021.103402 (2021).

53. Guo, B. *et al.* Cryo-EM Revealing the Origin of Excessive Capacity of the Se Cathode in Sulfide-Based All-Solid-State Li–Se Batteries. *ACS Nano* **16,** 17414–17423. doi:10.1021/acsnano.2c08558 (2022).

54. Lazić, I. *et al.* Single-particle cryo-EM structures from iDPC–STEM at near-atomic resolution. *Nature Methods* **19,** 1126–1136. doi:10.1038/s41592-022-01586-0 (2022).

55. Zhou, L. *et al.* Low-dose phase retrieval of biological specimens using cryo-electron ptychography. *Nature Communications* **11,** 2773. doi:10.1038/s41467-020-16391-6 (2020).

56. Elad, N., Bellapadrona, G., Houben, L., Sagi, I. & Elbaum, M. Detection of isolated protein-bound metal ions by single-particle cryo-STEM. *Proceedings of the National Academy of Sciences* **114,** 11139–11144. doi:10.1073/pnas.1708609114 (2017).

57. Velazco, A., Jannis, D., Béché, A. & Verbeeck, J. Reducing electron beam damage through alternative STEM scanning strategies. Part I – Experimental findings. *arXiv.* doi:10.48550/arxiv.2105.01617 (2021).

58. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* **180,** 519–30. doi:10.1016/j.jsb.2012.09.006 (2012).

59. Zheng, S. Q. *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat Methods* **14,** 331–332. doi:10.1038/nmeth.4193 (2017).

60. Zhang, K. Gctf: Real-time CTF determination and correction. *J Struct Biol* **193,** 1–12. doi:10.1016/j.jsb.2015.11.003 (2016).

61. Zivanov, J. *et al.* New tools for automated high-resolution cryo-EM structure determination in RELION-3. *Elife* **7,** e42166. doi:10.7554/elife.42166 (2018).

62. Scheres, S. H. W. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nature Methods* **9,** 853–854. doi:10.1038/nmeth.2115 (2012).

63. Van Schayck, J. P. *et al. Charging of vitreous samples in cryo-EM mitigated by graphene - Supplementary Figures and Movies* (June 2023). doi:10.6084/m9.figshare.23244299.v1.

64. Rosenthal, P. B. & Henderson, R. Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy. *Journal of Molecular Biology* **333,** 721–745. doi:10.1016/j.jmb.2003.07.013 (2003).

Appendix A Supplemental Information

Supplementary movies have been deposited at: van Schayck, J. P. *et al. Charging of vitreous samples in cryo-EM mitigated by graphene - Supplementary Figures and Movies* (June 2023). doi:10.6084/m9.figshare.23244299.v1

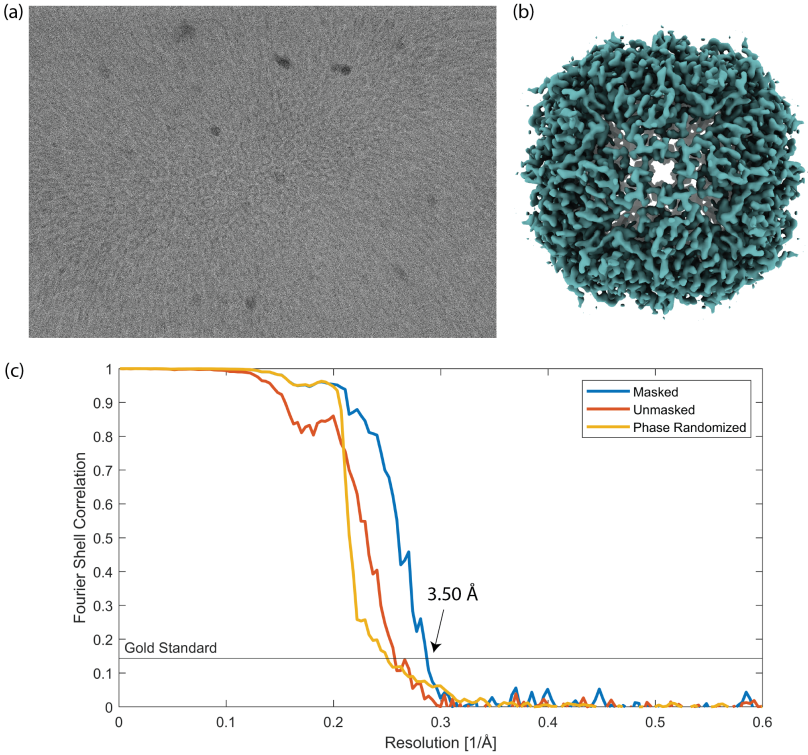


Figure 4.6: Single-particle analysis of the BfrB data set on a Quantifoil grid without graphene, under conditions where the beam (900 nm) was smaller than the hole (1.2 µm) and not exposing the supporting foil. (**a**) A micrograph of highly concentrated (50 mg mL^{-1}) BfrB in vitreous ice. (**b**) 3D reconstruction from 85,707 particles at 3.50 Å resolution. **c**) Gold-standard Fourier shell correlation (FSC) before (red line) and after (blue line) masking, and the phase-randomised FSC (yellow line).

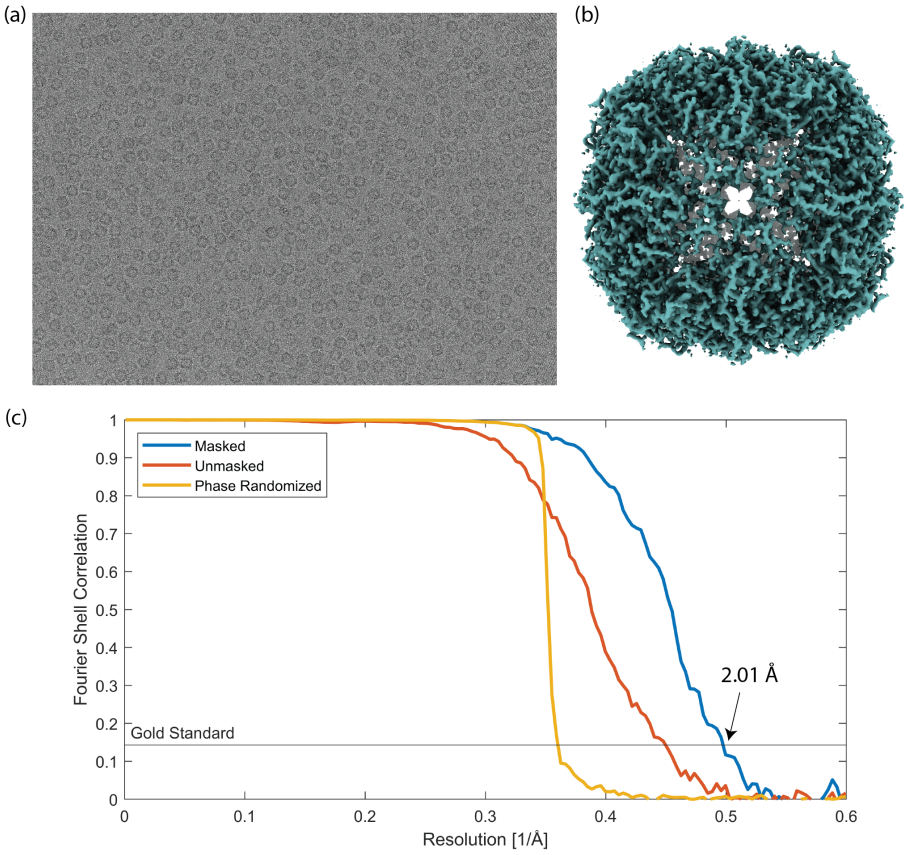


Figure 4.7: Single-particle analysis of BfrB data set on Quantifoil grids with a graphene layer, under conditions where the beam (900 nm) was smaller than the hole (1.2 µm) and not exposing the supporting foil. (**a**) A micrograph of diluted (5 mg mL^{-1}) BfrB in vitreous ice. (**b**) 3D reconstruction from 146,626 particles at 2.01 Å resolution. (**c**) Gold-standard Fourier shell correlation (FSC) before (red line) and after (blue line) masking, and the phase-randomised FSC (yellow line).

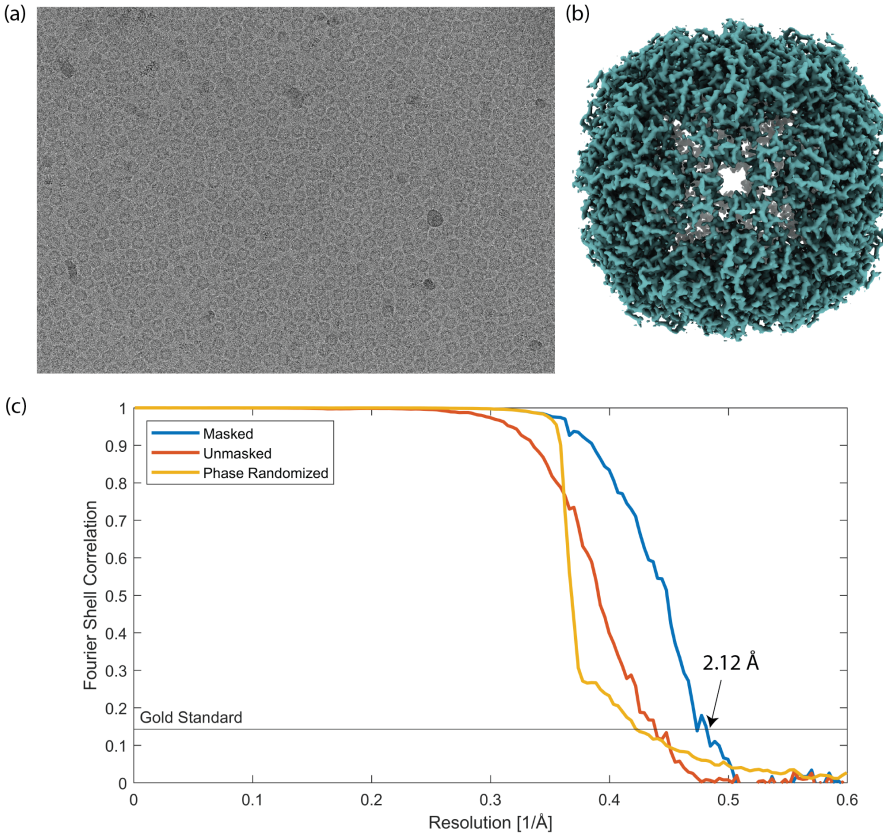


Figure 4.8: Single-particle analysis of BfrB data set on normal Quantifoil grid without graphene, under conditions where the beam (1.9 µm) was larger than the hole (1.2 µm). (**a**) A micrograph of highly concentrated (50 mg mL^{-1}) BfrB in vitreous ice. (**b**) 3D reconstruction from 596,238 particles at 2.12 Å resolution. (**c**) Gold-standard Fourier shell correlation (FSC) before (red line) and after (blue line) masking, and the phase-randomised FSC (yellow line).

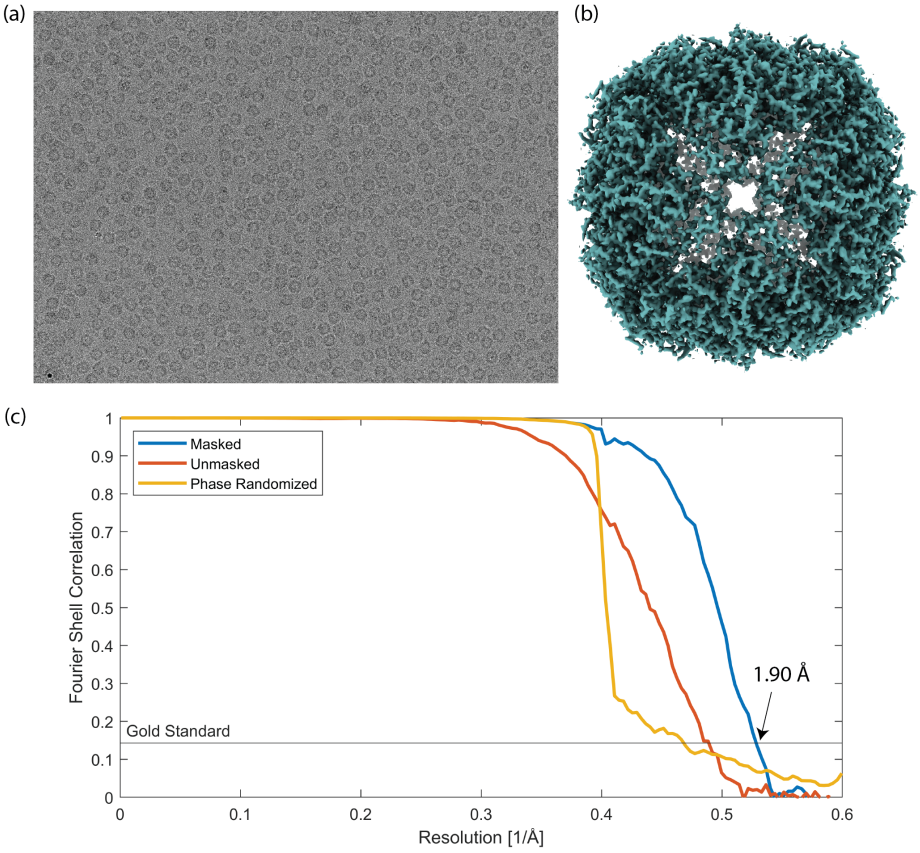


Figure 4.9: Single-particle analysis of BfrB data set on Quantifoil grids with a graphene layer, under conditions where the beam (1.9 µm) was larger than the hole (1.2 µm). (**a**) A micrograph of diluted (5 mg mL^{-1}) BfrB in vitreous ice. (**b**) 3D reconstruction from 494,154 particles at 1.90 Å resolution. (**c**) Gold-standard Fourier shell correlation (FSC) before (red line) and after (blue line) masking, and the phase-randomised FSC (yellow line).

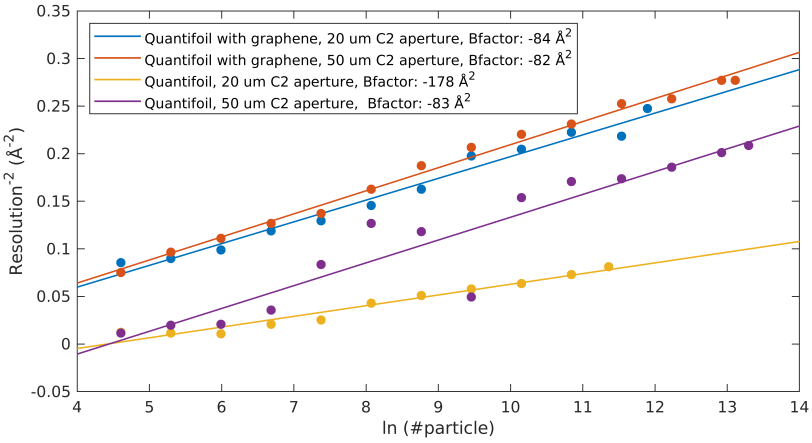


Figure 4.10: The Rosenthal-Henderson B-factor plot showing resolution versus number of particles. B-factors were estimated by fitting a straight line through the inverse of the resolution squared versus the natural logarithm of the number of particles for a range random subsets of full particle list. [64]

5

First-line Research Data Management for Life Sciences: a Case Study

Adapted from: **Van Schayck, J. P** & Coonen, M. First Line Research Data Management for Life Sciences: a Case Study. *International Journal of Digital Curation* **16,** 13. doi:10.2218/ijdc.v16i1.761 (2022).

Abstract

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 performed 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. Here, 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. 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 creating shared knowledge between DataHub and the data stewards, including insights how to focus support on the most reusable datasets. This model has shown to be very beneficial given limited time and personnel. We found that co-hosting 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. Overall, utilising the expertise and communication channel of the embedded data stewards was also instrumental in our RDM success. Looking into the future, we foresee the need to further embed the role of data stewards into the lifeblood of the research organisation, along with policies on how to finance long-term storage of research data. The latter, to remain feasible, needs to be combined with a further formalising of appraisal and reappraisal of archived research data.

5.1 Introduction

Modern life sciences studies depend on the collection, management and analysis of comprehensive datasets. The focus in life sciences is no longer on the collection of a single sample but on arrays of samples analysed and collected in parallel. Researchers have opted for multimodal approaches in their experiments because multiple techniques are required to reveal better insights into dynamic molecular mechanisms underlying biochemical processes. The limiting factor in research nowadays is the pace at which researchers can analyse their data rather than the amount of samples that can be processed [1]. Life sciences research is often conducted by relatively small research groups. These small teams have sometimes been dubbed 'small science' opposed to the 'big science' of large-scale astronomy or physics projects [2]. While this is undoubtedly an oversimplification of reality [3], the term 'small science' can be used to quickly sketch the difficulties faced by these research groups in managing their data. For example, reliable collection and transformation of data into open data formats as well as using community standards for metadata can be a bottleneck when tenured expertise and IT infrastructure are lacking. Thus, overall it has proven difficult for individuals in these small- and medium-sized research labs to manage their ever-growing mountain of research data [4].

Parallel to the increasing difficulty for researchers to manage their data, there has been an increasing urgency to openly publish research data [5]. Initiatives like FAIR and Open Science, among others, have been pushing to share annotated research data openly and in a structured way [6]. These initiatives are mostly imposed top-down on researchers through, for example, the requirements in data management plans (DMPs). They set very high goals and standards, some of which can feel very far from the current daily practice of researchers [7, 8].

In response to both these initiatives and the challenges faced by researchers, research institutes have formed or strengthened existing research data management (RDM) support groups. These groups are based at academic libraries, IT departments, (bio)-informatics research groups and/or other existing research support structures that institutes may have had in place [9]. Historically, academic libraries have had a central role in the archiving and curation of research output. For decades, they have been at the forefront of research digitalisation in the form of digital repositories or other digital services. This practice made it logical for academic libraries to also step into the field of RDM, primarily serving an advisory role as opposed to providing technical RDM services [10, 11].

One of the key challenges for RDM support groups is translating the aforementioned top-down initiatives into day-to-day practice for researchers, while at the same time not losing touch with the primary goal of moving their research forward. In this case study, we reflect on the lessons learnt by the RDM sup-

port group DataHub Maastricht (hereafter DataHub) from Maastricht University and the Maastricht University Medical Centre in the Netherlands. We present a brief history of DataHub and its focus on life sciences by its support of the data-intensive research institutes in Maastricht. We present several strategies and initiatives that have yielded success, including organising incentives, the use of disciplinary data stewards and how to prioritise RDM support. From a technological perspective we present a method for reducing storage costs, the effect we have had in supporting tailored domain specific platforms and the need for cross-institutional access to services. We also review caveats and future prospects for RDM support in life sciences in general.

5.2 Background

Maastricht University is a medium-sized and relatively young (45 years old) university in the south of the Netherlands, with approximately 4,400 employees and 20,000 students. The university has six faculties, with over half its employees located at the Faculty of Health, Medicine and Life Sciences (FHML). This faculty has close links with its next-door neighbour the hospital of Maastricht; together they form the Maastricht University Medical Centre MUMC+.

DataHub (initially dubbed Research IT) was founded in 2015 within the FHML IT support department. Appointed by the Board of MUMC+, DataHub's focus is on supporting and improving RDM for both the hospital and the faculty. DataHub consists of a small core team of data and software engineers. This core team is complemented by disciplinary data stewards, many of whom have joint appointments at DataHub and a research group.

Shortly after being founded, DataHub set out to design and build an infrastructure to support its RDM goals. This infrastructure became the Maastricht Data Repository (MDR). At its core, the MDR was built using the integrated Rule-Orientated Data System (iRODS). Briefly, iRODS provides storage virtualisation in one common directory namespace, authentication and authorisation, combined with flexible metadata on object and collection level all under the control of server-side policies written in either its own rule language or Python [12]. iRODS can be used as a generic RDM platform or to build highly specialised workflows. It is also known for its capability to handle high volumes of data. For example, the Wellcome Trust Sanger Institute uses iRODS to serve more than 30 PiB of molecular sequencing data to hundreds of internal users [13, 14].

The use of iRODS has gained much traction in the Dutch RDM landscape over the last five years and a thriving community around it has sprung into being [for example: 15–17]. Over 10 institutes are now deploying iRODS, including SURF (the national cooperative association of education and research institutes

for digital services) that provides several iRODS-based services. Several Dutch institutes have become iRODS consortium members, and iRODS user conferences have been organised twice in Utrecht (in 2017 and 2019).

The MDR serves as a generic data repository for researchers. Data can be ingested to the MDR by users via so called 'drop zones' which are accessible via Windows network shares and WebDAV. Users initiate these drop zones via a web interface where detailed metadata can be added. For certain metadata fields there is an ability to select ontology-controlled terms. Once ingested, a drop zone becomes a collection within a research project and is assigned a persistent Handle identifier. Access to project data is controlled on either a user or group level with three distinct roles (manager, contributor, viewer). Depending on the policy assigned to a project, its data are either stored on premises (university and hospital) or remote on offline tape library. Data ready for publication can be published directly from within the MDR to DataverseNL, a national instance of Dataverse [18]. Several applications make use of the different iRODS APIs for domain specific workflows implemented on top of the MDR.

In the five years after the initial conception of the MDR, DataHub broadened its scope of RDM support. One of the challenges faced by the DataHub team was to position the MDR not only as endpoint for inactive data but also as an active repository in all phases of the research data life cycle. Different research domains have different requirements for the active phases of the research data life cycle, which DataHub needed to consider (detailed in Technological Lessons below) when providing guidance and support.

As of this writing, the MDR hosts 276 TiB of data, across 272 research projects and is used by 339 researchers (Figure 5.1) from approximately 10 different research departments within MUMC+. Two examples are the Maastricht Multi-Modal Molecular Imaging Institute (M4i) and the MERLN Institute for Technology-Inspired Regenerative Medicine. M4i uses imaging mass spectrometry and cryogenic electron microscopy to study the molecular world, while MERLN uses high-content screening microscopy to study the interaction between biomaterials and tissue. Both institutes require very data-intensive research methods but have different RDM needs. As such, they both approached DataHub for RDM guidance and support.

5.3 Organisational lessons

RDM and its related support do not take place in isolation. They are part of the broader research support ecosystem of an institute. Typically, the support that can be provided is limited by the available personnel and time, in turn often dictated by financial limits. We identified the following lessons in building and

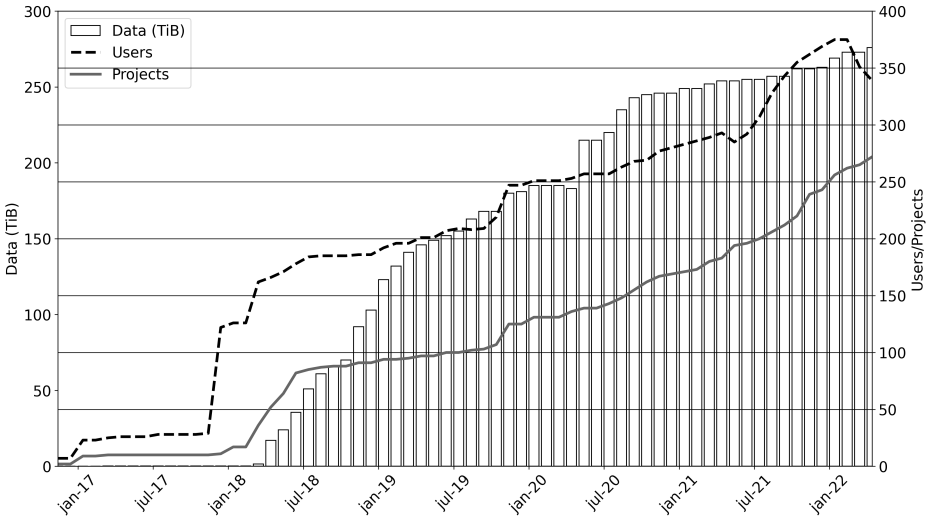


Figure 5.1: Growth of the Maastricht Data Repository over 5 years by data (bars), research projects (solid curve) and users (dashed curve).

improving the organisational model of DataHub.

5.3.1 Appointing disciplinary data stewards

A data steward can be defined as a person responsible for keeping the quality, integrity, and access arrangements of data and metadata in a manner that is consistent with applicable law, institutional policy, and individual permission [19]. In recent years, the role of the data steward has become more and more clearly defined within the RDM and Open Science community and is recognised as a key instrument for moving RDM forward [20]. Due to the diversity of requirements between, or even within, research domains in the life sciences, there is also a need for disciplinary data stewards [21]. Mons already stated this need for a high granularity of data stewards and recommended to strive towards one data steward for every 20 researchers [20].

The DataHub team has made efforts to encourage the embedding of disciplinary data stewards at different research groups. DataHub has strived to embed data stewards at the level of the research group rather than the faculty; currently, there are about 10 data stewards employed for about 300 researchers. The embedding has taken place in a variety of ways, but preferably in the form of a shared position between the research group and DataHub. Stewards have often been recruited

from the ranks of the research groups themselves, but occasionally also newly filled positions have been realised. Due to the mutual benefit for DataHub and the research group, this has always been possible to achieve in a funding neutral way.

The shared placement has had two important effects: (1) the data steward is able to help researchers with their day-to-day practice of data management, data analysis and general IT issues and (2) the data steward possesses or obtains valuable domain-specific knowledge and can translate this knowledge to and from the DataHub infrastructure. Overall, the data stewards have a signalling role by providing early coaching of researchers into RDM. Our experiences show that DMPs can provide good first points of contact to talk about RDM between the data steward and the researcher. Data stewards are encouraged to make use of DMPMaastricht, an instance of DMPOnline [22], for this purpose.

In addition to supporting researchers, data stewards act as stakeholder for DataHub's research software engineers, who work according to the Scrum framework [23]. Scrum is a software development methodology where development takes place incrementally and focus on user value is paramount in all phases of the process. It is a way to get business requirements delivered into working code effectively. In Scrum there are multiple roles defined: the product owner is taking care of stakeholder management and transforming business ideas into user stories, the development team is committed to turn user stories into working code or features and the Scrum master is guiding the process and resolving any impediments. Data stewards, in their role as stakeholders, work together with the product owner to define and select the user stories that bring the most user (i.e., researcher) value to improvements of the MDR.

5.3.2 Providing incentives to start RDM

The benefits of organising and sharing data are not always immediately clear, and the gains to be made are sometimes not obvious [24]. Regularly, the benefits of RDM only become obvious months or even years after implementation. As famously said by the archivist Jason Scott 'metadata is a love note to the future'. This 'love note' does not only apply to metadata but to research data in general. Short-term members of research teams will have moved on before seeing the benefits of the efforts they put into organising and/or sharing their data [10, 25]. Within small research groups, where the bulk of data is generated by PhD students with fixed short-term contracts, this limited time perspective can especially be dominant.

Over the years, we have learnt that it is instead most effective to incentivise the principal investigators (PI) of research groups to instigate RDM. With the PI of a group on-board and directing this policy, it is much easier to get her/his PhD

students to participate. A clear incentive for the PIs can be a financial one, for example, in the reduction of storage costs (see technological lesson 'Reducing storage costs through tiered storage').

We have also seen potential pitfalls in the hope to realise new incentives for researcher. We explored the possibility of making a certain technology or feature in the RDM service available as a possible incentive to users to start RDM. Thus far, we have experienced this strategy to be ineffective. For example, we implemented a fully featured semantic search engine, using Ontoforce DISQOVER, on top of the MDR to allow researchers to easily find and retrieve datasets, both their own and of their colleagues. While this feature added value to existing users, it did not propel researchers to initiate RDM and MDR usage. Mainly because finding someone else's data does not encourage you to add your own necessarily. Another example is that the DataHub RDM infrastructure, when used properly, could provide a track-record of all the data's stages in the research data life cycle. Whereas this traceability is important to the Faculty Board, we found that researchers did not use/benefit from it and some reported it just to be a burden.

While single technological features could be a stimulus for RDM initiation, the two we looked at here were generally not effective (as incentive). We will continue to identify others in the future that could be more appealing, and encourage other RDM services to do so.

5.3.3 Reducing bureaucracy: no wrong door policy

For a researcher, RDM can be experienced as yet another (bureaucratic) topic to be covered while doing their 'real work' [24]. Similar to any information they need to provide about legal, ethical or (bio)-safety concerns for their research, RDM can just feel like more paperwork. Not knowing where to go for RDM guidance might add to the researcher's feeling about RDM bureaucracy. We have learnt to focus on minimising this feeling by adapting the policy of 'no wrong door'.

At MUMC+, various groups provide RDM support in one form or another. These include (but are not limited to) the university library, a software engineering department, a data science research group, a clinical trial centre, and ourselves at DataHub. By intergroup collaboration, shared personnel between these groups and active encouragement, we established clear channels and an atmosphere where researcher's requests are quickly directed to the right person at the most appropriate group to handle the solution.

An alternative approach could be to create an overarching layer or support desk to handle all first-line requests. However, this structure does not align with the decentralised disciplinary data stewards who also have flexible roles between the various groups that provide RDM support. It may turn out in the future that

an overarching support desk may still be beneficial, but this would require close collaboration with the same data stewards.

5.3.4 Prioritising efforts on the most reusable data

The RDM community is at a stage where, for the foreseeable future, far more research data are being generated than can realistically be supported and assisted in making fully FAIR and Open. Making data truly interoperable on a Linked-Data level, as intended by the FAIR principles, is a daunting task.

In addition, at DataHub more support requests are coming in than can be handled. In effect, it means that choices have to be made as to which research or datasets are to be supported and which are not. DataHub decided to use the criterion of a dataset's potential reuse value once published; that is, additional work put into making such a dataset more FAIR or Open would be greatly amplified in the future by its reuse.

DataHub has taken several approaches to keep the focus on the potential reuse of data: (1) encourage the search for and use of domain-specific repositories. These often offer data type specific functionality and visualisation, provide better indexes and mandate more detailed metadata when compared to generic repositories. Submission to such a repository will automatically make data more FAIR, even if this means only part of the research data of a study is published; (2) Identify, through the embedded data stewards, the research projects that may have the highest potential for future reuse. This early knowledge allows RDM to be designed and implemented as the data are being generated; (3) Use an approach (here: Agile/Scrum) that directs software development efforts to features that have the highest user value. Additionally, one is encouraged to critically think how new features contribute to the effectiveness of their RDM services.

The topic of prioritising efforts on reusable data is related to the question Christine Borgman asked: 'If data sharing is the answer, what is the question?'. She asked the rightful question whether the effort of making data available is worth the effort put into it. It should remind us that in the triangle of RDM, FAIR and Open Science we should keep the use to which data can be put to at the heart of our activities [7].

5.4 Technological lessons

We have learnt that establishing and providing appropriate technology for RDM are essential, but very challenging. The requirements for researchers within different life sciences domains are very diverse and can even differ greatly within

the same research group. At the same time, any technology has limited flexibility to support this diversity of use.

5.4.1 Reducing storage costs through tiered storage

A clear incentive for researchers to start RDM is reducing their data storage costs. As is common worldwide, Maastricht researchers often pay directly for storage costs from the research grants they obtain. The consequence of this is that researchers often choose for the lowest costs and store their data on external hard drives or self-managed network storage solutions, which is the opposite of good RDM practice [10]. This decision makes it far more likely for research data to become unavailable after publishing [26, 27]. Therefore, opportunities to reduce storage costs with the additional benefit of RDM functionality will be attractive.

At DataHub, we have formulated this financial incentive in two ways. Firstly, we have prioritised the principle that storage is a public utility, similar to water and electricity, and should be heavily subsidised by the faculties. Up to 100 GiB of storage is free of charge. Above this, storage is offered at cost price or lower, while additional RDM features on top of this are free. Secondly, we have developed a tiered storage system. Most networked storage is expensive because it keeps all data available at high speed at all times. For RDM purposes, data can often be migrated ('tiered') to less expensive and slower performing storage. However, this option must remain transparent in use and easy to execute to be an attractive choice for researchers.

Using iRODS, it was possible to meet these requirements and implement this as functionality of the MDR (Figure 5.2). Data is stored long-term on the tape library of SURF in Amsterdam, which is offered at a very competitive price. Due to the way tape storage functions, it cannot handle datasets with many small files. Usually, this is overcome by bundling small files together, for example in tarballs. We chose to implement an iRODS policy to migrate only files over 256 MiB to tape, which results in datasets that are stored mixed between remote tape and on on-premise solutions. The choice for 256 MiB was made based on technical grounds of the tape archive. However, for the whole MDR, 70% of the volume of data is contained in files over 256 MiB of size, demonstrating that bulk of the data storage is at the most economical tier. Transparency for end users is maintained: they still see their datasets, presented by the catalogue provider, in the same way as before. Currently, dataset tiering is manually triggered by project managers in the web interface of the MDR. Looking into the future, DataHub would like to extend this functionality to perform tiering automatically based on access time.

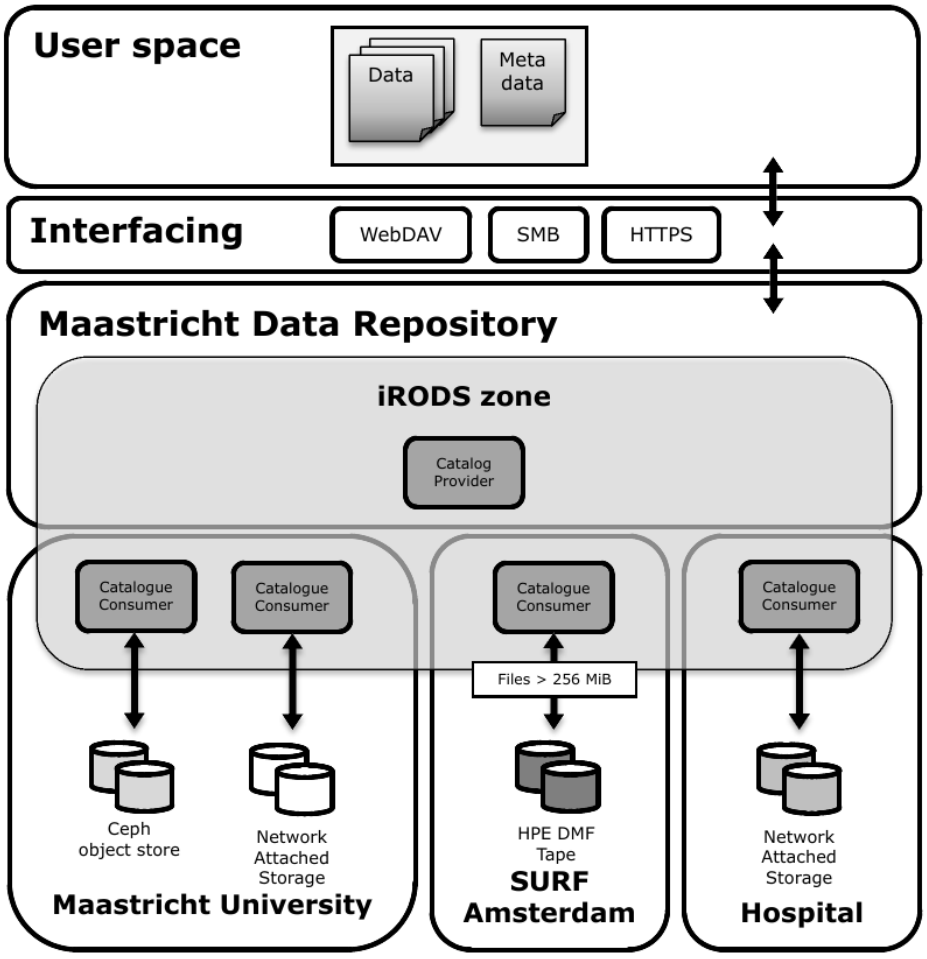


Figure 5.2: The Maastricht Data Repository uses iRODS for the transparent tiering of data across multiple storage solutions and geographical locations (bottom row). Data and its metadata are organised in projects and collections in iRODS and can be stored and retrieved by different interfaces: SMB, WebDAV or HTTPS. The relational database in iRODS's catalogue provider stores system metadata, logical paths and authorisations, among other things. Data storage is distributed via different catalogue consumers to various on-premise or remote locations based on rules and user requirements. One of these rules (policies) is that only files over 256 MiB in size can be stored on tape.

5.4.2 Supporting data-intensive and diverse life sciences research

Life sciences research has become more data intensive in recent years. Dealing with multi-terabyte datasets with millions of files places more strain on all IT infrastructures, especially bandwidth and storage capacity as well as optimised user experience of the tools that are provided. For example, performing uploads and downloads through a web browser is not realistic for terabyte-size datasets.

One of the answers to this challenge was building many of our RDM workflows on top iRODS, which enabled us to support a high volume of data. For example, data transfers during the ingest process go via Windows network shares, thereby bypassing the web browser. However, the generic nature of iRODS limits its functionality, especially for the needs of the diverse subdomains of the life sciences: it does not allow visualisation or specific file format support.

As an additional solution, we implemented support for domain-specific RDM platforms. Various very successful RDM platforms exist for different life sciences subdomains, for example, XNAT for radiology, OMERO for microscopy and MOLGENIS for molecular genetics [28–30]. These three platforms are successful in their respective domains: i.e., they are providing established community standards and/or methods for researchers to share and manage their research data. While these platforms are regularly setup by research groups on a project basis, long-term support in both funding and expertise is required for continued success.

In implementing these domain-specific platforms, we found that different expertise was required at different levels. DataHub provided support where professional IT skills are required, while the data stewards mainly interacted with the researchers and developed specific workflows on these platforms. We have used this model successfully for XNAT for multiple research groups and are now in the process of implementing it for OMERO.

Obtaining long-term funding is still a challenge for these sorts of local, very domain-specific infrastructural RDM platforms. However, recently, the Dutch Science Foundation (NWO) has started to recognise the need for this as well in their aim to support Open Science. Through their large infrastructure roadmap grant-scheme DataHub has been able to secure funding for supporting OMERO for the next few years.

5.4.3 Offering cross-institutional collaboration

Researchers engaged in international collaborations are functionally limited by software their institute provides (e.g., network shares not accessible by third parties). As a solution, federated access has been developed separately for several RDM platforms. In addition to being very costly, this process also led to

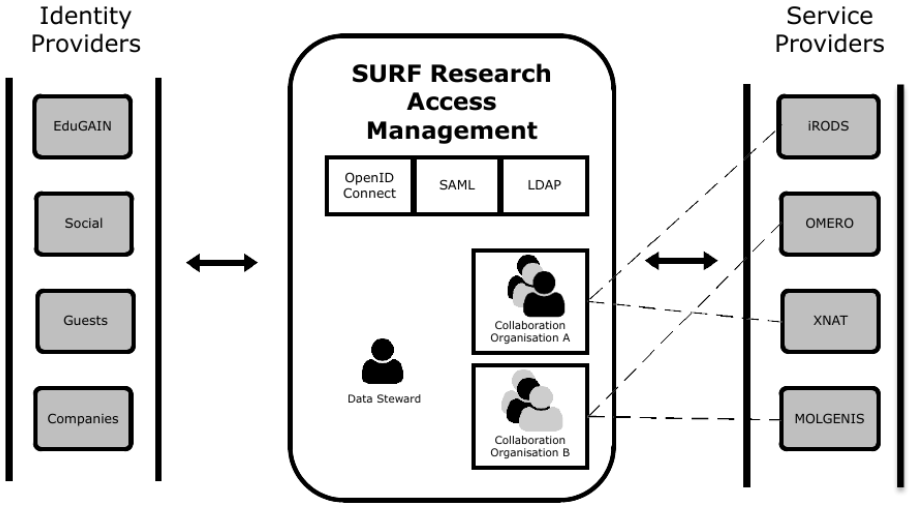


Figure 5.3: SURF Research Access Management (SRAM) is a federated authentication proxy with self-service group management. It proxies an existing identity from, for example, an institutional account via OpenID Connect, SAML or LDAP to a service provider. Users are organised by data stewards into collaboration organisations. Access to service providers can be configured per collaboration organisation.

suboptimal user experiences because users are burdened with learning new interfaces, and/or remembering/storing accounts for all the services they use [31]. Researchers thus often resort to consumer (cloud) sharing applications for which governance does not conform to their institute's policies or national privacy policies.

In response, the European e-infrastructure collaboration GÉANT, among others, has put considerable effort in the development of EduTEAMS to offer generic solutions for federated authentication and authorisation. Several efforts to incorporate these technologies are underway in the Netherlands. One of such is SURF Research Access Management (SRAM) (Figure 5.3), which provides the generic federated authentication proxy through EduTEAMS. In SRAM the collaborative organisation (CO), which can be a research consortium or a research group, is the functional centrepiece of the system. COs are given a digital home in SRAM where CO-members can see the services they collaborate in, while data stewards can manage these members and add services to their CO. DataHub has been collaborating with SURF in the design and implementation of SRAM and has taken SRAM into production as its federated authentication provider and user/group

management for the MDR in 2021. Thereby, secure data collaboration between researchers from trusted identity providers has been realised.

Looking broader, there are several parallel developments taking place in the access and authorisation infrastructures space, notably the Elixir-AAI project, the European Open Science Cloud and Internet2's COManage/CILogon [31, 32]. Making services quickly available for international access and offering simple self-service group management are key features to effectively support research groups engaged in international collaborations, as typically found in the life sciences.

5.5 Conclusion

Our experience shows a few key lessons for effective institutional first-line RDM support for the life sciences. Firstly, the requirements for effective RDM differ considerably between the different life sciences domains, affecting the technology and infrastructure to be provided and preventing generalised solutions. This diversity is reflected and addressed in the use of the disciplinary data stewards, but also in the need to provide professional IT support for very tailored RDM platforms. Secondly, prioritising where/how to provide RDM support is an effective way to deal with limited support resources. At DataHub, we chose to focus on research data that are most reusable. This focus can be achieved by disciplinary data stewards knowing the research projects that have the highest potential reuse of their data. Thirdly, finding incentives to motivate researchers to work on their RDM, with the exception of a financial one, remains a challenging task. We found that offering technological features have little effect in motivating RDM uptake. Our limited success with incentives may lead to the conclusion that implementing mandatory requirements and/or punitive measures for RDM practice may still be required. Finally, RDM must allow for cross-institutional collaboration by default by using the appropriate authorisation and authentication infrastructure.

5.6 Outlook

Looking ahead, we identified several technological and organisational points that will require attention or will play an increasingly important role in the future for institutional RDM support.

5.6.1 Technological

A topic regularly overlooked in RDM is the archival and maintenance of research software [33]. Without the software used to analyse the data, it may be impossible to actually reuse the data. This is still a very challenging topic with many possible pitfalls in its solutions, but it needs more attention. Additionally, definitive metadata schemas, even for a particular domain, do not exist and will likely keep changing over time. Currently, only partial technical solutions, in terms of entry, migration, search and presentation, exist to deal with these ever-changing metadata schemas [34]. We are also exploring and developing this further in the use of flexible metadata schemas in iRODS (Chapter 6).

5.6.2 Organisational

Our findings underline the important role of data stewards to facilitate RDM embedding in the organisational structure. As such, ongoing efforts to professionalise the role of data stewards at institutions in the Netherlands [19] are critical. In particular, training paths and career perspectives should be defined for data stewards. Secondly, the question of who finances the long-term storage of research data needs to be answered clearly. Currently, in Maastricht, there is an ongoing discussion about the benefits of a clear financial model to pay for the long-term storage of research data. One example model is a discounted, one-time, lump-sum payment for data storage at the end of the project. Finally, data stored must undergo a constant form of curation. At the current rate of data growth, published or unpublished, it is impossible to store everything indefinitely. Therefore, constant appraisal and reappraisal of stored research data, with a continued focus on reusable data, should be a priority for institutional RDM. This practice is only feasible with strong RDM in place and in particular better metadata about the data.

5.7 Methods

This paper is based on five years of participant observation by the authors. The authors' findings were further validated by using the results from qualitative semi-structured interviews with key people around DataHub Maastricht. A total of four respondents were interviewed, comprising two long-term DataHub core team members, one Faculty Board member and one data steward. The recordings of the interviews have been deposited in the Maastricht Data Repository and are available upon request [35].

5.8 Acknowledgements

This work has been supported by the Limburg Invests in its Knowledge Economy (LINK) programme from the Province of Limburg, the Netherlands. We would like to thank the respondents of our interviews for their cooperation. We are very grateful to Hang Nguyen and Raimond Ravelli for critically reading this manuscript.

5.9 References

1. Mons, B. *Data Stewardship for Open Science* doi:10.1201/9781315380711 (Chapman and Hall/CRC, 2018).
2. Heidorn, P. B. Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends* **57,** 280–299. doi:10.1353/lib.0.0036 (2008).
3. Darch, P. T. & Sands, A. E. Beyond Big or Little Science: Understanding Data Lifecycles in Astronomy and the Deep Subseafloor Biosphere. http://hdl.handle.net/2142/73655 (2015).
4. Borgman, C. L. *et al.* Data Management in the Long Tail: Science, Software, and Service. *International Journal of Digital Curation* **11,** 128–149. doi:10.2218/ijdc.v11i1.428 (2016).
5. Borgman, C. L. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology* **63,** 1059–1078. doi:10.1002/asi.22634 (2012).
6. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3,** 160018. doi:10.1038/sdata.2016.18 (2016).
7. Higman, R., Bangert, D. & Jones, S. Three camps, one destination: the intersections of research data management, FAIR and Open. *Insights the UKSG journal* **32.** doi:10.1629/uksg.468 (2019).
8. McQuilton, P. *et al.* Helping the Consumers and Producers of Standards, Repositories and Policies to Enable FAIR Data. *Data Intelligence* **2,** 151–157. doi:10.1162/dint_a_00037 (2020).
9. Cox, A. M., Kennan, M. A., Lyon, L. & Pinfield, S. Developments in research data management in academic libraries: Towards an understanding of research data service maturity. *Journal of the Association for Information Science and Technology* **68,** 2182–2200. doi:10.1002/asi.23781 (2017).

10. Tenopir, C. *et al.* Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE* **6,** e21101. doi:10.1371/journal.pone.0021101 (2011).

11. Cox, A. M., Kennan, M. A., Lyon, E. J., Pinfield, S. & Sbaffi, L. Progress in Research Data Services. *International Journal of Digital Curation* **14,** 126–135. doi:10.2218/ijdc.v14i1.595 (2019).

12. Xu, H. *et al.* iRODS Primer 2: Integrated Rule-Oriented Data System. *Synthesis Lectures on Information Concepts, Retrieval, and Services* **9,** 1–131. doi:10.2200/s00760ed1v01y201702icr057 (2017).

13. Chiang, G.-T., Clapham, P., Qi, G., Sale, K. & Coates, G. Implementing a genomic data management system using iRODS in the Wellcome Trust Sanger Institute. *BMC Bioinformatics* **12,** 361. doi:10.1186/1471-2105-12-361 (2011).

14. Clapham, P. *Informatics Support Group – Wellcome Sanger Institute* https://www.sanger.ac.uk/group/informatics-support-group/.

15. Lee, H.-C., Oostenveld, R., van den Boogert, E. & Maris, E. *Neuroimaging Research Data Life-cycle Management* in (2017), 17–19. https://irods.org/uploads/2017/irods_ugm2017_proceedings.pdf#page=17.

16. Staiger, C., Smeele, T. & van Schip, R. *A national approach for storage scale-out scenarios based on iRODS* in (2017), 55–63. https://irods.org/uploads/2017/irods_ugm2017_proceedings.pdf#page=59.

17. Zondergeld, J. J. *et al.* FAIR, safe and high-quality data: The data infrastructure and accessibility of the YOUth cohort study. *Developmental Cognitive Neuroscience* **45,** 100834. doi:10.1016/j.dcn.2020.100834 (2020).

18. Crosas, M. The Dataverse Network®: An Open-Source Application for Sharing, Discovering and Preserving Data. *D-Lib Magazine* **17.** doi:10.1045/january2011-crosas (2011).

19. Jetten, M. *et al.* Professionalising data stewardship in the Netherlands. Competences, training and education. Dutch roadmap towards national implementation of FAIR data stewardship. doi:10.5281/zenodo.4623713 (2021).

20. Versweyveld, L. We need 500.000 respected data stewards to operate the European Open Science Cloud - News blog - e-Infrastructures Reflection Group. http://e-irg.eu/news-blog/-/blogs/we-need-500-000-respected-data-stewards-to-operate-the-european-open-science-cloud (2016).

21. Teperek, M., Cruz, M. J., Verbakel, E., Bohmer, J. & Dunning, A. Data Stewardship addressing disciplinary data management needs. *International Journal of Digital Curation* **13,** 141–149. doi:10.2218/ijdc.v13i1.604 (2018).

22. Getler, M., Sisu, D., Jones, S. & Miller, K. DMPonline Version 4.0: User-Led Innovation. *International Journal of Digital Curation* **9,** 193–219. doi:10.2218/ijdc.v9i1.312 (2014).

23. Schwaber, K. & Beedle, M. *Agile software development with Scrum* 158 (2002).

24. Wilms, K. L., Stieglitz, S., Ross, B. & Meske, C. A value-based perspective on supporting and hindering factors for research data management. *International Journal of Information Management* **54,** 102174. doi:10.1016/j.ijinfomgt.2020.102174 (2020).

25. Doucette, L. & Fyfe, B. *Drowning in Research Data: Addressing Data Management Literacy of Graduate Students* in (2013). http://hdl.handle.net/11213/18087.

26. Abrams, S., Kratz, J., Simms, S., Strong, M. & Willett, P. Dash: Data Sharing Made Easy at the University of California. *International Journal of Digital Curation* **11,** 118–127. doi:10.2218/ijdc.v11i1.408 (2016).

27. Ashiq, M., Usmani, M. H. & Naeem, M. A systematic literature review on research data management practices and services. *Global Knowledge, Memory and Communication* **71,** 649–671. doi:10.1108/gkmc-07-2020-0103 (2022).

28. Marcus, D. S., Olsen, T. R., Ramaratnam, M. & Buckner, R. L. The extensible neuroimaging archive toolkit. *Neuroinformatics* **5,** 11–33. doi:10.1385/ni:5:1:11 (2007).

29. Li, S. *et al.* Metadata management for high content screening in OMERO. *Methods* **96,** 27–32. doi:10.1016/j.ymeth.2015.10.006 (2016).

30. Velde, K. J. V. D. *et al.* MOLGENIS research: Advanced bioinformatics data software for non-bioinformaticians. *Bioinformatics* **35,** 1076–1078. doi:10.1093/bioinformatics/bty742 (2019).

31. Linden, M. *et al.* Common ELIXIR Service for Researcher Authentication and Authorisation. *F1000Research* **7,** 1199. doi:10.12688/f1000research.15161.1 (2018).

32. Basney, J. *et al. CILogon: Enabling Federated Identity and Access Management for Scientific Collaborations* in. **351** (Sissa Medialab, 2019), 031. doi:10.22323/1.351.0031.

33. Howison, J. & Bullard, J. Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology* **67,** 2137–2155. doi:10.1002/asi.23538 (2016).

34. Philipson, J. The Red Queen in the Repository. *International Journal of Digital Curation* **15,** 16. doi:10.2218/ijdc.v15i1.646 (2020).

35. Schayck, J. P. v. *Background interviews for "First line research data management for the life sciences: a case study"* (2021). https://hdl.handle.net/21.12109/P000000190C000000001.

6

Providing validated, templated and richer metadata using a bidirectional conversion between JSON and iRODS AVUs

Adapted from: **Van Schayck, J. P**, Smeele, T., Theunissen, D. & Westerhof, L. Providing validated, templated and richer metadata using a bidirectional conversion between JSON and iRODS AVUs. *iRODS User Group Meeting 2019 Proceedings*, 9–17. `https://irods.org/uploads/2019/irods_ugm2019_proceedings.pdf` (2019).

Abstract

A frequently recurring question in research data management is to structure metadata according to a standard and to provide the corresponding user interface to it. This has only become more urgent since the introduction of the FAIR principles which state that metadata should use controlled vocabularies and meet community standards.

The iRODS data grid technology is well positioned as a core layer within an infrastructure to manage research data. One of its strengths is the ability to attach any number of attribute, value, unit (AVU) triples as metadata to any iRODS object. This makes iRODS adaptable to very diverse use cases in research data management. However, the challenge of working with more structured metadata is not being addressed by the default capabilities of iRODS. Our aim is to develop a new method for storing richer, templated and validated metadata in AVUs.

JSON is a popular, flexible and easy to use format for serialising (nested) data, while maintaining human and developer readability. Furthermore, a JSON Schema can be used to validate a JSON structure and it can also be used to obtain a dynamically generated form on the basis of this schema. This combination of functionalities makes it an excellent format for metadata. We have therefore designed and implemented a bidirectional conversion between JSON and AVUs. The conversion method has been implemented as Python iRODS rules that allow to set and retrieve AVU metadata on an iRODS object using a JSON structure. Optionally, a policy can be installed to validate metadata entry and updates against the JSON Schema that governs the object.

With this work we provide other iRODS developers with a generic method for conversion between JSON and AVUs. We are encouraging others to use the conversion method in their deployments.

6.1 Introduction

Over recent years there has been an increasingly urgent call for the practice of Open Science [1]. Driven by the core values in research of transparency and reproducibility, the sharing of research data has become a hot topic. Additional reasons for sharing research data are to better leverage investments in research by promoting reuse and making those data that can be considered public assets, available to the public [2]. To facilitate this accountability and reuse of research data, the findable, accessible, interoperable and reusable (FAIR) principles of research data have been introduced and broadly taken up [3]. Briefly, the FAIR principles suggest for research data to be globally and uniquely identifiable, and associated with searchable metadata ('findable'); these identifiers should point to (meta)data using an open protocol ('accessible') and that this data uses a formal representation language using widely applicable ontologies ('interoperable'); finally, data should be provided with cross-references, provenance and license information ('reusable'). Furthermore, the FAIR principles state that all this should be provided in both human and machine-readable form to facilitate automated pipelines and the increasing need for automated analysis and large-scale data research. However, the FAIR guidelines did not define any form of implementation recommendations for these principles.

The iRODS data grid technology is well-positioned as a core layer within an infrastructure to manage research data [4]. iRODS features support for preservation properties that are important for research data management such as authenticity, integrity and chain of custody. Most of these properties are met using metadata to annotate data objects and collections. Within iRODS, a basic building block for managing metadata is the AVU, short for *attribute-value-unit*. It consists of three string-typed fields that as a triple represent a single metadata property of the data. While typically more than one AVU is needed to communicate and to document properties of research data[1], currently the iRODS support for AVU *composition* is somewhat limited. Microservices operate either on a single AVU or on an attribute-value map structure that does not allow a unit component to be specified. Attributes with multiple values, nested structures and atomic operations on a composition of AVUs are not supported. There is also no way to specify a template or structure that AVUs associated to an object should adhere to. These limitations may hinder the implementation of the FAIR principles or could lead to *ad hoc* solutions that are not transferable to other systems.

In contrast, many applications have adopted the data interchange standard JavaScript Object Notation (JSON) to efficiently exchange and operate on composed data structures [5]. JSON is lightweight and can be used across many

[1] see for instance the DataCite specification at `https://schema.datacite.org`

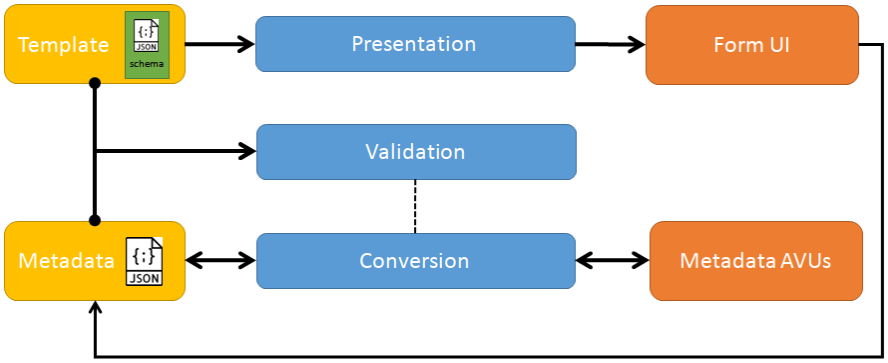


Figure 6.1: Overview of the schematic layers between JSON, AVUs, JSON-schema and its process in conversion, validation and presentation.

programming languages. JSON data structures consist of an unordered collection of name/value pairs referred to as an *object*. Values are primitive data types, or an ordered list of values, or another JSON object. To improve interoperability, in 2017 IETF has published a more restrictive version of the JSON standard [6]. For instance, this version requires the name of name/value pairs to be unique, so that programming languages can conveniently implement JSON using map constructs.

JSON can be used to serialize arbitrary metadata, resulting in an equally arbitrary set of AVUs. For the purpose of adhering to the FAIR principles however, we seek to restrict the metadata that documents research data to a well-defined set of composed, related and consistent properties. The semantics of this metadata structure can be modeled as a *template*. Such a template can be applied to validate operations that attempt to modify any of the AVUs within the *namespace* of the template.

We propose to use JSON Schema as a technique to represent a metadata template within the context of iRODS. The metadata template will act on a composition of AVUs that is represented by a JSON data structure. JSON Schema is a draft internet standard that aims to define the structure and content of JSON formatted data [7]. Using a JSON Schema definition, applications such as iRODS can validate and interact with instances of JSON formatted data.

The application of a JSON Schema based template does not need to be limited to iRODS server-side validation. Figure 6.1 shows how client applications can opt to use the same JSON Schema in a presentation layer to *dynamically* render

forms that facilitate data entry of metadata. For instance the React[2] component react-jsonschema-form[3] is used by Utrecht University, The Netherlands to render metadata entry forms in its data management application Yoda [8]. DataHub at MUMC+ and Maastricht University, The Netherlands seeks to implement a similar solution based on the CEDAR Workbench [9]. DataHub's use case is focused on semantically linked (meta)data. Therefore not only should the JSON structure be governed by a JSON Schema, in addition its vocabulary and structure must conform to the W3C JSON-LD recommendation [10].

Hence, our research can be applied on three levels. At the foundation level, a conversion method supports the use of JSON to manage arbitrary compositions of AVUs in iRODS and to exchange these compositions efficiently with client applications (Methods section). The optional second level validates the exchanged JSON against a metadata template defined in JSON Schema. Both the first and second levels are implemented in the iRODS server (Results section). A third level can optionally be implemented as part of a client application. It uses the (same) metadata template for dynamic form-based user interactions. In the Discussion section, the main advantages and disadvantages we found using the proposed methods are discussed. Finally, the research results are summarized and an outlook is provided in the Conclusion section.

6.2 Methods

6.2.1 Bidirectional conversion between JSON and AVU structures

We intend to represent a set of iRODS AVUs as a JSON structure and vice versa and use the serialized data in communications between the iRODS server and client applications.

Before creating the conversion method we set the following five design goals.

1. The conversion method must be a bijective function to ensure lossless conversions between JSON and AVU structures in both directions. Any JSON structure that is compliant with the JSON specification should be supported. The method should provide support for Unicode characters, nested structures, and ordered lists.

2. It must be easy to identify corresponding JSON objects and AVU attributes. This means that, especially for simple JSON structures, it should be trivial to retrieve a JSON element from the AVUs without first back-converting the JSON.

[2] `https://reactjs.org`
[3] `https://github.com/rjsf-team/react-jsonschema-form`

```
{
  "title": "Hello World!",
  "parameters": {
    "size": 42,
    "readOnly": false
  },
  "authors": ["Foo", "Bar"],
  "references": [
    {
      "title": "The Rule Engine",
      "doi": "1234.5678"
    }
  ]
}
```

Listing 6.1: Example of a JSON structure

3. The conversion method should be lean and efficient. We seek to avoid an explosion in the number of AVUs as a result of representing a nested JSON structure.
4. The method should be compatible with existing use cases that operate directly on AVUs.
5. The conversion method should be compatible with JSON-LD use cases.

Using the design goals set as requirements we arrived, over several iterations, at a working design for the conversion. An example JSON structure and its converted counterpart in AVUs are listed in Listing 6.1 and Table 6.1. This example will be used to explain how the conversion method works. Further examples can be found in the online repository[4].

The proposed conversion method repurposes the unit field of the AVU to encode JSON variable type and structure information. This also reduces the chance of collisions with existing AVUs that presumably have an empty unit field. Currently, nearly all the iRODS microservices that facilitate AVU operations, for example `msiAssociateKeyValuePairsToObj`, do not allow rule developers to specify content for the unit field. As a result of this limitation, the unit component of the AVU has rarely been used in an operational implementation of iRODS.

The AVU unit field comprises of four components, separated by an underscore character, except for the last component where a hash is used. The first component

[4] `https://github.com/MaastrichtUniversity/irods_avu_json`

Table 6.1: Conversion of the JSON structure of listing 6.1 in its counterpart AVUs.

Attribute	Value	Unit
title	Hello World!	root_0_s
parameters	o1	root_0_o1
size	42	root_1_n
readOnly	False	root_1_b
authors	Foo	root_0_s#0
authors	Bar	root_0_s#1
references	o2	root_0_o2#0
title	The Rule Engine	root_2_s
doi	1234.5678	root_2_s

indicates a *namespace* carried by all the AVUs that belong to this set. Conversion operations will affect only AVUs that are part of the selected namespace. In addition, it facilitates that iRODS objects are annotated with multiple JSON structures, each identified by their own namespace. In example Table 6.1 the namespace is `root`.

The second component is an *object sequence number* that keeps track of AVUs that are part of the same JSON object. The top level JSON object is assigned sequence number '`0`'. In the example this includes `title`, `parameters`, `authors` and `references`. Note that the element `parameters` holds a nested object as its value. The next sequence number '`1`' is assigned to this nested object and note the sequence number in its (otherwise unused) AVU value component, prefixed by the character `o`. All elements of the nested object, in this example `size` and `readonly`, have the object sequence number '`1`' in their unit field.

An important design goal of the conversion method is the support of different variable types within JSON. AVUs only allow string values, while JSON supports various primitive types. The third component of the AVU unit field is used to indicate the JSON *type* of the value. See table 6.2 for an overview of the supported types. A special case is the `empty array` type that indicates the presence of an array without any members. To achieve a lean conversion, we only create AVUs to represent *members* of an array. We have to make an exception for an empty array, which otherwise would not have any AVU representation at all. Without this provision, a later conversion from AVU back to JSON would not be able to recreate the empty array.

The fourth, optional component of the AVU unit field is used to denote an *ordered index* of the element. This component is separated from its predecessor using a hash character `#`. The JSON specification includes the array type. This is an ordered list of elements of any type. As AVUs are unordered, the last component

Table 6.2: Overview of JSON variable types and their corresponding type string.

Type	Unit	Value	Remarks
string	s	The literal string	
object	o + id	o + object_id	The AVU-value field is not used
boolean	b	'True' or 'False'	
number	n	Float or int	Converted to string
null	z	'.'	AVUs do not allow empty values
empty string	e	'.'	AVUs do not allow empty values
empty array	a	'.'	For convenience during conversion

of the unit field denotes the array index to maintain order in arrays.

Summarising, the AVU unit field has been used for the following purposes: 1. defining the JSON namespace, 2. the object sequence number, 3. the value type and 4. the array index. A regular expression to capture these components of the unit field is shown in the Listing 6.2.

```
^([a-zA-Z0-9_])+_([0-9]+)_([osbanze])((?<=o)[0-9]+)?((?:#[0-9]+?)*)
```

Listing 6.2: A regular expression to parse the components of the unit field

6.2.2 Validation of JSON structure using a JSON Schema template

The conversion method discussed above allows client applications to store JSON structures efficiently within the iRODS server. This method is agnostic to the structure and the semantics of the metadata that is exchanged. For some use cases this may suffice. Many use cases, however, require that metadata stored or exchanged is compliant with a certain standard. We will now propose a validation method to fulfil this need.

The validation method must meet the following three design goals.

1. It must be able to assess that a *stored or exchanged set* of metadata meets predefined quality levels with respect to structure and semantics as typically documented in metadata standard. The metadata schema can vary per iRODS object. For instance, objects that belong to a data set related to the Geosciences may require geospatial location annotations whereas objects related to History disciplines may depend on chronological classification metadata.

2. It should also be possible to annotate a single object with multiple disjunct sets of metadata. The metadata template can vary per set of metadata.

3. In line with the conversion method, the validation method should again be compatible with existing use cases that operate directly on individual AVUs.

For the purpose of validation, we shall consider sets of metadata that annotate an iRODS object rather than individual AVUs. These sets are easily identified by their namespace identifier which is incorporated in the unit component of the AVU. This makes the validation compatible with existing use cases as AVUs without a namespace will not be affected by the validation and protection policies.

We select the draft internet standard JSON Schema to specify a metadata template used to validate sets of metadata [7]. JSON Schema conveniently supports the description of quality properties of individual metadata elements as well as qualities that span across elements, for instance dependency relationships. Incoming metadata will be in JSON representation and can be validated directly against a template.

6.3 Results

6.3.1 Conversion method implementation

The implementation has been developed for iRODS version 4.2.x [11]. Since iRODS 4.2 does not expose any microservices to modify the unit field of an AVU triple, custom iRODS microservices have been developed. The conversion scheme has been developed as Python 2.7 and Python 3 module[5] named `irods_avu_json`. The outcome of the conversion of the example JSON is shown in Table 6.1.

The `irods_avu_json` module has in itself no interaction with or dependency on iRODS. To expose this functionality within an iRODS installation we developed an iRODS ruleset [6]. The developed ruleset uses the Python Rule Engine recently released with iRODS 4.2 to import the irods_avu_json module.

An overview of the functionality being exposed by the ruleset is summarized in Table 6.3. All ruleset functions allow specifying any iRODS object type (collection, object, user, group or resource) in a similar fashion as that iRODS AVUs can be attached to any iRODS object. Furthermore, all functions also expect the JSON namespace to know which AVU set to operate on.

[5] `https://github.com/MaastrichtUniversity/irods_avu_json`
[6] `https://github.com/MaastrichtUniversity/irods_avu_json-ruleset`

6.3.2 Validation method implementation

The special `$schema` AVU denotes whenever a JSON Schema is attached to an iRODS object. We implemented two ways for the JSON Schema to be specified in the value field of this AVU. (1) An (public or private) URI pointing to the stored JSON schema. Optional caching of this URI has been implemented for performance reasons. (2) By specifying 'i:' in front of a path the JSON Schema is directly retrieved from within iRODS. Care must be taken that this iRODS object is accessible for anyone allowed to modify the iRODS object to which the JSON Schema is attached. Other methods for storing the JSON Schema could be devised and implemented at a later point.

Note that both the iRODS server and client applications can benefit from using the metadata template to check the validity of any metadata that is being exchanged. Therefore a best practice is that the template reference is an absolute URI and the metadata template itself is available at an internet-accessible location.

Validation of the JSON structure set by `setJsonToObj()` is triggered by the presence of the `$schema` attribute and the same JSON namespace being present on the iRODS object. After retrieving the JSON Schema contents, the validation is performed using the `jsonschema` Python module. Any validation errors will be passed back to the caller of `setJsonToObj()`. To ensure only the full and validated JSON object is being stored first all existing AVUs of the same JSON namespace are removed before the new one are set. This also ensures that any no longer existing parts of the JSON object are removed during a `setJsonToObj()` operation.

The irods_avu_json-ruleset further implements validation of a JSON object by implementing a policy enforcement points (PEPs), which are executed during the modification of AVUs. Whenever a `$schema` AVU is present on the iRODS object and the AVU being modified is part of the same JSON namespace the operation is disallowed. Modification of these AVUs can only be performed through `setJsonToObj()`. Because `setJsonToObj()` would also trigger the same PEPs, the PEPs check whether execution is coming from `setJsonToObj()` and has been validated against the JSON Schema. This is achieved through setting a Python global variable that is preserved in the rule engine memory.

6.4 Discussion

We successfully used the conversion method and the accompanying validation method in several pilot use cases. We found several limitations and problems with the current implementation of the conversion and validation method.

The implementation of the conversion method currently requires all existing

Table 6.3: Functionality exposed by the irods_avu_json-ruleset

Function	Description
`setJsonToObj`	Set a JSON to an iRODS object.
`getJsonFromObj`	Retrieve a JSON from an iRODS object.
`setJsonSchemaToObj`	Attach a JSON Schema to an iRODS object.
`getJsonSchemaFromObj`	Retrieve a JSON Schema from an iRODS object.

AVUs relating to a JSON namespace to be removed before the new JSON is added. This may lead to performance issues with very large JSON structures. Furthermore, the addition of all JSON related AVUs is not an atomic operation, meaning that collisions may occur if multiple clients modify the same iRODS object at once.

The current implementation of validation uses the metadata PEPs to prevent any non-validated AVUs to be created or modified. These PEPs directly wrap around their AVU modification microservice counterparts. Therefore a single microservice call can, using a wildcard, operate on multiple AVUs. This means logic created for the PEPs is rather convoluted. Furthermore, the chosen implementation of using a global Python variable to bypass the PEPs when `setJsonToObj()` is being called breaks when the call for `setJsonToObj()` is initiated from a catalog consumer instead of the catalog provider. This is because in that case of the call being initiated from the catalog consumer the global variable is not in memory when the PEP is being executed on the catalog provider.

The performance issue, the non-atomic operation of the current conversion and the issue with the PEPs can all be tackled by the introduction of a multi-AVU atomic core iRODS functionality. Such a microservice could handle the entire operation of converting a JSON structure and its validation at once.

While not directly shown in this work the use of JSON Schema can be extended beyond the validation level that is currently implemented. As mentioned before, an important feature is the auto-generation of web forms from the JSON Schema. Several implementations of libraries capable of this functionality exists and we have explored several of those. Furthermore, we envision that the use of JSON Schema for presentation can be further extended to for example auto-generated search forms.

6.5 Conclusion

We set out to add a generic toolset to iRODS for handling richer, templated and validated metadata. We have developed a bidirectional conversion between

JSON and AVUs and validation provided through JSON Schema. Both methods have been validated to meet their design goals via a proof of concept followed by an application-level implementation. The methods developed provide new starting points for iRODS developers in facilitating research data management that adheres to the FAIR principles.

6.6 Acknowledgements

This work was sparked by the discussions in the iRODS metadata template working group and we would like to thank them for their feedback. We would like to thank all members of the DataHub Maastricht team for their helpful feedback and comments. Finally, we thank Raimond Ravelli, Peter Peters and Michel Dumontier for their critical reading of this manuscript.

6.7 References

1. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nature Human Behaviour* **1,** 0021. doi:10.1038/s41562-016-0021 (2017).
2. Borgman, C. L. *Big data, little data, no data: Scholarship in the networked world* (MIT press, 2015).
3. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3,** 160018. doi:10.1038/sdata.2016.18 (2016).
4. Moore, R. Towards a Theory of Digital Preservation. *International Journal of Digital Curation* **3,** 63–75. doi:10.2218/ijdc.v3i1.42 (2008).
5. ECMA. ECMA-404: The JSON Data Interchange Syntax. *Ecma International* **Standard** (2017).
6. Bray, T. IETF RFC 8259: The JavaScript Object Notation (JSON) Data Interchange Format. *Internet Engineering Task Force (IETF)* (2017).
7. Andrews, H. & Wright, A. JSON Schema A Media Type for Describing JSON Documents. *Internet Engineering Task Force (IETF)* **Internet-Draft.** https://tools.ietf.org/pdf/draft-handrews-json-schema-01.pdf (2018).
8. Smeele, T. & Westerhof, L. Using iRODS to manage, share and publish research data: Yoda. *iRODS User Group Meeting 2018 Proceedings.* https://irods.org/uploads/2018/irods_ugm2018_proceedings.pdf (2018).

9. Gonçalves, R. S. *et al.* The CEDAR Workbench: An Ontology-Assisted Environment for Authoring Metadata that Describe Scientific Experiments. *Lecture Notes in Computer Science* **10588,** 103–110. doi:10.1007/978-3-319-68204-4_10 (2017).

10. Sporny, M., Longley, D., Kellogg, G., Lanthaler, M. & Lindström, N. JSON-LD 1.0. *W3C Recommendation* **16,** 41 (2014).

11. Rajasekar, A., Moore, R., Wan, M. & Schroeder, W. Policy-based Distributed Data Management Systems. *Journal of Digital Information* **11.** https://journals.tdl.org/jodi/index.php/jodi/article/view/756 (2010).

7
General Discussion

In this thesis, I have looked at two different topics. The first topic is the design, integration, optimisation, and application of the Timepix3 as a digital recording device for low-dose cryo-electron microscopy (cryo-EM). The second topic examines the management of life sciences research data and the organisation thereof. The results and outlooks obtained will be discussed below in the same order and concluded with some final remarks.

7.1 The characterisation, optimisation and integration of the Timepix3 as detector for cryo-EM

At the time of this writing, the highest resolution obtained using cryo-EM single particle analysis (SPA) is a 1.22 Å reconstruction of mouse apoferritin [1]. While this specific record will undoubtedly be broken again in the near future, it is interesting to look at how this specific record was achieved. Nakane and his colleagues at the MRC Laboratory of Molecular Biology (LMB) in Cambridge used a cold-field emission gun (CFEG) operating at 300 kV. The CFEG has a very small energy spread of 0.3 eV, reducing the impact of chromatic aberrations of the objective lens. A Falcon4 direct electron detector was used, with a built-in energy filter. The energy filter, in combination with the small initial energy spread of the CFEG, allowed the inelastically scattered electrons to be removed, which could deteriorate the signal-to-noise ratio (SNR) of the image. In their outlook, they note that they are unsure if the SNR improvement can be attributed solely to the removal of inelastically scattered electrons or whether other effects, for example, increased amplitude contrast, play a role. To what extent is it simply an effect of using 300 kV acceleration voltage?

Looking back at previous SPA resolution records [overview in 2], the highest resolution records were made, without exception, at an acceleration voltage of 300 kV. This fact is surprising, as there are several indications that operating at 100 kV could provide significant benefits over operating at 300 kV. These benefits are the improved ratio of elastic to inelastically scattered electrons as well as the increased amplitude contrast [3–6]. The first should effectively lead to less radiation damage, while the second makes the particles more visible. These benefits are in particular true when the sample is thin enough ($\ll$ 100 nm), e.g. not much thicker than the thickness of particles, such that multiple scattering of electrons is avoided. In addition to these benefits in imaging, 100 kV microscopes are cheaper to manufacture and operate. These benefits and practical aspects of cryo-EM at 100 kV have been explored in recent years both theoretically and practically [7, 8].

Although using a lower acceleration voltage such as 100 kV has numerous ben-

efits, it begs the question: why does an acceleration voltage of 300 kV continue to produce record-breaking resolutions? Typical cryo-EM SPA samples, such as apoferritin, are 30 to 50 nm thin, therefore perfectly adequate for the use of energies < 300 kV. The effect of chromatic aberrations can be larger at 100 kV, as electron guns have a certain absolute energy spread, which becomes, in relative terms, more prominent at lower energies. Could it be due to energy-dependent charging effects, causing (apparent) motion of the sample early on in the exposure? The sole more obvious reason is the performance of the widely used commercially available backthinned MAPS detectors: their performance is best at 300 kV. At lower energies, MAPS detectors suffer from more noise due to backscattering of electrons in the epilayer. This backscattering cannot be counteracted by further backthinning as this would make the detector too fragile. While there may be other reasons why the best resolution SPA reconstructions have thus far been obtained at 300 kV rather than at lower energies, the role that the backthinned MAPS detectors play in this – optimised for 300 kV – is the least disputed.

Direct electron detectors that are based on a different technology than backthinned MAPS detectors have the potential to overcome this limitation on energy. In Chapter 2 we have characterised and optimised the use of the Timepix3 Hybrid Pixel Detector (HPD) as a digital recording device for 200 kV cryo-EM workflows (as well as 300 kV). Our characterisations, which included both simulations and experiments, corroborate some of the earlier findings with the use of hybrid pixel detectors in EM [9–16]: (1) the intrinsic radiation hardness of HPDs allows for experiments with high electron flux, and (2) a sufficiently energised incident electron will spread its charge into several pixels, thus degrading the MTF of the detector. We set out to develop methods to better exploit the first, while mitigating the second.

Our results in Chapter 2 showed that a convolutional neural network trained on simulated Timepix3 data can be used to accurately localise the impact position of an electron with less than a pixel uncertainty on the sensor layer of a Timepix3 detector. By using the spectroscopic per-pixel properties of the Timepix3 we were able to deduct more information of the incident electron trajectory, and model the most probable impact pixel, with sub-pixel accuracy. This method significantly improved the MTF of the detector.

In Chapter 3 we show how we integrated Timepix3 into a fully automated cryo-EM workflow. We demonstrated this by resolving a protein to 3.0 Å resolution in a SPA workflow. The Timepix3 detector ran, without human intervention, for 72 h, collecting data from approximately 3000 foil holes and acquiring ~30.000 navigational images.

The results of Chapter 2 and 3 combined show the viability of the use of HPDs, both for 200 kV and 300 kV cryo-EM workflows. However, they also showed

several areas that need improvement before HPDs could support a larger variety of workflows. Let me highlight some areas where improvements could be made.

The resolution of 3.0 Å was less than what we achieved when using the commercial Falcon3 (2.5 Å, Table 3.1) on the same protein sample and microscope. Several reasons can be given for this difference. Using simulations, we showed that the use of a CNN trained on both ToT and ToA data performed best. However, for experimental data, the improvement in MTF when using both ToT and ToA compared to using ToT alone (Figure 2.9), was very small. Furthermore, the non-CNN method Highest-ToA was also close in performance. This difference in performance of the CNN between the simulations and the experimental data could be related to (1) a detrimental effect caused by the systematic row and column patterns seen in the ToA data (Figure 2.12), and (2) the ToA simulation is not accurate enough and too different from the experimental data. In general, it could be expected that the training of the neural networks using simulated data only had compromised the localisation accuracy of our method and thereby the maximum achievable resolution.

We have experimented with training the neural network directly with experimental knife-edge data (Dirk Bongers, unpublished results). In our hands, doing so accurately proved to be difficult, primarily because the training information, i.e. an edge of pixels where the beam is blocked, differs from the desired output information of the model, i.e. where the electron has hit the pixel. We tried to solve this discrepancy using a custom loss function for training of the neural network, but never arrived at a model outperforming the neural network trained on simulated incident position data. It remains unclear whether this result was due to the method being unsuitable or our implementation thereof lacking. Ideally, it would be possible to train our neural networks directly on experimental data. For example, we could move a subpixel-sized beam over the Timepix3 pixels with some *a priori* knowledge position of the beam on those pixels. Such experiments were not possible with our setup, as we lacked the necessary STEM unit on our microscope.

Another area of improvement is the deterministic blur, described in Chapter 3 and implemented somewhat pragmatically. Thermo Fisher seems to apply deterministic blur in their electron-counting algorithms for the Falcon detectors, whereas it is not used by Gatan for the K2/K3 detectors. Little information can be found in the literature — in particular within the EM literature — that describes the fine details of using deterministic blur. Consequently, we feel that there is not a solid enough foundation for employing this properly. The DQE curves we present assume a certain DQE(0), as it was practically difficult to perform the accurate Faraday cup measurements at very low currents, which would be needed to obtain an experimental DQE(0).

The SPA resolution we obtained was also restricted due to the relatively small number of pixels in our detector. The Timepix3 reconstruction presented in Chapter 3 used ∼10k particles, compared to ∼124k for the Falcon3 reconstruction. We had used a Timepix3 detector in a quad configuration, which had an active area of 3×3 cm^2. This configuration gave 512×512 active pixels and only 50×50 nm field of view at a typical pixel size of 1 Å. This resulting area is very small compared to commercial detectors such as Falcon3 and K3. The Timepix3 chips cannot be tiled together to larger configurations than a quad, as otherwise the size of dead areas would become excessive. Increasing the number of effective pixels using super resolution (1024 × 1024) proved to be possible (Chapter 3), but still leaves the Timepix3 detector far smaller with respect to the number of pixels compared to its commercial imaging competitors.

Looking ahead, some of these improvements should come from Timepix4, which became available at the end of 2021 [17]. Timepix4 offers a maximum of 2.5 Ghit s^{-1} with a size of 448×512 pixels (compared to 120 Mhit s^{-1} for a Timepix3 in quad configuration). Additionally, the Timepix4 can be tiled seamlessly, by utilising a through-silicon via (TSV) connection to let the data channels pass through the wafer of the chip instead of occupying space next to the active area of the chip. This configuration eliminates the dead space between chips and can greatly increase the field of view compared to Timepix3. At the time of writing, there is on-going work at Maastricht University to characterise and test the Timepix4.

We believe, like others, in the good prospects for further 100 keV SPA experiments. Our integrated setup (Chapter 3) is unique in that Timepix3 allows for such experiments. Full-scale SPA experiments, to the best of our knowledge, have not been conducted yet on Timepix- or Medipix-family chips in the 20 to 120 keV energy range. A pixel pitch of 55 μm would still lead, for a 60 kV or higher electron, to the sharing of charge between pixels [14]. The event-localisation method, as described in this thesis, may work as long as each incident electron produces hits in multiple pixels; however, the training of CNN model would become more and more difficult alongside the smaller the number of pixels used to localise an individual event.

Still in the design phase, but to become available in the next few years will be Medipix4. Similar to the Timepix4, this chip will feature TSV connectors making it tileable without dead space. As currently designed, Medipix4 will have a 75 μm pixel pitch and feature, just like Medipix3, a Charge Sharing Mode (CSM) to electronically combine the signal shared between several pixels. Paton *et al.* showed in their work with Medipix3 that the combination of CSM and higher Z sensor material, such as gallium-arsenide (GaAs), has much potential [15]. CSM is limited to summing the charge of the directly surrounding pixels,

but when complemented with the limiting effect of the higher-Z material on charge spreading, this combination proved to be give near-ideal DQE and MTF for energies well above 100 kV. The added ease of operating and integrating a frame-based readout of the Medipix series (compared to the data-driven Timepix series) may make Medipix4 with a GaAs sensor layer an ideal combination for a multitude of 100 to 200 kV cryo-TEM workflows.

In recent years, several other developments have also taken place for MAPS-based EM detectors. The German-based company TVIPS develops MAPS sensors coupled to a scintillator layer, which deteriorates the MTF and DQE of the detector, but the coupling performs very well for electron diffraction workflows that benefit from the higher read-out speed and radiation hardness. In 2023, the UK-based company Quantum Detectors announced the Quantum C100: a MAPS detector with 50 μm pixel pitch and 2048 × 2048 pixels. Such a setup has the benefit of a larger field of view. The large pixel size should allow for good DQE and MTF, which possibly makes it an excellent choice for 100 kV SPA experiments, which do not require radiation hardness. At the time of writing, no published performance measures were available for the Quantum C100.

During our work to integrate the Timepix3 (Chapter 3), in an attempt to increase throughput, we tried to record in five separate spots within a 2 μm hole of a sample grid. This setup led one of the recording spots to have the total beam area well within the hole, without touching the supporting foil layer. This led to quite severe (apparent) motion of the particles visible in the micrographs. Surprisingly, micrographs collected in the same hole directly before or after the distorted micrograph did not show severe motion of the particles. This observation was the initial step of our work of Chapter 4, in which we show that the addition of a graphene coating can alleviate this (apparent) motion. Overall, our results indicate that the apparent motion is not due to a plastic formation (referred to as 'doming') of the sample, but due to a micro-lensing effect.

Our results in Chapter 4 also indicated that early-stage rapid sample motion was not reduced by the addition of the graphene layer (Figure 4.5A). This rapid sample motion has been speculated to be due to the release of mechanical stress introduced during the freezing of the sample [18–20]. It has also been related to radiation damage occurring in the first few microseconds of the exposure. Even though faster detectors, such as the Timepix3 and Timepix4, are now available, it will be difficult to assess what happens to the sample in these early microseconds using SPA alone. Fundamentally, the signal generated in SPA will always need to overcome the shot noise limit to be visible. To answer those questions, we need to look beyond SPA cryo-EM.

7.1.1 Looking beyond cryo-EM SPA

One of the important limits of cryo-EM is radiation damage done to the sample by the electron beam. This means that techniques that are able to more efficiently create signal per amount of dose they deposit in the sample have a great advantage. Scanning transmission electron microscopy (STEM), which involves a small convergent beam scanning through the sample, can offer several of such advantages. Compared to TEM, STEM can have some inherent properties making it more dose efficient [21, 22]. First, whereas in TEM the resolution limit is caused by an inserted aperture and lens aberrations, in STEM the semi-angle of the focused beam controls the apparent resolution and the depth of focus. When aberrations are present, they will only affect the contrast transfer function (CTF) shape but will not limit the resolution. Second, unlike TEM imaging methods that produce additional phase contrast by defocusing the specimen, creating additional CTF correction artefacts in the process, STEM techniques reach the highest contrast of the image in focus. When combined, STEM can be an attractive, more dose-efficient alternative to SPA cryo-EM.

However, traditional angular dark field or bright field STEM still suffers from dose inefficiency because the total number of electrons collected in the dark field is several orders of magnitude smaller than in the bright field. Several alternative approaches using STEM are under investigation for application in cryo-EM. One of them is integrated differential phase contrast (iDPC). In this technique, the centre of mass (COM) of the convergent beam on the detector plane is recorded, and integrated, while scanning the sample [23]. From the resulting data the phase information can be retrieved, and a micrograph reconstructed. Recently, Lazić *et al.*[24] achieved several sub-nanometer reconstructions of biological macromolecules using iDPC. Their in-focus micrographs required no further CTF correction before being used for further SPA processing.

Another STEM imaging modality is ptychography. Compared to iDPC, this method computationally reconstructs the image from a series of diffraction patterns. It is based on the measurement of a matrix of correlated diffraction images as a function of the position of the probe. It allows the reconstruction of both the phase and the amplitude of an object by reverse calculating it from the reciprocal space data [25]. For dose-resistant samples, ptychography has been able to achieve the highest possible resolution [26]. Ptychography requires scanning the sample while taking diffraction patterns at each scan spot [27]. A full micrograph requires a fast detector that is able to withstand the direct beam in diffractive mode, while still offering enough dynamic range per pixel. Furthermore, speed is an important practical limitation in cryo-electron ptychography. For a 410 × 410 nm field of view and a pixel size of 1 Å, the latest commercially available EMPAD detector (128 × 128 pixels) would allow for 0.98 ms dwell time

and a total exposure time of 394 s. For the same field of view and pixel size, a Timepix3 detector (256 × 256 pixels) would allow for 2.8 ms dwell time and a total exposure time of 287 s. For the Timepix4 detector (448 × 512 pixels) and the same field of view and pixel size, one would be able to obtain 0.1 ms dwell time and a total exposure time of 2.6 s. This improvement amounts to two orders of magnitude.

All in all, several cryo-STEM modalities are being investigated and offer a promising outlook to continue the cryo-EM Resolution Revolution for ever smaller and dynamic macromolecules. For now, for the above-mentioned modalities to become viable, additional work in detectors and processing algorithms is required.

7.2 Management of life sciences research data and the organisation thereof

During our process of characterising and integrating Timepix3, we faced several challenges with regard to the availability and openness of hardware, software, and data. As described in Chapter 3, we were hindered by the lack of open descriptions of how to integrate Timepix3 into the Thermo Fisher TEM hardware and software platform. We were also surprised by the lack of data and open software that are available along with published DQE and MTF characterisations of detectors. This dearth was true for both commercial and scientific sources of detector characterisations. We experienced that the results of these characterisations can vary tremendously depending on the exact data used and the exact numerical implementation of the algorithm. This experience exemplifies how effective research data management (RDM) is a key component of reproducible research and research integrity. How to organise effective RDM, on a technical and organisational level, is the second aspect I have examined in this thesis.

In Chapter 5 a case study was made of the DataHub Maastricht RDM support group and its lessons learnt after being in operation for five years. We looked at both technical solutions, in terms of flexible and powerful software to deal with the volume of data and metadata, and solutions on an organisational level. In Chapter 6 details and an implementation plan to improve the description and structure of metadata in the RDM platform iRODS were given. Our results of Chapter 5 showed that several technological and organisational lessons can be learnt to improve the support for RDM at the institution level. We identified that the availability and training of disciplinary data stewards can play a key role in being able to bridge the gap between the high-level policies and the day-to-day practises of individual researchers. Our observations support similar findings

by Teperek *et al.* [28] and corroborate the high granularity of data stewards Mons[29] stated as necessary.

A recurring theme in discussions around RDM is providing the right carrots to researchers to start and continue their RDM. Our results of Chapter 5 showed that reducing storage costs could be an effective way to motivate researchers. However, our limited success with other incentives may also lead to the conclusion that implementing mandatory requirements or policies may be the only way forward. It is well known that punitive measures are counterproductive in the long run. A bundle of case studies on RDM in several institutes also indicates the limited success in using policies [30]. Instead, it shows many examples of how positive one-to-one support of researchers, in the form of domain data stewards, data champions, or starting data communities, can have a positive effect on RDM uptake.

An obvious disadvantage of this highly granular support is that it will be hard and expensive to truly scale it to everyone in the research community realistically. A more scaleable approach which has shown promising signs to be effective in promoting RDM is the rise of more public digital repositories and especially domain-specific repositories [31]. The availability of public digital research repositories such as FigShare, Zenodo, and BioStudies among others, most of which are at least partly free, have made it increasingly feasible to publish data with at least some form of metadata. Next to these generic repositories, domain-specific digital repositories allow for specific and detailed metadata, and enforce those standards on their users. As noted in a survey, the data repository staff reported that data quality is an important criterion before allowing data to be published to their repository [32]. This data quality can make domain-specific repositories also more applicable for uptake in their respective research field. Currently, re3data.org, a searchable registry for digital research repositories, lists 1900 digital research repositories. A number that nearly doubled from 1000 listed repositories in 2014 [33]. However, in a recent survey, only 30% of the researchers reported that they are depositing their data in a digital repository of some sort [34].

Where does this discrepancy in apparent availability of repositories and their usage come from? It has been stated that much of the success of a domain-specific repository depends on its longevity, exact implementation of its metadata schema, and uptake within a research community [31]. An important driving force can, for example, be a research community *de facto* requiring the availability of data in a specific repository that covers their research domain.

The myriad of factors that contribute to the success of a domain-specific repository is illustrated by the history of the Protein Data Bank (PDB), one of the oldest still active digital repositories [35]. Founded in 1971, and merged with several other repositories, the World Wide PDB (wwPDB) currently hosts more than

200,000 structures of biological macromolecules and its identifiers were referenced by nearly 600,000 papers in 2021 just alone [36]. These are staggering numbers, but it took a long time to reach this level of acceptance and use [37]. Mandatory PDB deposition of structure data as a condition of publication was advocated by key opinion leaders, and recommended guidelines were published in 1989. The PDB metadata schema was developed over a series of official workshops in the 1990s, leading to the mmCIF format, which is currently still in use [38]. The PDB also continued to evolve, starting initially with X-ray diffraction but later integrated modalities such as nuclear magnetic resonance (NMR) spectroscopy, widening its impact and remaining relevant. In essence, the historical account of the PDB underscores the importance of longevity and widely accepted metadata schemes for digital repositories. Nevertheless, deriving a universal blueprint for all digital repositories from this particular case is challenging. Conducting a systematic assessment of the effectiveness of domain-specific repositories and evaluating various implementation approaches would be an avenue for future research.

The advance of centralised RDM services, the use of data management plans, the publication and adoption of the FAIR principles, and the rise of more digital repositories have brought a huge and well-deserved spotlight on the practise of good RDM and Open Data practises. However, a part that has been somewhat neglected is the sharing of the accompanying software and pipelines together with the data. The data of a research paper are often just the outcome or input of the research. The calculations and data pipelines to get to the resulting figures are nearly as important for the ability to accurately reproduce and reuse research. Often these details cannot be captured with enough detail in the written text of the method section of a paper; instead, they require the publication of the research software itself.

However, several unsolved challenges surround the publication of research software. It is certainly possible to share and publish the source code of scripts and software written by the researchers themselves. For example, platforms like GitHub allow the sharing of code, and the digital repository Zenodo has made it possible to make a permanent record from a code base stored on Github. But most software made today depends on a large number of dependencies and runtimes to be able to operate. It takes considerable effort and knowledge to properly document which software dependencies and runtime environments have been used. This interrelationship may mean that it can become increasingly difficult to run a certain piece of software even just a few years after publication. Other challenges surrounding research software are the use of proprietary software, for which the source code is not available or difficult to obtain due to costs or availability. These days, analyses can also be performed in increasingly complex pipelines of multiple pieces of software and configuration parameters. Together,

these challenges make it very difficult to capture exactly how the data have been processed from input to output across all steps, in order to reproduce the result. Generic tools such as software containerisation using Docker, and domain-specific tools such as Scipion [39] for cryo-EM are providing (part of the) answers to these challenges, but are also leaving many unanswered. I would urge policy makers and the RDM community at large to pay more attention, in the coming years, to the open publication of research software, both from a technological and a policy standpoint.

The data and software produced as part of this thesis have been published in a variety of different digital repositories (see Published Work). I would like to highlight the paper by Lamers *et al.*[40], to which I contributed the RDM, analyses, and publication of the EM data. The paper, published in Science in May 2020, shows how the SARS-CoV-2 virus can productively infect human enterocytes, a gut tissue. Due to the COVID-19 pandemic that raged at the time, all data analyses were performed remotely from home. Through the use of OMERO, an RDM platform for microscopy data that I had installed and configured at Maastricht University, we were able to do collaborative annotation of the images remotely from home and run the analyses remotely within OMERO. After composing the manuscript, we worked together with the curators of the Image Data Repository (IDR), a domain-specific repository for microscopy data, to get the data published. To facilitate reuse, the data had to be reformatted from an homemade BigTiff format to the IDR recommended community standard OME.TIFF, for which I spent a couple of days rewriting software code. The data have been published in the IDR under the identifier IDR0083 [41], marking the first time detailed *in situ* electron microscope images of human tissue infected with SARS-CoV-2 were publicly shared. Since then, the data have been reused at least a couple of times [42, 43].

This particular case of sharing of research data can be considered quite successful, but also show that a considerable amount of, mostly invisible, work often needs to be done before one can publish the data or software. From personal experience, it can be a real chore to perform proper RDM. This can range from re-formatting data, to collecting and writing down metadata, or writing data documentation. And whilst this (meta)data has been called a love note to the future, many feel the efforts invested in it are often not receiving the recognition it deserves. From a recent systematic review conducted by Devriendt *et al.*[44], they concluded that improving altmetrics, such as a h-index for data sharing, can be a powerful incentive. However, they comment that such an index can only work when accompanied by improvements to the practise of data sharing via persistent identifiers (such as DOIs). More importantly, they also require changes to policy and practises at institutes, funding agencies, and publishers. Combined, they would allow the sharing of data by researchers to be recognised and have a

meaningful impact on, for example, their career trajectories.

In conclusion, we need both highly granular RDM support, and ways to scale up this support to a larger audience; we need to improve the methods to capture the research process workflows from sample to data; and we need to recognise researchers' efforts in publishing their data. However, for the foreseeable future, we will produce more research data than we can realistically manage completely FAIR and open. Thus, as an RDM community, let us focus our efforts pragmatically on those data.

7.3 Concluding remarks

Research on life processes is complex. When considering the cryo-EM structural biology research field presented in this thesis, for example, no single person can have combined detailed understanding of how an electron microscope works, how a detector works, how proteins are purified and prepared, how the software and network infrastructure works, and how the algorithms work that process microscope data into 3D structures. However, I can state without reservations that our biology, life itself, is far more complex than anything humans have ever designed or built. For example, we do not fully understand how plants convert light into chemical energy, let alone how the human brain functions.

During the writing of this thesis, two remarkable advances in artificial intelligence (AI) were presented. The first was AlphaFold, a deep neural network that predicts the structures of proteins based only on their sequence. Released in 2017, AlphaFold emerged as the winner of the Critical Assessment of Structure Predictions (CASP) competition in the same year. The second was ChatGPT, a generative predictive neural network language model, released by the company OpenAI in the autumn of 2022. Its results took the world by storm, of which the outcome at the time of writing is not yet clear.

Both advances can be considered disruptive innovations. AlphaFold has opened a new era to structural biology. With more than 200 million predicted proteins, AlphaFold opens up a world of new possibilities in research, such as *in silico* site-directed mutagenesis experiments. ChatGPT and other language models have the potential to revolutionise the use of general search engines (such as Bing or Google) and in how they present, summarise, and adapt their results to us. Similarly, such a language model may also be able to find, retrieve, and summarise the most relevant research data for us. This ability would overcome many of the challenges described in this thesis, for example, with regard to the use of semantics in the description of metadata.

However, these innovations also come with a caveat. ChatGPT researchers and other AI researchers have admitted to not fully understand how the models

come to their results. The models are more complex than we can understand at this time. Thus, rightly so, there is a wide call for using caution in applying such technology too broadly, despite all the benefits they may bring us. They can also not overcome the fact that for such models garbage-in is garbage-out; i.e. AlphaFold heavily relies on highly curated PDB models to feed its neural network with accurate training data [37]. This means that while both AlphaFold and ChatGPT can be used with great effect to summarise or extrapolate our findings, they will need to be supported and fed, for the foreseeable future, with structured and curated input. Something to which I hope this thesis, even if it is just a microscopic bit, will have contributed.

7.4 References

1. Nakane, T. *et al.* Single-particle cryo-EM at atomic resolution. *Nature* **587,** 152–156. doi:10.1038/s41586-020-2829-0 (2020).
2. Fukuda, Y., Stapleton, K. & Kato, T. Progress in spatial resolution of structural analysis by cryo-EM. *Microscopy* **72,** 135–143. doi:10.1093/jmicro/dfac053 (2022).
3. Glaeser, R. M. Limitations to significant information in biological electron microscopy as a result of radiation damage. *J Ultrastruct Res* **36,** 466–82. doi:10.1016/s0022-5320(71)80118-1 (1971).
4. Langmore, J. P. & Smith, M. F. Quantitative energy-filtered electron microscopy of biological molecules in ice. *Ultramicroscopy* **46,** 349–373. doi:10.1016/0304-3991(92)90024-e (1992).
5. Henderson, R. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q Rev Biophys* **28,** 171–93. doi:10.1017/s003358350000305x (1995).
6. Majorovits, E., Angert, I., Kaiser, U. & Schröder, R. R. Benefits and Limitations of Low-kV Macromolecular Imaging of Frozen-Hydrated Biological Samples. **110.** doi:10.1016/j.bpj.2016.01.023 (2016).
7. Peet, M. J., Henderson, R. & Russo, C. J. The energy dependence of contrast and damage in electron cryomicroscopy of biological molecules. *Ultramicroscopy* **203,** 125–131. doi:10.1016/j.ultramic.2019.02.007 (2019).
8. Naydenova, K. *et al.* CryoEM at 100 keV: a demonstration and prospects. *IUCrJ* **6,** 1086–1098. doi:10.1107/s2052252519012612 (2019).
9. Faruqi, A. R., Cattermole, D. M., Henderson, R., Mikulec, B. & Raeburn, C. Evaluation of a hybrid pixel detector for electron microscopy. *Ultramicroscopy* **94,** 263–276. doi:10.1016/s0304-3991(02)00336-4 (2003).

10. McMullan, G. *et al.* Electron imaging with Medipix2 hybrid pixel detector. *Ultramicroscopy* **107,** 401–413. doi:10.1016/j.ultramic.2006.10.005 (2007).

11. Nederlof, I., van Genderen, E., Li, Y.-W. & Abrahams, J. A Medipix quantum area detector allows rotation electron diffraction data collection from submicrometre three-dimensional protein crystals. *Acta Crystallographica Section D: Biological Crystallography* **69,** 1223–1230. doi:10.1107/s0907444913009700 (2013).

12. Krajnak, M., McGrouther, D., Maneuski, D., Shea, V. O. & McVitie, S. Pixelated detectors and improved efficiency for magnetic imaging in STEM differential phase contrast. *Ultramicroscopy* **165,** 42–50. doi:10.1016/j.ultramic.2016.03.006 (2016).

13. Van Genderen, E. *et al.* Ab initio structure determination of nanocrystals of organic pharmaceutical compounds by electron diffraction at room temperature using a Timepix quantum area direct electron detector. *Acta Crystallographica Section A: Foundations and Advances* **72,** 236–242. doi:10.1107/s2053273315022500 (2016).

14. Mir, J. A. *et al.* Characterisation of the Medipix3 detector for 60 and 80keV electrons. *Ultramicroscopy* **182,** 44–53. doi:10.1016/j.ultramic.2017.06.010 (2017).

15. Paton, K. A. *et al.* Quantifying the performance of a hybrid pixel detector with GaAs:Cr sensor for transmission electron microscopy. *Ultramicroscopy* **227,** 113298. doi:10.1016/j.ultramic.2021.113298 (2021).

16. Jannis, D. *et al.* Event driven 4D STEM acquisition with a Timepix3 detector: Microsecond dwell time and faster scans for high precision and low dose applications. *Ultramicroscopy* **233,** 113423. doi:10.1016/j.ultramic.2021.113423 (2022).

17. Llopart, X. *et al.* Timepix4, a large area pixel detector readout chip which can be tiled on 4 sides providing sub-200 ps timestamp binning. *Journal of Instrumentation* **17,** C01044. doi:10.1088/1748-0221/17/01/c01044 (2022).

18. Glaeser, R. M. Chapter Two Specimen Behavior in the Electron Beam. *Methods in Enzymology* **579,** 19–50. doi:10.1016/bs.mie.2016.04.010 (2016).

19. Wu, C., Shi, H., Zhu, D., Fan, K. & Zhang, X. Low-cooling-rate freezing in biomolecular cryo-electron microscopy for recovery of initial frames. *QRB Discovery* **2,** 213–221. doi:10.1017/qrd.2021.8 (2021).

20. Wieferig, J.-P., Mills, D. J. & Kühlbrandt, W. Devitrification reduces beam-induced movement in cryo-EM. *IUCrJ* **8,** 186–194. doi:10.1107/s2052252520016243 (2021).

21. Lazić, I. & Bosch, E. G. T. Chapter Three Analytical Review of Direct Stem Imaging Techniques for Thin Samples. *Advances in Imaging and Electron Physics* **199,** 75–184. doi:10.1016/bs.aiep.2017.01.006 (2017).

22. Zhang, Y. *et al.* Single-particle cryo-EM: alternative schemes to improve dose efficiency. *Journal of Synchrotron Radiation* **28,** 1343–1356. doi:10.1107/s1600577521007931 (2021).

23. Lazić, I., Bosch, E. G. T. & Lazar, S. Phase contrast STEM for thin samples: Integrated differential phase contrast. *Ultramicroscopy* **160,** 265–280. doi:10.1016/j.ultramic.2015.10.011 (2016).

24. Lazić, I. *et al.* Single-particle cryo-EM structures from iDPC–STEM at near-atomic resolution. *Nature Methods* **19,** 1126–1136. doi:10.1038/s41592-022-01586-0 (2022).

25. Hoppe, W. Beugung im inhomogenen Primärstrahlwellenfeld. I. Prinzip einer Phasenmessung von Elektronenbeungungsinterferenzen. *Acta Crystallographica Section A* **25,** 495–501. doi:10.1107/s0567739469001045 (1969).

26. Chen, Z. *et al.* Electron ptychography achieves atomic-resolution limits set by lattice vibrations. *Science* **372,** 826–831. doi:10.1126/science.abg2533 (2021).

27. Rodenburg, J. Ptychography and Related Diffractive Imaging Methods. *Advances in Imaging and Electron Physics* **150,** 87–184. doi:10.1016/s1076-5670(07)00003-1 (2008).

28. Teperek, M., Cruz, M. J., Verbakel, E., Bohmer, J. & Dunning, A. Data Stewardship addressing disciplinary data management needs. *International Journal of Digital Curation* **13,** 141–149. doi:10.2218/ijdc.v13i1.604 (2018).

29. Versweyveld, L. We need 500.000 respected data stewards to operate the European Open Science Cloud - News blog - e-Infrastructures Reflection Group. http://e-irg.eu/news-blog/-/blogs/we-need-500-000-respected-data-stewards-to-operate-the-european-open-science-cloud (2016).

30. Clare, C. *et al. Engaging researchers with data management: The cookbook* doi:10.11647/obp.0185 (Open Book Publishers, 2019).

31. Sansone, S.-A. *et al.* FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology* **37,** 358–367. doi:10.1038/s41587-019-0080-8 (2019).

32. Kindling, M. & Strecker, D. Data Quality Assurance at Research Data Repositories. *Data Science Journal* **21.** doi:10.5334/dsj-2022-018 (2022).

33. Weisweiler, N. L. & Strecker, D. *Celebrating 10 Years of re3data – The Registry of Research Data Repositories* in (Zenodo). https://doi.org/10.5281/zenodo.6697943.

34. Tenopir, C. *et al.* Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLoS ONE* **15,** e0229003. doi:10.1371/journal.pone.0229003 (2020).

35. Crystallography: Protein Data Bank. *Nature New Biology* **233,** 223–223. doi:10.1038/newbio233223b0 (1971).

36. Burley, S. K. *et al.* Protein Data Bank: A Comprehensive Review of 3D Structure Holdings and Worldwide Utilization by Researchers, Educators, and Students. *Biomolecules* **12,** 1425. doi:10.3390/biom12101425 (2022).

37. Burley, S. K. & Berman, H. M. Open-access data: A cornerstone for artificial intelligence approaches to protein structure prediction. *Structure* **29,** 515–520. doi:10.1016/j.str.2021.04.010 (2021).

38. Fitzgerald, P. M. D. *et al.* International Tables for Crystallography. *International Tables for Crystallography,* 295–443. doi:10.1107/97809553602060000745 (2013).

39. De la Rosa-Trevín, J. M. *et al.* Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. *Journal of Structural Biology* **195,** 93–99. doi:10.1016/j.jsb.2016.04.010 (2016).

40. Lamers, M. M. *et al.* SARS-CoV-2 productively infects human gut enterocytes. *Science* **369,** 50–54. doi:10.1126/science.abc1669 (2020).

41. Van Schayck, J. P. *SARS-CoV-2 productively Infects Human Gut Enterocytes* 2020. doi:10.17867/10000135.

42. Son, R. *et al.* Morphomics via Next-generation Electron Microscopy. *arXiv.* doi:10.48550/arxiv.2111.14373 (2021).

43. Moore, J. *et al.* OME-NGFF: a next-generation file format for expanding bioimaging data-access strategies. *Nature Methods* **18,** 1496–1498. doi:10.1038/s41592-021-01326-w (2021).

44. Devriendt, T., Shabani, M. & Borry, P. Data Sharing in Biomedical Sciences: A Systematic Review of Incentives. *Biopreservation and Biobanking* **19,** 219–227. doi:10.1089/bio.2020.0037 (2021).

Societal impact

Structural Biology has had a profound societal impact on medicine, particularly in the context of vaccine development for infectious diseases. The relatively recent Resolution Revolution in cryo-electron microscopy (cryo-EM) has enabled researchers to investigate the intricate 3D structures of biological molecules, including viruses and their constituent proteins.

With the outbreak of the COVID-19 pandemic caused by the SARS-CoV-2 virus, cryo-EM played a crucial role in understanding the structure of the virus, specifically the spike protein responsible for viral entry into host cells. High-resolution structures obtained through cryo-EM allowed researchers to identify key regions on the spike protein that are crucial for interaction with human receptors, which subsequently facilitated the design of vaccines targeting these specific regions [1]. In the past, determining the structure of biological molecules was a time-consuming and technically challenging task. However, cryo-EM has significantly reduced the time required to obtain high-quality structures, allowing researchers to quickly characterise new strains of the virus and adapt existing vaccines accordingly.

The work in this thesis will have contributed to the continuous improvement of cryo-EM as a technology. An illustrative example of this progress is the integration of graphene as a substrate material for sample carriers (Chapter 4). This finding may effectively address the challenges of electrostatic sample charging encountered by researchers, offering the potential to enhance the overall operational efficiency of cryo-EM. Furthermore, investigation of the use of the Timepix3 HPD chip (Chapter 2 and Chapter 3) holds the promise of democratising access to expensive cryo-EM techniques in the coming years. Microscopes designed for 100 kV are a lot more affordable, making it more accessible for smaller labs to obtain such a device. Furthermore, the use of 100 kV has shown great promise, but is still hindered by an effective detector for this energy range. A role that the Timepix3 or future HPD chips could fulfil (Chapter 7). This is further underlined by the recent market introduction of Thermo Fischer's less expensive 100 kV Tundra electron microscope.

In general, public funding of expensive lab techniques, such as electron microscopes, has been a topic of debate. For example, is it feasible for multiple laboratories in the Netherlands to operate a cryo-EM facility? As an alternative to public funding, researchers are encouraged or even mandated to engage in public-private partnerships. E.g. this thesis was (partially) funded by a grant supporting a public-private partnership, in this case, to develop and characterise the Timepix3 chip as detector for electron microscopy. This successful partnership

led to the publication of a patent that was licenced by the private partner [2]. However, it is imperative to acknowledge that the engagement of public-private partnerships to secure funding for these endeavours carries certain inherent complexities and potential risks. One prominent concern revolves around the potential influence that private entities may wield over the research trajectory, potentially skewing scientific pursuits toward commercial interests. The use of patents, while offering a route for private partners to commercialise scientific breakthroughs, further introduces a layer of potential conflict that necessitates vigilance.

The second aspect of this thesis is the description, storage, and management of life science research data and the organisation thereof to help make research data more findable, accessible, interoperable, and reusable (FAIR). The Open Science and FAIR movement have been instrumental in promoting data reuse and addressing the issue of irreproducibility. The reuse of research data has a profound societal impact that extends far beyond the academic community. When researchers share their data for others to reuse and validate, it fosters a culture of transparency, collaboration, and open dissemination of knowledge. Embracing FAIR principles ensures that research data are well organised, described with standardised metadata, and made available through easily accessible repositories. One of the most significant benefits of data reuse is the potential to build on existing research, saving time, resources, and reducing redundancy. This, in turn, can lead to accelerated scientific progress and the development of more robust and reliable conclusions.

On the flip side, the costs of irreproducible science can be substantial and detrimental to society. When research results cannot be independently verified due to a lack of access to the underlying data or flawed methodologies, it can lead to misleading or even harmful applications in real-world contexts. The consequences can range from wasted funding and resources to public trust erosion in scientific findings. Furthermore, in fields where research results impact public policy or medical decisions, irreproducibility can have severe consequences on human well-being.

However, we should be careful not to make the sharing of data a goal on its own. I am careful not to advocate for RDM and data sharing for their own sake. In all cases, the effective use to which research data can be put should be at the heart of our activities. In some cases, a subset of processed data can serve (nearly) the same purpose as the full raw data archive. For example, when the full raw data archive makes the data unwieldy to use. In other cases, this may mean that it is better to store a sample in a freezer than to store (all) data on a server. For example, reanalysis of the sample may be cheaper and may even be done with improved technology in the future. So, whilst (expensive) digital infrastructures and specialised personnel can be part of the complicated data

management practises I am encouraging, they should never replace a pragmatic focus on improving the efficiency and inclusivity of the research process and helping new research questions be answered.

References

1. Wigge, C., Stefanovic, A. & Radjainia, M. The rapidly evolving role of cryo-EM in drug design. *Drug Discovery Today: Technologies* **38,** 91–102. doi:10.1016/j.ddtec.2020.12.003 (2020).
2. **Van Schayck, J. P.** & Ravelli, R. B. G. https://data.epo.org/gpi/EP3525229A1(2018).

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.

Acknowledgements

So here it is, my acknowledgements to the many people who have helped me along the way. I am extremely grateful to have had the opportunity to meet and interact with so many wonderful people over this eight-year-long journey.

First of all, I would like to thank the **committee members** for dedicating their valuable time to the critical reading of this thesis.

Dear **Peter** and **Carmen**. Thank you for the opportunity you gave me and your guidance throughout the years. Peter, I appreciate your encouragements to keep my research as applicable as possible and relevant for a broad (scientific) audience. I have enjoyed that challenge (and hopefully fulfilled some of it). Carmen, thank you for reminding me of the biology behind the techniques we were pursuing. Thank you both for your continued enthusiasm and encouragements in difficult times and the trust you put in me to develop my scientific skills. I have learnt so much from both of you.

Dear **Yue**, I am not sure where I would have been without you. Your loyalty, friendship and expertise have been absolutely crucial in getting this thesis to the finishing line. Not to mention your Chinese snacks and food. I am honoured to have you as my paranymph. Love to Rong.

Beste **Maarten**, dank voor je vriendschap en de mooie jaren samen. Ik denk dat wij elkaar altijd erg goed aanvulden, en daardoor bereikten we samen meer dan ieder voor zich. We realiseerden ons dat, denk ik, al vanaf de eerste dag. Dank voor je perfectionisme en de structuur die je bij DataHub bracht. Misschien heb ik dat niet vaak genoeg tegen je gezegd, maar dat is onmisbaar gebleken. Ik ben heel erg trots en blij dat je mijn paranimf wilt zijn. Al het beste, ook voor Lian, Loek en Eef.

Dear **Casper** and **Abril**. Thank you for being here with me since the beginning. Describing those pioneering years in the lab to outsiders is always a challenge, but you were there and understand. Both of you hold a special place in my heart. Love to Nuria, Nynke and Sophie.

Beste **Pascal**. Dankjewel voor het schijnbaar oneindige vertrouwen dat je in mij had. Ik kon bij je terecht als ik ergens mee zat, en je ontzorgde me door problemen bij mij weg te houden. Je bent voor mij echt een voorbeeld van dienend leiderschap.

Beste **Eric** en **Erik**. MOL3DEM forever! Dankjewel dat jullie mijn collega's-op-afstand wilden zijn. Ik heb genoten van ons kleine appgroepje dat we vol spamden met een bijna constante stroom aan grafieken, afbeeldingen, ideeën en gedachten. Dank voor jullie enthousiasme om elk probleem dat we tegenkwamen met frisse moed op te lossen.

Dear **Hang**, thank you for your friendship, advice and professional expertise. Your ability (and speed!) to make sense of a text (and return it in red :P), never ceased to amaze me. Love to Mattas and Ofelia.

DataHub-collega's, ik wil jullie enorm bedanken voor jullie kameraadschap, en het mooie werk dat we samen hebben opgeleverd. Jullie waren mijn tweede werkfamilie, waar ik altijd naar toe kon gaan. Dankjewel **Daniel, Mirjam, Jonathan, Rogier, Tim, Laurent** en **Dean**. Ook dank aan iedereen van de grotere DataHub-familie, in het bijzonder **Dennie** en **Ralph**, voor het delen van inzichten en expertise in de wonderlijke wereld van onderzoeksdatabeheer.

I would like to thank all current and former Nanoscopy and Microscopy CORE lab members. Thank you for making me feel welcome, for sharing the positive moments, and for standing together during challenging times. Thank you **Frank** for your friendship and assistance, in and outside of work. I wish we could have done more projects together. We formed a good team. Thank you **Anita** for your endless practical support. Thank you **Navya** for sharing your amazing story, and success in finishing your thesis. Thank you **Willine, Audrey, Hans** and **Helma** for making the lab a great place to work. Thank you **Chris** for friendship, collaborations and coming up with the title for this thesis. Thank you **Kèvin** for sharing your expertise. Thank you **Giancarlo, Axel, Nino** and **Kristof** for your camaraderie. **Sneha**, good luck in finishing your thesis. Thank you **Jan-Erek, Ralph, Hirotoshi, Marie-Helene, Maria, Michelle, Dora** and **Delei** for our time together and your team spirit.

I would like to thank the people at PSI and C-CINA for making me feel welcome the couple of times I visited. Thank you **Jan-Pieter** and **Erik** for your help and advice in this project.

A very special thanks also to current and former members of the Amsterdam Scientific Instrument team. Your expertise, engineering and guidance made this project possible in the first place. You often went far beyond what was required of you to help me. Thank you **Dmitry, Bram, Igor, Erik, Jord, Mathijs** and many others.

I also would like to thank the IDEE team for their invaluable help in many areas of the mechanical and electrical engineering of this project. Special thanks to **Maurice** for operating a clean server room together.

I would like to thank members of our fellow M4i Mass Spec division. Thank you **Ron** for our collaborations in the Medipix projects, and the trust you put in me to represent M4i at Medipix meetings. Thank you **Benjamin** for collaborations in IT and data. Thank you **Gert, Ian** and **Anjusha** for exchanging your knowledge and experience in operating and using the Timepix.

Also a special thanks to the Medipix team at CERN. This project would not have existed without your work. I cannot cease being amazed at how such a small team can pull off such a humongous task. Thank you **Michael, Erik, Xavi** and all others. I also would like to thank **Martin** at Nikhef for being a guiding light in all Timepix3 curiosities.

I would like to thank the students I have had the pleasure to supervise. Thank you **Dirk, Hugo** and **Lucas** for your fresh insights, your hard working and knowledge you brought to the project. The outcome would have looked different without your input.

Thank you to the U2Connect team for their sharing of knowledge in everything iRODS and related. Especially **Ton** and **Lazlo**, with whom the brainstorming for and writing of Chapter 6 was an absolute pleasure.

Beste **Sef** en **Wesley**. Dank voor jullie expertise in een gebied waar ik totaal geen kaas van heb gegeten. Ik heb jullie financiële kennis (vaak extreem last-minute!) bij de subsidieaanvragen enorm gewaardeerd.

I have really enjoyed the opportunity and input I could give to the NLBioImaging and NEMI data management team. I feel a bit sad for leaving so much of the plans we had unfinished, but feel encouraged to see the work continue. Thank you **Ben, Rohola, Marc, Katy, Eric, Lennard, Morris** and the many, many others who were involved in these efforts.

Mijn dank is ook groot aan het FHML-ICT team. Dankjewel **Erik, Pieter, Pascal, Patrick** en **Bart** voor jullie ondersteuning, zowel aan Nanoscopy, MCL als aan DataHub. Ik heb jullie inzet enorm gewaardeerd.

Ik wil mijn huidige KNMI-collega's enorm bedanken voor hun warme welkom en om mij wegwijs te maken in de wonderlijke wereld van de meteorologie. Dank voor jullie geduld terwijl ik deze thesis heb afgemaakt. Dankjewel **Annegies, Jeffrey, Rosina, Lukas, Wim, Tim** en heel veel anderen!

I would also like to make a very special appreciation to all the Eri(c|k)s that helped me throughout these years. Thank you **Eric, Erik, Erik, Erik, Erik, Erik, Eric, Jan-Erek** and **Erik**. You know who you are. I always knew which Eri(c|k) I could count on, and you never let me down!

Beste **Marcel** en **Marcel**. Dank voor jullie vriendschap. Jullie waren er vanaf het begin wekelijks bij. Jullie en ons project WeDeclare vormden een welkome afleiding (soms te veel!). Dank voor jullie geduld met mij terwijl ik deze thesis afmaakte. Ik hoop dat we nog velen jaren aan onze hobby verder kunnen werken. Ik weet zeker dat we uiteindelijk aan de juiste Balie staan.

Lieve **Pieter**. Dank voor je lange vriendschap. Het was bijzonder om dat eerste jaar wekelijks bij jou te logeren. Voor ons alle twee een bijzondere tijd. Ik waardeer het enorm dat we alles met elkaar kunnen delen.

Dankjewel **Angelo, Roxanne, Loes, Maaike, Gerben, Marjolijn, Bart, Brenda, Boris, Karin, Jeroen, Leoni, Lars, Jouke, Mariëlle, Pieter, Nathalie, Wim Joost, Elvy, Anne, Chantal, Sanne, Leonoor, Marijke, Ron, Ilse** en andere vrienden voor de mooie zeiltochtjes, kampeerweekenden, feestjes en vooral jullie vriendschap.

Tessa, Daan, Jan, Dorien, Peter, Anne-Maaike, Jeroen, Gianna, Paul en **Ceciel**, ik wil jullie bedanken voor jullie liefde en steun. Dank voor jullie interesse in waar ik mee bezig ben geweest. Ik snap dat het nog steeds even onbegrijpelijk is wat ik heb gedaan. **Bart**, dankjewel voor het mede-schrijven van het gedicht aan het begin van deze thesis.

Als laatste wil ik een paar mensen bedanken die extra bijzonder voor mij zijn.

Lieve **Raimond**. Zonder jou was ik nooit aan deze thesis begonnen. Zonder jou had ik deze thesis nooit afgemaakt. Ik dank je voor je geduld, je positieve kijk, je enthousiasme, je collegialiteit en vriendschap. Naar mijn gevoel zijn we acht jaar geleden een gesprek begonnen, en is dat gesprek nooit opgehouden. In onze gesprekken heb je me geleerd kritisch maar optimistisch te denken en je hebt me geleerd het grotere plaatje nooit uit het oog te verliezen. Je was altijd bereid om tijd vrij te maken om mij verder te helpen. Al koos je daar wel soms de meest gekke tijdstippen voor (en een kop cappuccino met suiker). Je advies was opbouwend, en je ontnam me nooit mijn vrijheid. Je gaf me juist alle ruimte om eigen ideeën te ontwikkelen en paden af te lopen. Het is verschrikkelijk jammer dat we niet samen de voltooiing van deze thesis kunnen vieren. Weet dat ik het niet zonder je had gekund, en dat het een grote eer voor mij is geweest om het samen met jou te doen. Heel veel liefs aan jou, en aan Maaike, Seppe en Noé.

Lieve **papa** en **mama**. Ik heb zoveel aan jullie te danken. Jullie onvoorwaardelijke liefde en steun betekenen heel veel voor mij. Pap, dankjewel voor je interesse in waar ik mee bezig ben en je adviezen wanneer ik die nodig had. Mam, dankjewel voor je wijsheid en al je hulp met de jongens. Het was heel bijzonder om deze afgelopen zeven jaar weer opnieuw zo dicht bij elkaar te wonen, en zoveel met elkaar te delen.

Lieve **Nolan** en **Woud**. Hier is die dan eindelijk echt: het saaie boekje van papa. Ik hoop dat jullie dit ooit lezen, en dan vooral aan het leuke feestje van papa terugdenken. Het is geweldig om jullie groter te zien worden, en jullie je eigen avonturen te zien beleven. Nu dit boekje af is, beloof ik dat we er nog meer samen gaan beleven.

Tot slot wil ik mijn lieve **Pauline** bedanken: mijn steun en toeverlaat. Lieve Pauline, al jaren moet je het doen met mijn standaard excuus 'dat boekje' als ik je vroeg om je geduld, hulp of begrip, terwijl ik zelf vaak genoeg stresserig en ongeduldig ben geweest. Ik hoop dat je het me allemaal kan vergeven. Je (eigen)wijsheid, liefde en toewijding maken me (nog steeds) zielsgelukkig.

Paul van Schayck
Bennekom
1st of January 2024

Published work

Articles

1. Heshof, R., **Van Schayck, J. P**, Tamayo-Ramos, J. & de Graaff, L. H. Heterologous expression of Gaeumannomyces graminis lipoxygenase in Aspergillus nidulans. *AMB Express* **4.** doi:10.1186/s13568-014-0065-4 (2014).
2. Palovaara, J., Saiga, S., Wendrich, J. R., van 't Wout Hofland, N., **Van Schayck, J. P**, Hater, F., Mutte, S., Sjollema, J., Boekschoten, M., Hooiveld, G. J. & Weijers, D. Transcriptome dynamics revealed by a gene expression atlas of the early Arabidopsis embryo. *Nature Plants* **3,** 1. doi:10.1038/s41477-017-0035-3 (2017).
3. **Van Schayck, J. P**, Smeele, T., Theunissen, D. & Westerhof, L. Providing validated, templated and richer metadata using a bidirectional conversion between JSON and iRODS AVUs. *iRODS User Group Meeting 2019 Proceedings,* 9–17. https://irods.org/uploads/2019/irods_ugm2019_proceedings.pdf (2019).
4. Lamers, M. M., Beumer, J., van der Vaart, J., Knoops, K., Puschhof, J., Breugem, T. I., Ravelli, R. B. G., **Van Schayck, J. P**, Mykytyn, A. Z., Duimel, H. Q., van Donselaar, E., Riesebosch, S., Kuijpers, H. J. H., Schipper, D., van de Wetering, W. J., de Graaf, M., Koopmans, M., Cuppen, E., Peters, P. J., Haagmans, B. L. & Clevers, H. SARS-CoV-2 productively infects human gut enterocytes. *Science* **369,** 50–54. doi:10.1126/science.abc1669 (2020).
5. **Van Schayck, J. P**, van Genderen, E., Maddox, E., Roussel, L., Boulanger, H., Fröjdh, E., Abrahams, J.-P., Peters, P. J. & Ravelli, R. B. G. Sub-pixel electron detection using a convolutional neural network. *Ultramicroscopy* **218,** 113091. doi:10.1016/j.ultramic.2020.113091 (2020).
6. Zhang, Y., Lu, P.-H., Rotunno, E., Troiani, F., **Van Schayck, J. P**, Tavabi, A. H., Dunin-Borkowski, R. E., Grillo, V., Peters, P. J. & Ravelli, R. B. G. Single-particle cryo-EM: alternative schemes to improve dose efficiency. *Journal of Synchrotron Radiation* **28,** 1343–1356. doi:10.1107/s1600577521007931 (2021).
7. **Van Schayck, J. P** & Coonen, M. First Line Research Data Management for Life Sciences: a Case Study. *International Journal of Digital Curation* **16,** 13. doi:10.2218/ijdc.v16i1.761 (2022).

8. **Van Schayck*, J. P**, Zhang*, Y., Knoops, K., Peters, P. J. & Ravelli, R. B. G. Integration of an event-driven Timepix3 hybrid pixel detector into a cryo-EM workflow. *Microscopy and Microanalysis* **29,** 352–363. doi:10.1093/micmic/ozac009 (2022).
9. **Van Schayck*, J. P.**, Zhang*, Y., Pedrazo-Tardajos, A., Claes, N., Noteborn, W. E. M., Lu, P.-H., Duimel, H., Dunin-Borkowski, R. E., Bals, S., Peters, P. J. & Ravelli, R. B. G. Charging of Vitreous Samples in Cryogenic Electron Microscopy Mitigated by Graphene. *ACS Nano* **17,** 15836–15846. doi:10.1021/acsnano.3c03722 (2023).

* both authors contributed equally

Patents

1. **Van Schayck, J. P.** & Ravelli, R. B. G. https://data.epo.org/gpi/EP3525229A1(2018).

Software

1. Van Schayck, J. P., Roussel, L., Fröjdh, E., Schübel, A. & Kraphol, D. *M4I-nanoscopy/geant4medipix* version v1.0.1. Feb. 2020. doi:10.5281/zenodo.3667660.
2. Van Schayck, J. P. *M4I-nanoscopy/tpx3-event-localisation* version v1.1.0. Aug. 2020. doi:10.5281/zenodo.3980701.
3. Van Schayck, J. P. & van Genderen, E. *M4I-nanoscopy/tpx3-tot-correction* version 1.0.0. Feb. 2020. doi:10.5281/zenodo.3641268.
4. Van Schayck, J. P. *M4I-nanoscopy/tpx3EventViewer* version v1.0.3. Mar. 2020. doi:10.5281/zenodo.3693990.
5. Van Schayck, J. P. *M4I-nanoscopy/tpx3HitParser* version 1.0.1. Mar. 2020. doi:10.5281/zenodo.3693995.
6. Van Schayck, J. P. & Ravelli, R. B. G. *M4I-nanoscopy/tpx3HitParser* version v2.2.0. July 2022. doi:10.5281/zenodo.6874070.
7. Van Schayck, J. P., Zhang, Y. & Ravelli, R. B. G. *M4I-nanoscopy/mtf-nps-dqe* version v1.0.0. July 2022. doi:10.5281/zenodo.6867808.
8. Van Schayck, J. P. & Ravelli, R. B. G. *M4I-nanoscopy/tpx3EventViewer* version v2.0.0. July 2022. doi:10.5281/zenodo.6873946.

Data

1. Van Schayck, J. P., Zhang, Y. & Ravelli, R. B. G. *Integration of an event-driven Timepix3 hybrid pixel detector into a cryo-EM workflow* version v1.0.3 (Zenodo, Mar. 2022). doi:10.5281/zenodo.6851220.
2. Schayck, J. P. v. *Background interviews for "First line research data management for the life sciences: a case study"* (2021). https://hdl.handle.net/21.12109/P000000190C000000001.
3. Van Schayck, J. P., van Genderen, E., Maddox, E., Roussel, L., Boulanger, H., Fröjdh, E., Abrahams, J.-P., Peters, P. J. & Ravelli, R. B. G. *Sub-pixel accuracy in electron detection using a convolutional neural network* version v1.0.0 (Zenodo, Feb. 2020). doi:10.5281/zenodo.3635923.
4. Van Schayck, J. P., Zhang, Y., Pedrazo-Tardajos, A., Noteborn, W. E. M., Claes, N., Lu, P.-H., Duimel, H., Dunin-Borkowski, R., Bals, S., Peters, P. J. & Ravelli, R. B. G. *Charging of vitreous samples in cryo-EM mitigated by graphene - Supplementary Figures and Movies* (June 2023). doi:10.6084/m9.figshare.23244299.v1.
5. Van Schayck, J. P. *SARS-CoV-2 productively Infects Human Gut Enterocytes* 2020. doi:10.17867/10000135.
6. Van Schayck, J. P., Zhang, Y., Knoops, K., Peters, P. J. & Ravelli, R. B. G. *Integration of an event-driven Timepix3 hydrid pixel detector into a cryo-EM workflow* (EMBL-EBI, June 2022). doi:10.6019/empiar-11113.

About the author

Paul van Schayck (Nijmegen, 8 May 1987) completed his pre-university education (VWO) at Bonnefanten College Maastricht in 2005. After an initial pursuit of aerospace engineering at TU Delft, he found his passion in biotechnology and earned his BSc degree from Wageningen University in 2012. Enjoying his time in Wageningen, he continued to pursue a Master's degree in Cellular and Molecular Biotechnology.

Having learnt how to write code from a young age, Paul co-founded and ran the web development company Agar Hosting throughout his time in Wageningen as a student. During his internship at the plant breeding company Rijk Zwaan, he could combine his love for computers and biology by enabling Rijk Zwaan researchers to use OMERO as a data management tool for image analysis.

After obtaining his Master's degree in 2014, Paul continued to work for a short while at Rijk Zwaan and built the online transcriptomic browser AlbertoDB.org for Wageningen University. In 2015, he moved to Maastricht to work as a data engineer at the Nanoscopy division of the newly founded Maastricht MultiModal Molecular Imaging Institute (M4i) and in a joint position at ResearchIT, a research data management support group which would later become DataHub Maastricht. One year later, being fascinated by how electron microscopes work, he was given the opportunity to start his PhD trajectory.

During his time as a PhD student, Paul continued to work as software engineer at DataHub Maastricht, and as data steward at M4i, acting as a liaison between M4i and DataHub. In his role as data steward, he was heavily involved with the writing and implementation of the Large-scale Research Infrastructure grant applications NEMI and NLBioImaging, which seek to consolidate, among other things, the research data management of microscopy data across The Netherlands. He was also co-applicant on the successful grant application 4DEM to continue the detector work in Maastricht with the next-generation Timepix4.

Since 2022, Paul has been working as software engineer at the Dutch meteorological institute KNMI. He continues to employ his skills in computer programming and data management by working on the KNMI Data Platform.

Paul can be contacted at paul@vanschayck.nl.