UNIVERSITY OF LIVERPOOL

DEPARTMENT OF ENGINEERING

DEVELOPING COMPUTATIONAL IMAGING METHODS FOR QUANTITATIVE MULTI-DIMENSIONAL ELECTRON MICROSCOPY

ALEX WILLIAM ROBINSON

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

DECEMBER 4, 2023

Acknowledgements

This work would not have been possible without the support of my supervisors, Professor Nigel D. Browning and Professor B. Layla Mehdi. I'd like to thank Nigel immensely for the time he has spent guiding me through this process. His wisdom and knowledge speaks for itself, but his enthusiasm for the subject is remarkable. Thanks to Layla for her motivation, encouragement, knowledge, and ultimately for teaching me that I can always improve. This is something that I'll carry forward throughout my career.

I'd like to give an extended thank-you to my incredible colleagues. Firstly, to Dr. Daniel Nicholls for helping me to settle in and his endless supply of useless facts, and thanks to Jack Wells for answering my Teams messages in the early hours and for being a support throughout. Thanks to Dr. Amirafshar Moshtaghpour for being the finest academic role model, friend, and comedian, thanks to Ioannis Siachos for his friendship and support, thanks to Zoe Broad for her snack supply and wonderful sense of humour, thanks to Joseph John Burman for asking me questions that I later had to verify whether my answers were correct, thanks to Dr. Xiaodong Liu for bringing his FIB skills to our group, and thanks to Eneith Aguilar Ronquillo for putting up with my Spanglish which generally worsens at the pub. I'd especially like to thank Dr. Mounib Bahri who has taught me so much about how to operate a microscope, as well as for staying with me until the late hours of the night trying to do experiments. Thanks to Dr. Yoshie Murooka for his incredible range of knowledge, but also for bringing positivity and laughter to the office. A thanks also goes to Dan Somers for his support in making some of this research a potential tool for others.

Thanks to the people I've met at conferences and through collaboration. Special thanks go to Professor Angus I. Kirkland for giving me his time to help with papers and use of the RFI facilities, Professor Miaofang Chi for providing data and support, Dr. Ian MacLaren for letting

me visit his facility, and Dr. Giuseppe Nicotra for not only allowing me to use his wonderful microscope, but also his friendship and hospitality. Thanks to other academics, such as Dr. Chen Huang and Dr. Abner Velazco-Torrejon both at RFI, Dr. Gianfranco Sfuncia at CNR-IMM, and also a thanks to Dr. Jordan Hatchel at ORNL. Thanks to all the people that took their time to talk to me at conferences with special thanks to Professor Robert Klie, Dr. Ryo Ishikawa, Dr. Andrew Lupini, Professor Yimei Zhu, Professor Naoya Shibata, Dr. Lewys Jones, Professor Stig Helveg, and Dr. Colin Ophus.

I'd like to thank all my friends that have helped me along the way. Thanks to my oldest friend Josh and twin brother Luke who have been there for me through it all, often keeping me motivated when it got tough. Thanks to my good friends Lee and Chris who have taught me a lot about life and that hard work pays off. Thanks to my friends from golf, Kieran, Dan, Sean, Mark, Louis, Ross, Richie, Paul, Chris and Will.

I wouldn't have managed this without my remarkable family. Cleo and James, you have motivated me more than you can imagine, especially by bringing my wonderful niece, Amber, into the world. To Luke and Darcie, you have been support in hard times and I can't thank you both enough. To my supportive and loving grandparents, Nancy, Jean, Owen, and my late Grandad Alan, thank you for encouragement and always asking me about what I do. My grandad Alan always told me that if I did what I enjoyed then I'd never work a day in my life, and so far he has been right.

Last and not least, my parents, Neil and Joanne. Since the day I could walk (or the day I could talk), you have invested in my education. The sacrifices that you both made to help me get to this point can never be thanked enough. You've taught me to be strong in the face of adversity, and you've always believed in me when I didn't. I owe all of this to your love and support.

In memory of Alan William Robinson

Abstract

Scanning transmission electron microscopy (STEM) is a well established tool for identifying the complex structures of materials with atomic resolution, as well as the structure of biological specimen. The STEM is a multi-signal acquisition tool, capable of characterisation through a variety of techniques and strategies. Despite these benefits, STEM is limited to a subset of materials which are resistant to the electron beam itself, and there exists a much larger set of materials which change their underlying structure due to electron-specimen interactions. Typically, the beam current is reduced to overcome these limits. This reduces the signal-to-noise of the acquisition, and ultimately makes interpretation difficult.

One solution to this problem is to consider the use of compressive sensing to reduce the beam exposure for sensitive materials. By reducing the number of acquired probe locations, the data can also be acquired much faster, leading to more efficient characterisation of materials.

One mode which is of particular interest is 4-dimensional STEM (4-D STEM), where a wide range of images can be constructed from one dataset. This method, however, is limited by the readout speed of detectors and the volume of data which is acquired to gather results.

It is demonstrated in this thesis that multi-dimensional STEM acquisition, such as 4-D STEM, can be improved through compressive sensing and computational imaging approaches. The methods are also applied to STEM simulation, as well as standard 2-D STEM to improve image quality. The thesis also demonstrates the first acquisition of sub-sampled 4-D STEM data, showing increased acquisition speeds for frame rate limited detectors.

Contents

List of Figures

List of Tables

List of Notations and Definitions

Unless otherwise stated, the following list indicates notations and definitions used throughout this work.

j: the imaginary unit

$\boldsymbol{r}$: a vector

$\boldsymbol{A}$: a vector field

A: a scalar field

v_i: italic index indicates a variable *i.e.*, $v_1, v_2, \cdots$

v_{i}: non-italic index indicates a naming convention *i.e.*, v_{p} is a probe location

$\mathcal{X}$: an array

$\mathbb{N}_0$: natural numbers including zero

$\mathcal{F}$: Fourier transform

1 | Introduction

(Scanning) transmission electron microscopy (S/TEM) is a powerful tool for analysis of complex materials on the nano-scale and below. This is fundamentally down to brighter, smaller, and more coherent electron probes thanks to the development of aberration correctors [6]. This has allowed for higher resolution in typical imaging regimes, as well as the development of atomically resolved analytical methods such as electron energy loss spectroscopy (EELS) and energy dispersive x-ray spectroscopy (EDS). All of this is excellent when the samples considered are beam stable, yet there are a host of samples which cannot remain in their intended state under these intense probes. These *beam-sensitive* samples are therefore limited in signal-to-noise ratio (SNR), resolution, and analysis possibilities– solutions to this are therefore high priority in the development of STEM methods.

Ultimately, the problem comes down to electrons. There is a tight-rope which must be walked to acquire (i) high enough signal for accurate analysis and (ii) low enough signal that the sample remains undamaged. The most common solutions to this problem are to either reduce the probe current sufficiently (*i.e.*, minimise the number of electrons incident on the sample per second) or lower the dwell time (*i.e.*, the amount of time the probe is stationary at a location) such that (ii) is satisfied, with (i) being compromised significantly. This reduction in SNR (especially if too low) will lead to loss of resolution in analysis, which could lead to unreliable characterisation or lack thereof.

There are other more exotic solutions which can be employed when analysing beam-sensitive samples. For example, cryogenic electron microscopy (cryo-EM) began development over 40 years ago [7], and aims to reduce the negative influence of the beam by suspending the sample in a cryogenic state. At lower temperatures, the reaction rate is slower for the conversion of the sample to beam induced radicals. This means that exposure can be prolonged beyond that

which would be observed at room temperatures. However, performing cryo-EM experiments adds a layer of complexity and cost which makes some existing STEMs unsuitable.

Another potential solution is the application of computational or signal processing techniques. These methods aim to utilise state-of-the-art developments in signal processing to improve imaging at low dose. One class, known as image inpainting, aims to use the theory of compressed sensing (CS) to significantly reduce the amount of acquired data by *filling-in* the missing information from the sub-sampled data. This can now be considered as two problems– firstly how to acquire sub-sampled data, and secondly how to reconstruct the sub-sampled data.

CS-STEM is a maturing technique within the field. Its origins lie in the necessity to balance (i) and (ii) above such that beam-sensitive samples can be imaged with higher resolution than current methods permit. By only positioning the electron probe at a sub-set of the intended locations, the total electron dose can be significantly reduced. Furthermore, the time-to-acquire is reduced by the same factor which means that stage drift and other instabilities have less effect upon the final image. However, STEM was not designed with these methods in mind, and as such the ability to acquire a sub-sampled image requires an external scan generator. The job of the scan generator is to modify the scan coil voltages appropriately so that the probe can be positioned wherever the user intends. The signal(s) is then acquired, reshaped, and then sent through to an inpainting algorithm.

There are several ways to inpaint missing data such as interpolation based methods, dictionary learning with sparse coding methods, or deep learning methods. In STEM, often there is no one way of analysis which suits all cases, and the user must decide which method will yield the best analysis of their data. The same ideas should be employed when inpainting, as the results of different methods can vary on a case-by-case basis. In this work, two inpainting techniques are considered– a dictionary learning and sparse coding algorithm known as the beta process factor analysis (BPFA), and a kernel based interpolation technique with sparse regularisation called regularised local means inpainting (R-LMI). The later was developed as part of this work as an immediate solution to the time-restrictions of the BPFA method, however this issue has been solved by Jack Wells, as part of their research [8].

Although CS-STEM for standard imaging such as high-angle annular dark field (HAADF) has been shown to work well, the development for multi-dimensional STEM techniques re-

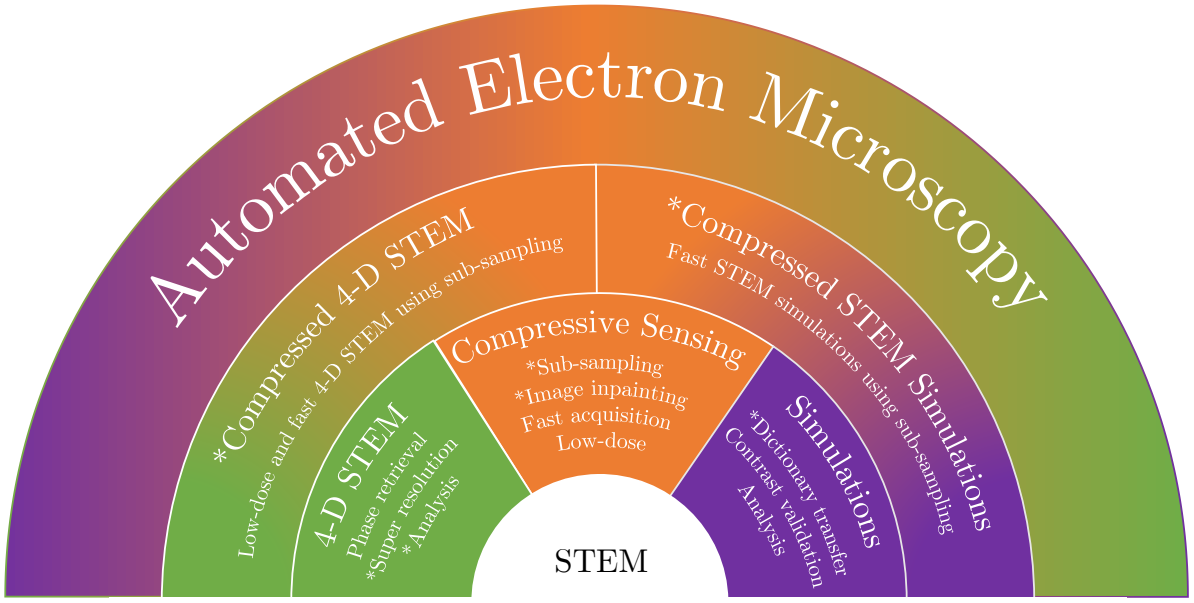

Figure 1.1: **Visualisation for the summary of this work.** This work aims to cover three main topics, (i) application of compressive sensing methods to STEM, (ii) application of compressive sensing methods to STEM simulation, and (iii) application of compressive sensing methods to 4-D STEM. Asterisks indicate novelty in this work.

mains lacking. Four-dimensional STEM (4-D STEM) is a popular, yet demanding technique within the field. In this method, a 2-D convergent beam electron diffraction (CBED) pattern is acquired at each location on a 2-D scanning grid, hence 4-D STEM. This means that data acquisition can quickly become a challenge to acquire, store, and process with only a small number of electron probes. This often restricts 4-D STEM to smaller fields of view (FOV), or larger scan step/pitch. It is not uncommon to find 4-D STEM datasets which exceed 100 gigabytes (Gb) in size, nor is it uncommon for acquisition times to far exceed several minutes. The later is down to the read-out speed of the camera used to image each CBED, with most typical cameras having frame rates on the order of 2000Hz, with the possibility of reaching $> 10,000$Hz with binning/windowing. For context, acquiring HAADF STEM images can be done with equivalent frames rates of $100,000$Hz (10μs dwell time), which is at least an order of magnitude faster. Not only that, the frame rate affects the lowest possible dwell time, which can be effectively 100μs up to $> 1m$s. This is going to lead to overexposure and potentially damage any sample under the illumination.

By extension, STEM simulation suffers similar issues when calculating the scattering of the probe. STEM simulation is often used as a verification method for observed contrast, and to validate the presence of defects, vacancies and interstitials. The multislice approximation is

a common method for calculating the exit wave-function of the probe after it has been projected through the theoretical sample potential, and similar to 4-D STEM, this requires a 2-D reciprocal space calculation at each 2-D scanning grid position. Typical multislice calculations can take on the order of minutes, potentially hours depending on parameter settings such as thermal diffuse scattering approximation, depth resolution, and reciprocal space resolution. Therefore, for STEM simulations to reach computational speeds where they could become useful *during* experimental acquisition, a new method must be developed that attempts to eliminate redundancy.

1.1 Chapter summaries

This thesis presents novel strategies for acquiring, inpainting, and analysing multi-dimensional electron microscopy data, and Fig. 1.1 gives a summary of the work contained within. The three main topics are conventional STEM, STEM simulations, and 4-D STEM. In all cases, a comprehensive background is given, as well as proposed methodology and results. The main contributions are outlined below.

In Chapter 2, a comprehensive overview of S/TEM is given, as well as a background containing the theory of electrons and scattering. In the context of this work, the aim is to provide a justification for why (i) the electron is a suitable mechanism for probing sub-atomic scales and (ii) how the underlying physics is then combined with engineering to develop a machine capable of doing so. It also considers the development of TEM, and how STEM and TEM are related through the principle of reciprocity. The different contrast mechanisms are discussed, as well as the physics which underpins the contrast transfer functions.

This chapter then leads into the drawbacks of S/TEM, mainly focussing on so-called *beam damage*. Here, common beam damage mechanisms are discussed such as knock-on damage and radiolysis which provide a motivation for developing methods which can mitigate.

Chapter 3 focusses on the theory of compressive sensing, more specifically image inpainting, in a general case. Here, the development of compressive sensing techniques within other fields provide analogies for application to S/TEM. This chapter is intended to be a point of reference for the reader to understand the underlying algorithms which are referenced throughout. The two main algorithms are (i) the BPFA and (ii) R-LMI techniques. Details on how these

algorithms can be optimised for STEM data are given, such as by incorporating sampling rate and image properties (such as the size of features in the image) into the models.

Chapter 4 is an overview of state-of-the-art CS-STEM application based on new methods developed up to this time of writing. This chapter covers how sub-sampled STEM data is acquired, processed, and analysed in experiment, as well as presenting work which aims to show that sub-sampling can out perform other low-dose techniques. This chapter also includes results of applying simulated to drive the recovery of experimental data through a technique known as simulated dictionary transfer. Here, by using prior knowledge, experimental data can achieve improved resolution by incorporating theory into experimental data.

Chapter 5 focuses on the application of CS to STEM simulation, and how certain redundancies can be eliminated through efficient calculation. Three main aspects of STEM simulation are explored, (i) the redundancy in real-space acquisition, (ii) the redundancy in reciprocal-space calculation, and (iii) how the frozen phonon model can be optimised by novel sampling strategies.

By incorporating CS with STEM simulations, the computation time can be significantly reduced without significant loss of information. This leads to the potential for real-time simulations to be performed alongside experimental acquisition, driving the recovery of experimental data in sync with analysis as discussed in Chapter 4.

Chapter 6 contains the second major project within this thesis, applying CS to 4-D STEM. This chapter outlines a detailed theoretical and experimental model for acquiring and inpainting sub-sampled multi-dimensional STEM data. The goal is to show that by using probe sub-sampling and detector down-sampling (*i.e.*, optimising sampling on the camera), the acquisition of 4-D STEM data can approach that of typical 2-D STEM acquisition. Results applied to simulated CS experimental data are given, as well as results when sub-sampling is used in practical acquisition of 4-D STEM.

Chapter 7 is an overview of other research undertaken as part of this thesis. This includes collaboration work, such as developing a robust variation of the ePIE algorithm to noise and sub-sampling (with Rosalind Franklin Institute), and the characterisation of a cadmium telluride-silicon interface using 4-D STEM (with CNR-IMM, Catania). The final section is an overview of the MAT 4-D STEM library/application which I developed as a user friendly

analysis tool for 4-D STEM data.

Chapter 8 presents conclusions which summarise each chapter above. This chapter also presents lines of research for future study, such as incorporating more signals into the acquisition process. This thesis is not intended to cover all aspects of electron microscopy, however serves as a basis for researchers wishing to understand the motivations and applications of CS-STEM in a multi-dimensional acquisition.

1.2 Journal publications

- **Robinson, A. W.**, Wells, J., Nicholls, D., Moshtaghpour, A., Chi, M., MacLaren, I., Kirkland, A.I. and Browning, N.D., 2023. Simultaneous High-Speed and Low-Dose 4-D STEM Using Compressive Sensing Techniques. Physical Review Letters, *with editors*

- **Robinson, A.W.**, Moshtaghpour, A., Wells, J., Nicholls, D., Broad, Z., Kirkland, A.I., Mehdi, B.L. and Browning, N.D., 2023. In silico Ptychography of Lithium-ion Cathode Materials from Subsampled 4-D STEM Data. arXiv preprint arXiv:2307.06138.

- Nicholls, D., Wells, J., **Robinson, A.W.**, Moshtaghpour, A., Kirkland, A.I. and Browning, N.D., 2023. Scan Coil Dynamics Simulation for Subsampled Scanning Transmission Electron Microscopy. arXiv preprint arXiv:2307.08441.

- **Robinson, A. W.**, Wells, J., Nicholls, D., Moshtaghpour, A., Chi, M., Kirkland, A.I. and Browning, N.D., 2023. Towards real-time STEM simulations through targeted subsampling strategies. Journal of microscopy, 290(1), pp.53-66.

- Browning, N.D., Castagna, J., Kirkland, A.I., Moshtaghpour, A., Nicholls, D., **Robinson, A. W.**, Wells, J. and Zheng, Y., 2023. The advantages of sub-sampling and Inpainting for scanning transmission electron microscopy. Applied Physics Letters, 122(5).

- Nicholls, D., Wells, J., **Robinson, A.W.**, Moshtaghpour, A., Kobylynska, M., Fleck, R.A., Kirkland, A.I. and Browning, N.D., 2023, June. A targeted sampling strategy for compressive cryo focused ion beam scanning electron microscopy. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

- **Robinson, A. W.**, Nicholls, D., Wells, J., Moshtaghpour, A., Kirkland, A. and Browning,

N.D., 2022. SIM-STEM Lab: Incorporating compressed sensing theory for fast STEM simulation. Ultramicroscopy, 242, p.113625.

- **Robinson, A. W.**, Nicholls, D., Wells, J., Moshtaghpour, A., Bahri, M., Kirkland, A. and Browning, N., 2022, May. Compressive scanning transmission electron microscopy. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1586-1590). IEEE. (Joint first authorship with Daniel Nicholls).
- Browning, N., Nicholls, D., Wells, J., and **Robinson, A. W.**, 2022. OPTIMAL SAMPLING AND RECONSTRUCTION STRATEGIES FOR SCANNING MICROSCOPES. Electronic Device Failure Analysis, 24(1), p.11–16.

1.3 Conferences

- Poster presentation: "Subsampling Methods for Fast 4-D STEM Acquisition", IMC20, Busan, S. Korea, 2023
- Platform presentation: "Sub-Sampled S(T)EM: Improving Real-Time Reconstruction Quality using Dictionary Transfer", IMC20, Busan, S. Korea, 2023 (on behalf of Jack Wells)
- Platform presentation: "Fast STEM Simulation Technique to Improve Quality of Inpainted Experimental Images Through Dictionary Transfer", Microscopy and Microanalysis, Minneapolis, USA, 2023
- Poster presentation: "Exploring Low-dose and Fast Electron Ptychography using l_0 Regularisation of Extended Ptychographical Iterative Engine", Microscopy and Microanalysis, Minneapolis, USA, 2023 (on behalf of Amirafshar Moshtaghpour)
- Poster presentation: "Advances in Probe Subsampling for 4D-STEM", MMC/EMAG, Manchester, UK, 2023
- Platform presentation: "*In silico* Ptychography of Lithium-ion Cathode Materials from Subsampled 4-D STEM Data", ISCS, Luxembourg, 2023
- Platform presentation: "Compressed 4-D STEM: From Conception to Implementation", MRS Fall Meeting, Boston, USA, 2022
- Platform presentation: "Compressed STEM Simulations", Microscopy and Microanaly-

sis, Portland, USA, 2022

1.4 Collaborations

- Collaborating with Prof. Angus I. Kirland, Dr. Amirafshar Moshtaghpour, and Dr. Abner Velasco at the Rosalind Franklin Institute in Oxford, UK. Part of this collaboration aims to progress the understanding of 4D-STEM for biological samples, and how CS may help with the reduction of beam influence. There are currently two journal papers pending submission based on the research together, as well as two accepted conference papers.
- Collaborating with Miaofang Chi at Oak Ridge National Laboratory, TN, USA. This collaboration aims to develop more understanding of how CS can be applied to focused-probe 4-D STEM. This has resulted in a conference paper so far.
- Collaborating with Dr. Ian MacLaren at the University of Glasgow, Glasgow, UK. This collaboration is working on techniques for rapid CBED acquisition for 4-D STEM.
- Collaborating with Prof. Roland Fleck at Kings College London, London, UK. This collaboration developed methods for applying CS to Focussed Ion Beam-Scanning Electron Microscopy (FIB-SEM) tomography. This has resulted in a conference paper.
- Collaborating with Dr. Giuseppe Nicotra at the Institute for Microelectronics and Microsystems, Catania, Italy. This collaboration involves the application of CS to multi-dimensional electron microscopy data acquisition, specifically simultaneous EELS and 4-D STEM acquisition.

1.5 Contributions

- The development of compressive sensing for STEM simulations to increase the speed of calculation. This work is demonstrated for state of the art algorithms such as the multislice and PRISM methods.
- The R-LMI inpainting algorithm for fast image recovery.
- A theoretical determination of the lower bound for patch size selection of sub-sampled STEM data for the BPFA algorithm.

- Development of compressive sensing for 4-D STEM. This thesis outlines strategies for the recovery of sub-sampled 4-D STEM data, as well as the application of both iterative and non-iterative algorithms.
- Improving the quality of iterative ptychography algorithms through sparsity promoting regularization. This work was done in collaboration with Amirafshar Moshtaghpour.
- A method to improve the resolution of experimental STEM using a technique known as dictionary transfer from simulated STEM data.
- The implementation of live inpainting of experimental STEM data during acquisition.

2 | Methods & Background

2.1 Overview

The intention of this chapter is to give a background to theory which is presented in the remaining of this work. The fundamental physics which underpins the existence of electron microscopy is discussed and derived, followed by its application to S/TEM. This chapter serves as a basis for the motivation of the remaining chapters, such as why CS methods can provide a solution to the underlying issues within STEM such as beam damage and long acquisition times in multi-dimensional acquisition.

2.2 Electrons, Scattering, and Theory

The humble electron sits proudly within the standard model of particle physics as a fundamental building block of matter. First discovered by J. J. Thomson in 1897 [9, 10], the electron has since been studied as well as any existing particle thanks to its observable interactions with the electromagnetic field. The photon, the massless quantum particle which pops into existence because of the electromagnetic field, was postulated back in 1905 by Albert Einstein [11]. The photoelectric effect was perhaps a catalyst for the study of so-called *matter waves* so that in Louis de Broglie's 1924 PhD thesis [12], he theorised that all matter can behave as both a wave and a particle– including the electron. It is important to note that the behaviour of something does not constitute what that thing is, it is our observations of macroscopic mechanics that determine these definitions. What an electron is, for all intents and purposes, is an excitation of the Dirac field [13]. It just so happens that the electron can potentially interact with other quantum fields to exhibit certain properties [14].

One important property is the mass of an electron and therefore its momentum when mov-

ing at a certain velocity. In a S/TEM, electrons are typically accelerated by a voltage, E, on the order of $60 - 300$kV. This means that their velocities, v, are on the order of the speed of light, c, which is given as

$$v = c\sqrt{1 - \left(1 + \frac{eE}{m_e c^2}\right)^{-2}} \tag{1}$$

where e is the elementary charge of an electron, and m_e is the rest mass of an electron. When combined with the (relativistically corrected) de Broglie equation reads

$$\lambda = \sqrt{1 - \frac{v^2}{c^2}} \frac{h}{m_e v} \tag{2}$$

where λ is the (relativistically corrected) electron wavelength and h is the Planck constant. For an electron accelerated by 300kV, this yields a wavelength of 1.97pm, roughly 50 times smaller than the radius of an atom. It is this that allows electrons to probe smaller dimensions that that of photons (specifically x-rays).

However, this view of electrons can be limiting when considering electron interactions. Another interpretation is based on the theory found in Richard P. Feynman's PhD thesis– the path integral formulation. What makes this powerful is that one does not have to interpret the electron a wave, and can remain in a particle based regime.

2.2.1 First principles and the Feynman path integral approach

The path integral approach is based upon the principle of least action [15], whereby a particle is most likely to take the path which minimises its action, *i.e.*, $\int KE - V dt$ where KE is the kinetic energy along a path for infinitesimal time intervals dt and V is the potential energy. Consider a single electron emitted from the source. For simplicity, it is assumed that the source is to be point-like in space to avoid ambiguity on its initial starting conditions. The electron is then accelerated, it interacts with the sample, and then it is measured on a 2-D detector at an arbitrary distance beyond the sample. From the observer's perspective, the electron left the source and then hit the detector, what it did in-between is undefined. By Feynman's theory [16], the electron actually took *all* paths from the source to the detector- interacting with the potential induced by electromagnetic lenses, the sample, and the field which itself induces.

It is important that this does happen, since it is this which gives rise to coherence and

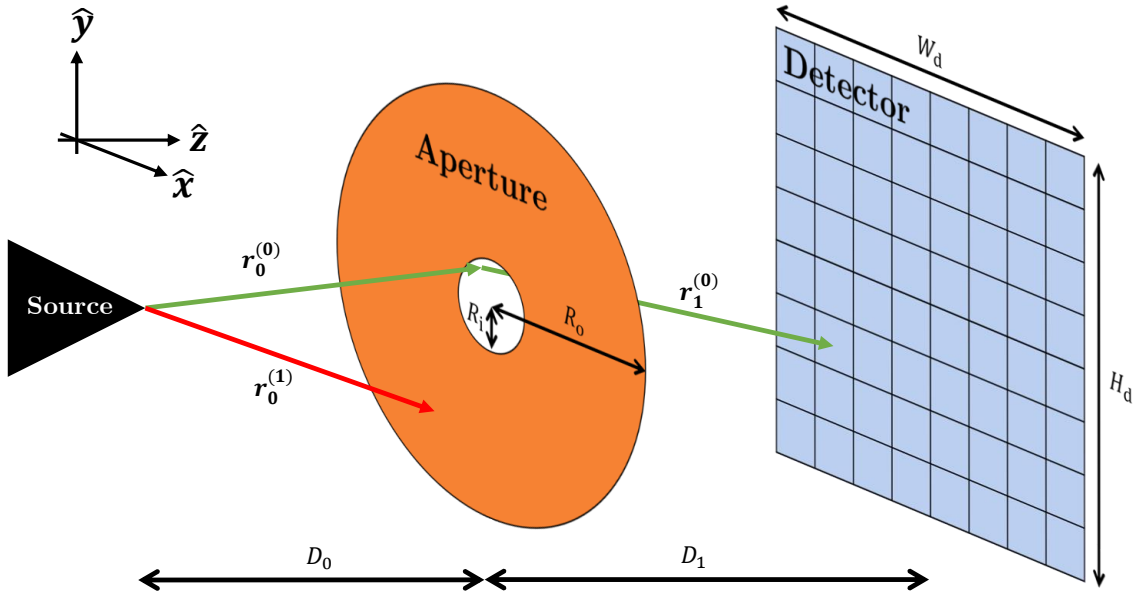

Figure 2.1: **Experimental set-up for simplistic view of Feynman Path Integral approximation to electron scattering.** Electrons leave the source (left) and travel along an initial step from the source to $\boldsymbol{r}_0^{(i)}$ to approach the plane of the aperture. If the electrons position is within the slit then it can take another step from $\boldsymbol{r}_0^{(i)}$ to $\boldsymbol{r}_1^{(i)}$ towards the detector, otherwise the combined path $\boldsymbol{r}^{(i)}$ does not contribute. The basis is indicated by the mutually orthogonal unit vectors $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{y}}$, $\hat{\boldsymbol{z}}$.

therefore phase contrast. In a simple example, consider the electron which is confronted by a spherical aperture as depicted in Fig. 2.1.

Assume an aperture at a distance D_0 from a point-like source and a detector at a distance D_1 from the aperture, where the position of the source is at location $\mathbf{0}$. The aperture has an inner radius R_i and an outer radius of R_o, such that initial steps to $\boldsymbol{r}_0^{(i)}$ can pass through the aperture if the amplitude of the intersection at the aperture falls within $\{0,\ R_\mathrm{i}\}$ (see $\boldsymbol{r}_0^{(0)}$ in Fig. 2.1) where the point of intersection is given as $h_\mathrm{a} \in \mathbb{R}^{1\times H_\mathrm{a}}$ and $w_\mathrm{a} \in \mathbb{R}^{1\times W_\mathrm{a}}$. If this criteria is not met, then it is ignored from calculation (see $\boldsymbol{r}_0^{(1)}$ in Fig. 2.1).

The second step *i.e.*, towards $\boldsymbol{r}_1^{(i)}$ from $\boldsymbol{r}_0^{(i)}$ propagates the electron from the aperture to the detector, where a detector location is fixed by $h_\mathrm{d} \in \mathbb{R}^{1\times H_\mathrm{d}}$ and $w_\mathrm{d} \in \mathbb{R}^{1\times W_\mathrm{d}}$. Note that all paths are treated as independent, unlike in classical wave theory where the total wavefront is computed. Given this, the i'th path is then computed according to positions $\boldsymbol{r}_0^{(i)}$ and $\boldsymbol{r}_1^{(i)}$,

Parameter	Value
D_0 (m)	0.5
D_1 (m)	0.001
R_i (nm)	30

Table 2.1: **Parameters for simulation of electron interaction with a spherical aperture.** The values for each parameter corresponding to Fig. 2.1 for the single slit experiment. Accelerating voltage is varied to show its effect.

$$\boldsymbol{r_0^{(i)}} = w_\mathrm{a}^{(i)}\hat{\boldsymbol{x}} + h_\mathrm{a}^{(i)}\hat{\boldsymbol{y}} + D_0\hat{\boldsymbol{z}} \tag{3}$$

$$\boldsymbol{r_1^{(i)}} = (w_\mathrm{d}^{(i)} - w_\mathrm{a}^{(i)})\hat{\boldsymbol{x}} + (h_\mathrm{d}^{(i)} - h_\mathrm{a}^{(i)})\hat{\boldsymbol{y}} + D_1\hat{\boldsymbol{z}} \ . \tag{4}$$

The next process is to calculate the action along each of the paths. As previously mentioned, paths are only computed which pass through the aperture *i.e.*, if $\sqrt{w_\mathrm{a}^2 + h_\mathrm{a}^2} < R_\mathrm{i}$. The action $\boldsymbol{S^{(i)}}$ along the i'th path is calculated according to the Lagrangian $\boldsymbol{\mathcal{L}^{(i)}}$,

$$\boldsymbol{\mathcal{L}^{(i)}} = \frac{1}{2} m_e \frac{|\boldsymbol{r^{(i)}}|^2}{(\Delta t)^2} \ , \tag{5}$$

given that Δt is the time permitted for the electron to move from along one path. It is important to note that this implies that the electron is free to travel faster than the speed of light along certain paths. The penultimate process is then to calculate $\boldsymbol{S^{(i)}}$ where,

$$\boldsymbol{S^{(i)}} = \frac{1}{2} m_e \left[\frac{|\boldsymbol{r_0^{(i)}}|^2}{\Delta t_0} + \frac{|\boldsymbol{r_1^{(i)}}|^2}{\Delta t_1} \right] \ . \tag{6}$$

The relative probability of an electron hitting the detector at position $w_\mathrm{p}, h_\mathrm{p}$ is then given as;

$$P(w_\mathrm{p}, h_\mathrm{p}) = A \left| \sum_{i=1}^{n} \exp\left[\frac{j}{\hbar} \boldsymbol{S}^{(i)}_{(w_\mathrm{p}, h_\mathrm{p})} \right] \right|^2 \ . \tag{7}$$

For the simulations, the parameters are given in Table 2.1. The accelerating voltage is varied from 60kV up to 300kV, and the results are given in Fig. 2.2.

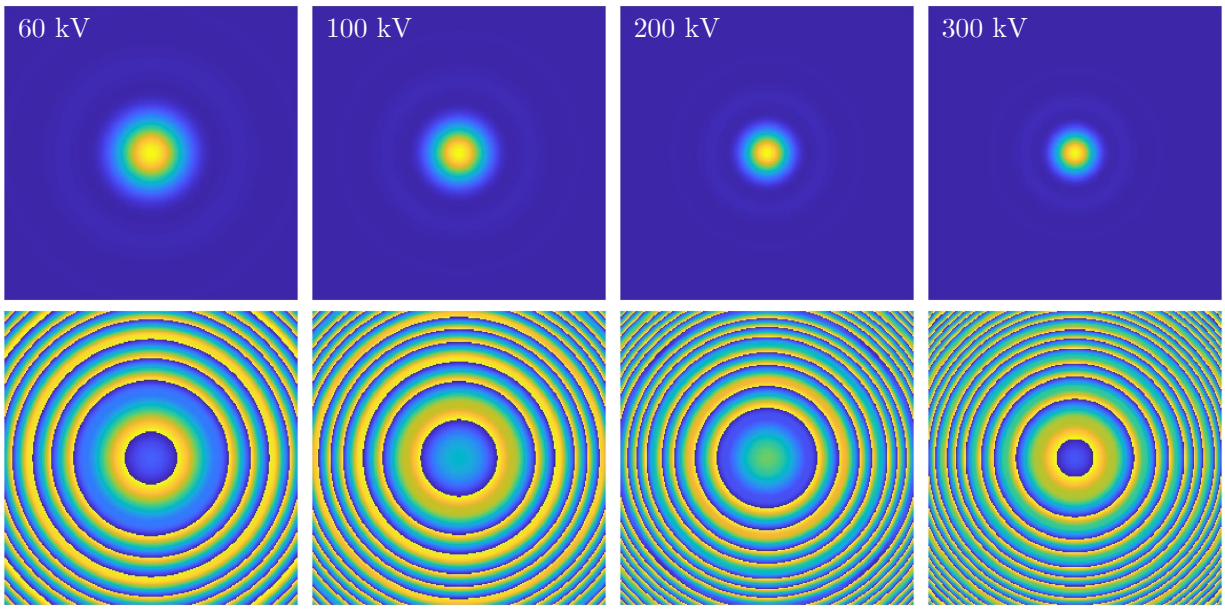

Figure 2.2: **Results of applying Feynman path integral approximation to the estimation of electron interaction with a spherical aperture.** The top row indicates the modulus square of the wave-function at the detector, whereas the bottom row indicates the corresponding phase at the detector.

By using the Feynman path integral, the wave-function can be approximated without requiring one to reinterpret the electron as anything other than a particle. This forms a basis by which comprehending image formation in S/TEM can be simplified to the likelihood an electron takes a certain path from the source to the detector. In the demonstration given, the problem is simplified to a two-step process, whereas in practice the number of steps and paths are infinite. This is a computationally heavy task to consider even one or two more steps, therefore the Feynman path integral approximation is generally avoided in typical wave-function calculations of this type.

2.2.2 The wave-function of a free electron

In S/TEM, it is vital to understand the electron wave-function prior to interacting with the sample and after interacting with the sample. This incident wave-function $\Psi_i \in \mathbb{C}$ determines the resulting image which is formed as a result of sample interaction [1]. To begin deriving the wave-function of the incident electron, the non-relativistic case is considered. The wave-function is initially assumed to be a function of space $\boldsymbol{r}$ and time t such that $\Psi_i \colon \boldsymbol{r},\ t \mapsto \Psi_i(\boldsymbol{r}, t)$.

The time-dependent Schrödinger equation is given as [17],

[1] Assuming infinite dose.

$$j\hbar\frac{\partial\Psi_{\mathrm{i}}}{\partial t} = -\frac{\hbar^2}{2m_e}\nabla^2\Psi_{\mathrm{i}} \ , \tag{8}$$

where ∇^2 is the Laplacian operator, $\hbar$ is the reduced Planck constant and m_e is the electron rest mass. Given that equation 8 is a function of time on the left-hand side of the equality, and a function of space on the right-hand side of the equality, a solution of the form $\Psi_{\mathrm{i}}(\boldsymbol{r},t) = \psi_{\mathrm{i}}(\boldsymbol{r})F(t)$ is sought. Substituting this into equation 8 and simplifying derives the time-independent Schrödinger equation as follows,

$$\begin{aligned} j\hbar\psi_{\mathrm{i}}(\boldsymbol{r})\frac{\partial F(t)}{\partial t} &= \left[-\frac{\hbar^2}{2m_e}\nabla^2\right]\psi_{\mathrm{i}}(\boldsymbol{r})F(t) \\ \frac{j\hbar}{F(t)}\frac{\partial F(t)}{\partial t} &= \frac{1}{\psi_{\mathrm{i}}(\boldsymbol{r})}\left[-\frac{\hbar^2}{2m_e}\nabla^2\right]\psi_{\mathrm{i}}(\boldsymbol{r}) \\ &= E \ , \end{aligned} \tag{9}$$

where E is a constant with units of energy. The solution for the left-hand side of equation 9 is found by solving the ordinary partial differential equation which has a solution of the form $F(t) = \exp(\alpha t)$,

$$\begin{aligned} F(t) &= \exp(\alpha t) \\ \frac{\partial F(t)}{\partial t} &= \alpha\exp(\alpha t) \\ &= -\frac{jE}{\hbar}\exp(\alpha t) \\ \implies \quad \alpha &= -\frac{jE}{\hbar} \end{aligned} \tag{10}$$

This returns the solution for the wave-function $\Psi_{\mathrm{i}}(\boldsymbol{r},t) = \psi_{\mathrm{i}}(\boldsymbol{r})\exp(-\frac{jE}{\hbar}t)$, which has observations given by $\Psi_{\mathrm{i}}(\boldsymbol{r},t)\overline{\Psi_{\mathrm{i}}(\boldsymbol{r},t)} = |\psi_{\mathrm{i}}(\boldsymbol{r})|^2$ where the bar notation indicates the complex conjugate. The right-hand side of equation 9 is then written as,

$$E\psi_{\mathrm{i}}(\boldsymbol{r}) = -\frac{\hbar^2}{2m_e}\nabla^2\psi_{\mathrm{i}}(\boldsymbol{r}) \ . \tag{11}$$

The above equation is assumed to have a solution of the form,

$$\psi_{\mathrm{i}}(\boldsymbol{r}) = \exp\left(j\boldsymbol{k}\cdot\boldsymbol{r}\right) \ , \tag{12}$$

such that,

$$\nabla^2\psi_{\mathrm{i}}(\boldsymbol{r}) = -k^2\psi_{\mathrm{i}}(\boldsymbol{r}) \tag{13}$$

where $k^2 = \frac{2Em_e}{\hbar^2}$ *i.e.*, $E = \frac{k^2\hbar^2}{2m_e}$. From this, a general solution to equation 11 is given as,

$$\psi_{\mathrm{i}}(\boldsymbol{r}) = A\exp\left[j(\boldsymbol{k}\cdot\boldsymbol{r} + \phi_0)\right] \ , \tag{14}$$

where ϕ_0 is a arbitrary linear phase shift corresponding to the initial conditions, A is an amplitude, and the probability P of finding the electron in the region $(\boldsymbol{r},\ \boldsymbol{r}+\Delta\boldsymbol{r})$ is,

$$P_{(\boldsymbol{r},\ \boldsymbol{r}+\Delta\boldsymbol{r})} = A^2\int_{\boldsymbol{r}}^{\boldsymbol{r}+\Delta\boldsymbol{r}} \psi_{\mathrm{i}}(\boldsymbol{r})\overline{\psi_{\mathrm{i}}(\boldsymbol{r})}d\boldsymbol{r} \ . \tag{15}$$

Equation 14 describes a plane wave, similar to that which illuminates a sample in TEM. If a detector was placed at the plane, the image formed would be exactly that described in 15. In practice, of course, a sample is introduced and a detector sits in the far-field some distance $L \in \mathbb{R}$ beyond the sample plane. The sample now influences the exit wave, and thus an image of the sample can be formed. Following on from the above descriptions, an image is simply the probability distribution that an electron hits a certain pixel, such that if one electron was emitted and detected, overtime the image would form.

2.2.3 Sample influence

Here, the quantum mechanical description of electron-specimen interaction is given and follows that described in [18]. For a fast electron incident on a crystalline sample, the Schrödinger equation is written as,

$$i\hbar\frac{\partial\psi(\boldsymbol{r},\boldsymbol{\rho},t)}{\partial t} = \left[-\frac{\hbar^2}{2m_e}\nabla^2_{\boldsymbol{r}} + H_c(\boldsymbol{\rho}) + H^{'}(\boldsymbol{r},\boldsymbol{\rho})\right]\psi(\boldsymbol{r},\boldsymbol{\rho},t) \ , \tag{16}$$

where $\boldsymbol{r}$ is the coordinate of the incident electron at time t, and $\boldsymbol{\rho}$ denotes the set of particles in the solid. The $-\frac{\hbar^2}{2m_e}\nabla^2_{\boldsymbol{r}}$ term is the kinetic energy operator, $H_c(\boldsymbol{\rho})$ is the Hamiltonian corresponding to the particles within the sample, and $H^{'}(\boldsymbol{r},\boldsymbol{\rho})$ is the Hamiltonian corresponding to the electron-specimen interaction.

Furthermore, assume that the wave-function can be written as follows,

$$\psi(\boldsymbol{r},\boldsymbol{\rho},t) \rightarrow \psi(\boldsymbol{r},\boldsymbol{\rho})F(t) \ , \tag{17}$$

such that the spatial and temporal components can be separated. This updates equation 16 as follows,

$$\begin{aligned} i\hbar\psi(\boldsymbol{r},\boldsymbol{\rho})\frac{\partial F(t)}{\partial t} &= \Big[-\frac{\hbar^2}{2m_e}\nabla^2_{\boldsymbol{r}} + H_c(\boldsymbol{\rho}) + H^{'}(\boldsymbol{r},\boldsymbol{\rho})\Big]\psi(\boldsymbol{r},\boldsymbol{\rho})F(t) \\ \frac{i\hbar}{F(t)}\frac{\partial F(t)}{\partial t} &= \frac{1}{\psi(\boldsymbol{r},\boldsymbol{\rho})}\Big[-\frac{\hbar^2}{2m_e}\nabla^2_{\boldsymbol{r}} + H_c(\boldsymbol{\rho}) + H^{'}(\boldsymbol{r},\boldsymbol{\rho})\Big]\psi(\boldsymbol{r},\boldsymbol{\rho}) \\ &= E \end{aligned} \tag{18}$$

where E is the total energy of the system (*i.e.*, constant). The solution is then the time independent Schrödinger equation given as,

$$\Big[-\frac{\hbar^2}{2m_e}\nabla^2_{\boldsymbol{r}} + H_c(\boldsymbol{\rho}) + H^{'}(\boldsymbol{r},\boldsymbol{\rho})\Big]\psi(\boldsymbol{r},\boldsymbol{\rho}) = E\psi(\boldsymbol{r},\boldsymbol{\rho}) \ . \tag{19}$$

The next step is to consider the meaning of $\psi(\boldsymbol{r},\boldsymbol{\rho})$ in more detail. This term is the wave-function of the system, *i.e.*, a superposition of wave-functions which can exist given discrete stationary states of the crystal. The wave-functions corresponding to these crystal states are denoted $a_m(\boldsymbol{\rho})$ (borrowed from [18]), such that when the crystal Hamiltonian $H_c(\boldsymbol{\rho})$ operates on these wave-functions, the energy of the stationary states is the eigenvalue of the operation, ϵ_m. This is summarised mathematically as,

$$\psi(\boldsymbol{r},\boldsymbol{\rho}) = \sum_m \phi_m(\boldsymbol{r}) a_m(\boldsymbol{\rho}) \ , \tag{20}$$

$$H_c(\boldsymbol{\rho}) a_m(\boldsymbol{\rho}) = \epsilon_m a_m(\boldsymbol{\rho}) \tag{21}$$

for $m \in M$ which is the set of possible stationary states. Assuming the initial state of the crystal to be $a_i(\boldsymbol{\rho})$ where $i \in M$, then this implies that $\phi_i(\boldsymbol{r})$ is the wave-function of the initial electron after an elastic scattering event, *i.e.*, the crystal does not change its energy state. The energy of this electron is then given as,

$$E_i = E - \epsilon_i \ . \tag{22}$$

In the case of inelastic scattering, the wave-function of the crystal is changed from $a_i(\boldsymbol{\rho})$ to $a_j(\boldsymbol{\rho})$ $(i \neq j)$, which in turn corresponds to an electron in state $\phi_j(\boldsymbol{r})$. The energy of this electron is then given as,

$$E_j = E - \epsilon_j = \frac{h^2}{2m_e} k_j^2 \ , \tag{23}$$

where k_m is the magnitude of the wave vector of the electron which has been scattered. Combining equations 22 and 23 means that the total energy loss of the incident electron corresponding to the inelastic excitation is therefore,

$$E_{\text{loss}} = E_i - E_j = \epsilon_j - \epsilon_i \ . \tag{24}$$

Of course, different scattering mechanisms within electron microscopy cause various signals to arise from the crystal. An inelastic event can result in various modes of excitation such as the excitation of electrons in orbitals, phonon excitation, or plasmon excitation. These scattering mechanisms shall be discussed in the following section, having now concluded the basic quantum mechanical description of electron-specimen interaction.

2.2.4 Electron scattering theory

As previously discussed, electrons will interact with the sample and as a result will potentially change their energy state. Since the energy of the system must be conserved, this gives rise to various signals which can be measured using specific detectors. Not all interactions are equal, and as such the likelihood of some interactions is more than that of others. In this section, elastic and inelastic scattering shall be discussed without considering the instrument. From a philosophical standpoint, the theory shall give rise to the experiment and the tool to perform these tasks.

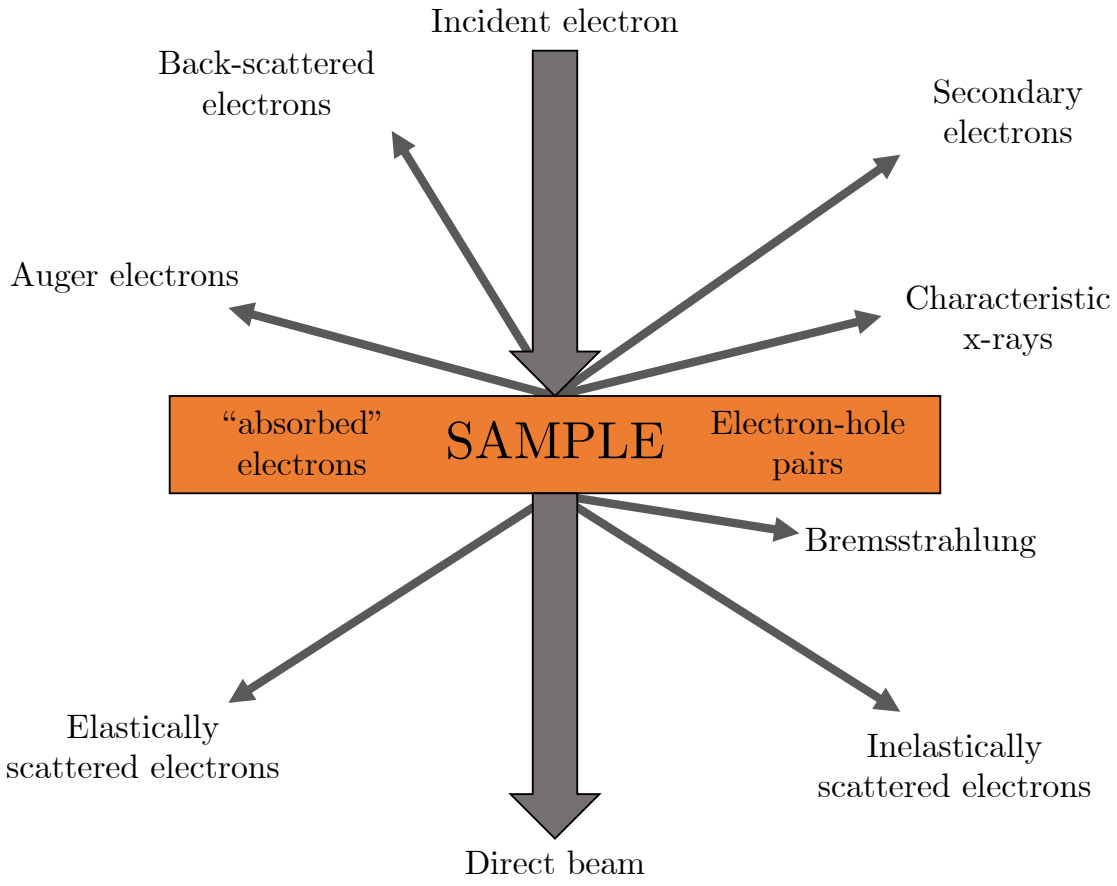

Figure 2.3: **Graphical description of possible electron scattering mechanisms when incident onto a sample.** Electrons can scatter in various ways when incident upon a sample, and the likelihood of these mechanisms are based upon the scattering cross-section for each mechanism. Figure inspired by Williams and Carter (1996), Fig. 1.3.

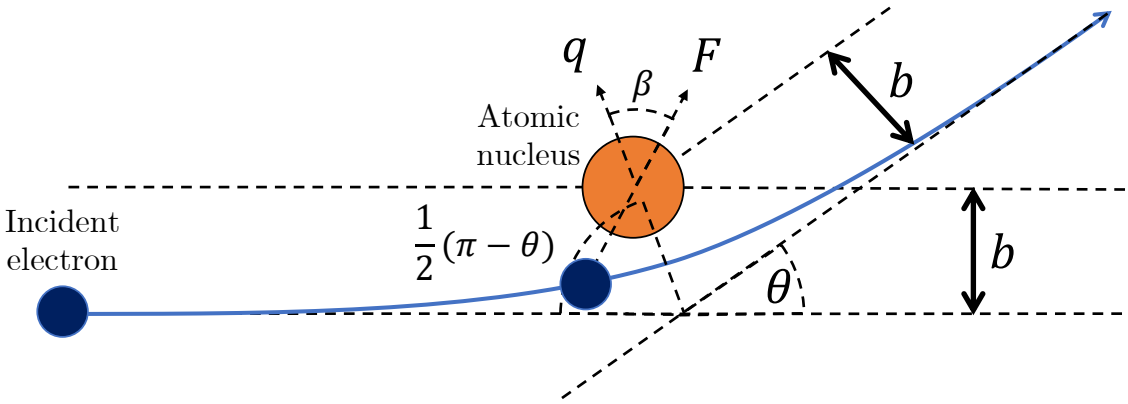

Figure 2.4: **Diagram showing the elastic scattering interaction when an electron approaches the nucleus of an atom.** As the electron approaches the atom, it is drawn towards it due to the Coulomb force. This causes the electrons path to deviate, with the strength of that deflection being proportional to the impact parameter, b.

Elastic scattering

Elastic scattering is defined as the conservation of the kinetic energy and momentum within the system. Elastic scattering arises from the electron-atom interaction (although dominated by the influence of the nucleus) and appropriate models can be used to determine the scattering cross-sections for both low-angle and high-angle scattering. High-angle scattering is often referred to as *quasi-elastic* since a small amount of energy is generally lost through phonon scattering. In this section, a derivation of the Rutherford scattering cross-section shall be given in the context of electron-nuclei interaction, then extended to include appropriate relativistic corrections and electron-electron shielding.

High-angle quasi-elastic scattering

The following derivation follows that for alpha-particle Rutherford scattering found in [19] with modifications for terms to account for electron interactions. Suppose a non-relativistic electron travelling at a velocity $\boldsymbol{v} \in \mathbb{R}$ is incident upon a stationary atom with mass number Z, as shown in Fig. 2.4. The electron is initially at a position vector $\boldsymbol{r}_{\mathrm{e}}^{\mathrm{initial}} = -r_x\boldsymbol{i} - b\boldsymbol{j}$ and the atom is assumed to be at the origin. To begin, only the interaction between the incident electron and the nucleus of the atom shall be considered. The initial kinetic energy $T \in \mathbb{R}$ of the system is therefore,

$$T = \frac{1}{2}m_{\mathrm{e}}v^2 \ , \tag{25}$$

where $v = |\boldsymbol{v}|$ is the magnitude of the electrons velocity. The Coulomb potential $V \in \mathbb{R}$ between the electron and the incident nucleus is given as,

$$V = -\frac{Ze^2}{r}\frac{1}{4\pi\epsilon_0} \ . \tag{26}$$

The next consideration is to derive the distance of closest approach $D \in \mathbb{R}$, *i.e.*, where the kinetic energy of the electron is equal to the Coulomb potential. This implies that,

$$T = -\frac{Ze^2}{4\pi\epsilon_0 D} \tag{27}$$

$$\rightarrow D = -\frac{Ze^2}{4\pi\epsilon_0 T} \ . \tag{28}$$

As previously stated, the momentum of the system must also be conserved, which also includes the angular momentum. The angular momentum is given as the cross-product of the displacement vector $\boldsymbol{r}$ and the momentum vector $\boldsymbol{p}$. The initial angular momentum $L \in \mathbb{R}$ is given as a function of the positions and momenta $\boldsymbol{p}_i$ of the electron and nucleus,

$$L = \sum_i \boldsymbol{r}_i \times \boldsymbol{p}_i \ , \tag{29}$$

where $\times$ denotes the cross product. This implies that the initial angular momentum of the system is,

$$\begin{aligned} L &= \boldsymbol{r}_{\mathrm{e}}^{\mathrm{initial}} \times m_{\mathrm{e}}\boldsymbol{v} \\ &= m_{\mathrm{e}} v b \ . \end{aligned} \tag{30}$$

At any given moment during the electrons approach towards the nucleus, the Coulomb force acts on the electron to change its angular momentum, where the angle by which the electron's momentum is change is given as $\beta \in \mathbb{R}$. At the point of closest approach, the momentum change is given by the vector $\boldsymbol{q} \in \mathbb{R}$, which has a magnitude $q \in \mathbb{R}$. Given that the initial and final momenta must be conserved, and that the nucleus' momentum is assumed to

be zero and unaltered, the initial momentum of the electron $\boldsymbol{p}^{\text{initial}}$, the final momentum of the electron $\boldsymbol{p}^{\text{final}}$ and the vector $\boldsymbol{q}$ are therefore related to the scattering angle by,

$$\boldsymbol{p}^{\text{initial}} = p\boldsymbol{i} \tag{31}$$

$$\boldsymbol{p}^{\text{final}} = p\cos(\theta)\boldsymbol{i} + p\sin(\theta)\boldsymbol{j} \tag{32}$$

$$\boldsymbol{q} = q\cos\left(\frac{1}{2}(\pi-\theta)\right)\boldsymbol{i} + q\sin\left(\frac{1}{2}(\pi-\theta)\right)\boldsymbol{j} \ , \tag{33}$$

which implies that,

$$p\sin(\theta) = q\sin\left(\frac{1}{2}(\pi-\theta)\right) \ . \tag{34}$$

The Coulomb force $F \in \mathbb{R}$ which is felt by the electron due to the nucleus is given as,

$$F = -\frac{Ze^2}{r^2}\frac{1}{4\pi\epsilon_0} \ , \tag{35}$$

and by combining this with equation 27 yields,

$$F = \frac{TD}{r^2} \ . \tag{36}$$

The component of this force along the direction of $\boldsymbol{q}$ is therefore,

$$F_{\boldsymbol{q}}(t) = \frac{TD}{r^2}\cos\beta(t) \ . \tag{37}$$

Given that a force along a vector is equal to the rate of change along that vector with respect to momentum, the value of q is the integral of equation 37,

$$q = \int \frac{TD}{r^2}\cos\beta(t)\,dt \ . \tag{38}$$

To solve this integral, it is helpful to reformulate the displacement vector by assuming that $\boldsymbol{r}(t) = r\cos(\beta(t))\boldsymbol{i} + r\sin(\beta(t))\boldsymbol{j}$. Given this, the momentum can be written as,

$$\begin{aligned}\boldsymbol{p} &= m_{\mathrm{e}}\frac{d}{dt}(\boldsymbol{r}(t)) \\ &= m_{\mathrm{e}}\frac{d}{dt}\big[r\cos(\beta(t))\boldsymbol{i} + r\sin(\beta(t))\boldsymbol{j}\big] \\ &= m_{\mathrm{e}}r\dot{\beta}\big[-\sin(\beta(t))\boldsymbol{i} + \cos(\beta(t))\boldsymbol{j}\big] \ . \end{aligned} \tag{39}$$

By the definition of angular momentum given earlier, the angular momentum can be written in terms of $\dot{\beta}$ as,

$$\begin{aligned} L &= \boldsymbol{r} \times \boldsymbol{p} \\ &= m_{\mathrm{e}}r^2\dot{\beta}\Big[\cos^2(\beta(t)) + \sin^2(\beta(t))\Big] \\ &= m_{\mathrm{e}}r^2\dot{\beta} \ . \end{aligned} \tag{40}$$

It is now possible to solve equation 38 through a substitution method, *i.e.*, $dt = \frac{d\beta}{\dot{\beta}}$ and given that angular momentum must be conserved by combining equations 30 and 40, $\dot{\beta}$ can be written as,

$$\dot{\beta} = \frac{bm_{\mathrm{e}}v}{m_{\mathrm{e}}r^2} \ . \tag{41}$$

The integral in equation 38 then becomes,

$$\begin{aligned} q &= \frac{TD}{bv}\int \cos(\beta)d\beta \\ &= \frac{TD}{bv}\sin(\beta) \ , \end{aligned} \tag{42}$$

with limits $\beta \in [-\frac{1}{2}(\pi-\theta), \frac{1}{2}(\pi-\theta)]$. Equation 42 is then reduced to,

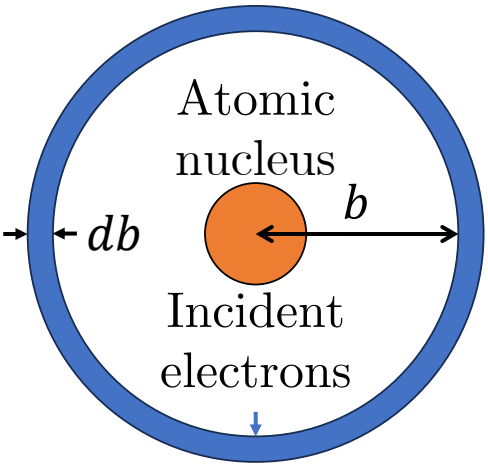

Figure 2.5: **Schematic for demonstrating electron flux in the context of impact factor.** The radial symmetry of the problem imposes that electrons passing by the nucleus with the same distance have the same impact parameter. This gives rise to the 2π factor in equation 46.

$$\begin{aligned} q &= \frac{TD}{bv}\Big[\sin(\beta)\Big]_{-\frac{1}{2}(\pi-\theta)}^{\frac{1}{2}(\pi-\theta)} \\ &= \frac{TD}{bv} 2\sin\left(\frac{1}{2}(\pi-\theta)\right) \ , \end{aligned} \tag{43}$$

which when combined with equation 34 and that the kinetic energy of the electron can be written as $T = p^2/2m_e$ forms the following equality,

$$\begin{aligned} 2p\sin\left(\frac{\theta}{2}\right) &= \frac{TD}{bv} 2\sin\left(\frac{1}{2}(\pi-\theta)\right) \\ &= \frac{Dp}{b}\sin\left(\frac{1}{2}(\pi-\theta)\right) \\ &= \frac{Dp}{b}\Big[\sin(\pi/2)\cos(\theta/2) - \sin(\theta/2)\cos(\pi/2)\Big] \\ \rightarrow \sin\left(\frac{\theta}{2}\right) &= \frac{D}{2b}\cos\left(\frac{\theta}{2}\right) \qquad (44) \\ \rightarrow \tan\left(\frac{\theta}{2}\right) &= \frac{D}{2b} \ . \qquad (45) \end{aligned}$$

Equation 45 is therefore a function to generate the scattering angle based on the distance of closest approach (which is a function of the kinetic energy of the electron) and the impact parameter b. From here, a scattering cross section can be derived.

Flux refers to the density of incident particles (in this case, electrons) passing through a unit area per unit time. It is a measure of the number of electrons impacting a specific target area in a given time interval. The diagram given in Fig. 2.4 has radial symmetry about the axis passing through the centre of the nucleus, as demonstrated in Fig. 2.5, meaning that the magnitude of θ is the same for all cases where the magnitude of b is the same. Following this, the number of electrons dN passing the nucleus with impact factors between b and $b + db$ where db is an infinitesimally small shift in b is related to the flux Φ by,

$$\begin{aligned} dN &= \Phi\left[\pi(b+db)^2 - \pi b^2\right] \\ &\approx \Phi 2\pi b\, db \ . \end{aligned} \tag{46}$$

ignoring terms $\mathcal{O}(db^2)$. The goal is to generate a function which returns the number of elastically scattered electrons to a certain angle θ given a certain flux. Equation 45 provides a function which relates the scattering angle to the impact factor, hence its derivation was important. Rewriting equation 45 as

$$b = \frac{D}{2}\cot\left(\frac{\theta}{2}\right) \tag{47}$$

and then differentiating with respect to θ yields,

$$\begin{aligned} \frac{db}{d\theta} &= \frac{D}{2}\frac{d}{d\theta}\cot\left(\frac{\theta}{2}\right) \\ &= -\frac{D}{2}\frac{1}{\sin^2(\theta/2}\frac{d}{d\theta}\left(\frac{\theta}{2}\right) \\ &= -\frac{D}{4\sin^2(\theta/2)} \end{aligned} \tag{48}$$

which can then be substituted into equation 46 such that

$$
\begin{aligned}
dN(\theta) &= \Phi 2\pi b \frac{D}{4\sin^2(\theta/2)} d\theta \\
&= \Phi 2\pi \frac{D}{2} \cot\left(\frac{\theta}{2}\right) \frac{D}{4\sin^2(\theta/2)} d\theta \\
&= \Phi \pi D^2 \frac{\cos(\theta/2)}{4\sin^3(\theta/2)} d\theta \ . \qquad (49)
\end{aligned}
$$

where the negative sign has been dropped. In the context of elastic scattering, the differential cross-section $d\sigma/d\Omega$ refers to the likelihood of an electron interacting with a nucleus and being scattered into a specific solid angle Ω, and is defined mathematically as,

$$
\frac{d\sigma}{d\Omega} = \frac{1}{\Phi} \frac{dN}{d\Omega} \qquad (50)
$$

By integrating this over the region contained by the solid angle, a scattering cross-section can be extracted, and hence a measure of how likely that scattering is. The differential solid angle is related to the differential angle by,

$$
\begin{aligned}
d\Omega &= \sin(\theta) d\theta d\phi \\
&= 4\pi \sin\left(\frac{\theta}{2}\right) \cos\left(\frac{\theta}{2}\right) d\theta \qquad (51)
\end{aligned}
$$

where ϕ is the azimuthal angle and given that the function should be azimuthally independent for high-angle scattering *i.e.*, only a function of θ, the differential azimuthal angle is integrated over which yields a factor of 2π. Rearranging equation 51 and substituting into equation 49 then gives,

$$
\begin{aligned}
dN &= \Phi \pi D^2 \frac{\cos(\theta/2)}{4\sin^3(\theta/2)} \frac{d\Omega}{4\pi \sin(\theta/2)\cos(\theta/2)} \\
&= \Phi \frac{D^2}{16\sin^4(\theta/2)} d\Omega \ . \qquad (52)
\end{aligned}
$$

This equation can be rearranged and substituted into equation 50 to yield the scattering

cross section for high-angle scattering in terms of θ as,

$$\frac{d\sigma}{d\Omega} = \frac{D^2}{16\sin^4(\theta/2)} \ . \tag{53}$$

In order to account for relativity, the D term can be modified such that the kinetic energy term contains relativistic corrections. Given that $T = p^2/2m_\mathrm{e}$, a relativistically corrected version of p can be included instead. Also, given that the electron wavelength $\lambda = h/p$, a relativistically corrected wavelength λ_R can be used such that,

$$\begin{aligned} T &= \frac{p^2}{2m_\mathrm{e}} \\ &= \frac{h^2}{2m_\mathrm{e}\lambda_\mathrm{R}^2} \ , \end{aligned} \tag{54}$$

and substituting this into the D term,

$$\begin{aligned} D &= -\frac{Z\lambda_\mathrm{R}^2}{2\pi^2}\frac{e^2 m_\mathrm{e}\pi}{\hbar^2\epsilon_0} \\ &= -\frac{Z\lambda_\mathrm{R}^2}{2\pi^2 a_0} \end{aligned} \tag{55}$$

where a_0 is the Bohr radius. Substituting equation 55 into equation 53 then gives a relativistically corrected version of the high-angle scattering cross-section equation,

$$\frac{d\sigma}{d\Omega} = \frac{Z^2\lambda_\mathrm{R}^4}{64\pi^4 a_0^2}\frac{1}{\sin^4(\theta/2)} \ . \tag{56}$$

The final correction to make is related to the screening effect caused by the electron cloud surrounding the nucleus. The electron cloud can make the nucleus appear slightly less positive and this effect is exaggerated at distances further from the nucleus, meaning the effect is more important at lower scattering angles. The screening parameter is given without proof as

$$\theta_0 = \frac{0.117Z^{1/3}}{E_0^{1/2}} \ . \tag{57}$$

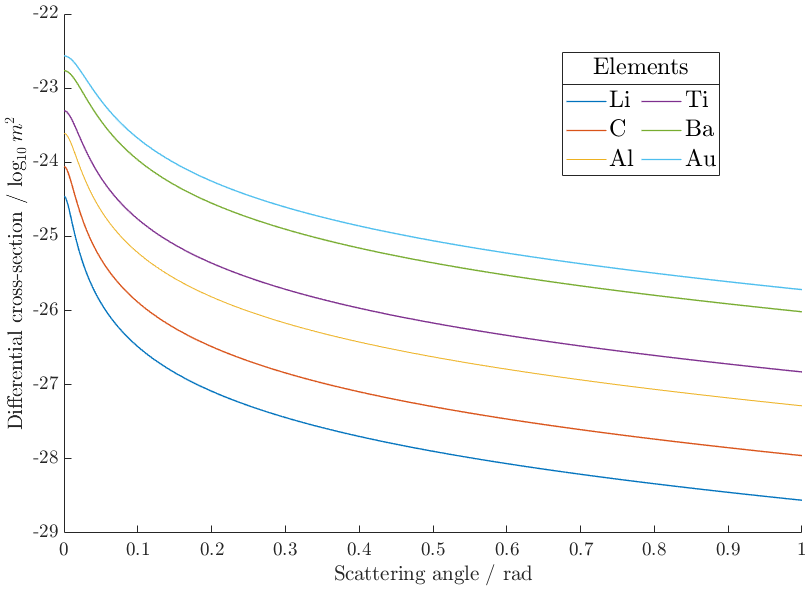

Figure 2.6: **Mott differential cross-section according to equation 59 as a function of scattering angle for various elements.**

The screening parameter is then used to modify the high-angle quasi-elastic scattering differential cross-section as follows,

$$\frac{d\sigma}{d\Omega} = \frac{Z^2\lambda_R^4}{64\pi^4 a_0^2} \frac{1}{\left[\sin^2(\theta/2) + (\theta_0/2)^2\right]^2} \quad . \tag{58}$$

The scattering cross-section can then be calculated by integrating equation 58 over appropriate limits. The cross-section derived is generally appropriate for high-angle scattering, assuming the beam voltage is $\approx 100keV$, and the nucleus has $Z < 30$. It is better to use what is known as the Mott cross-section to account for higher energies and heavier nuclei. The Mott cross-section simply extends the Rutherford cross-section by way of a linear correction term which reduces the magnitude of the cross-section for higher scattering angles. Without proof, this is given as,

$$\left(\frac{d\sigma}{d\Omega}\right)_{\text{Mott}} = \frac{Z^2\lambda_R^4}{64\pi^4 a_0^2} \frac{\cos^2(\theta/2)}{\left[\sin^2(\theta/2) + (\theta_0/2)^2\right]^2} \quad . \tag{59}$$

This cross-section then allows for measurement of high-angle quasi-elastic scattering, which is incoherent. However, the cross section is not appropriate where coherency effects change

the likelihood of scattering to certain angles through interference. Therefore, a different model must be used to account for this.

Low-angle elastic scattering

As previously mentioned, appropriate models can be used if the criteria fits such as in the case of high-angle scattering. For low-angle scattering, there are coherency effects which must be accounted for which eventually leads to diffraction and phase contrast imaging. The Rutherford model assumes the electron to be a particle, however, as is clear by now, the electron can also exhibit wave-like nature.

In order to account for this, the atomic scattering factor $f(\theta) \in \mathbb{C}$ is utilised. The scattering factor depends on:

- the wavelength of the incident electrons, λ
- the scattering angle, θ
- the atomic number, Z

and is used to calculate the scattering cross-section associated with low-angle elastic scattering. The atomic scattering factor is given without proof as,

$$f(\theta) = \frac{\left(1 + \frac{E_0}{m_e c^2}\right)}{8\pi^2 a_0} \left(\frac{\lambda}{\sin(\theta/2)}\right)^2 (Z - f_x(\theta)) \tag{60}$$

where $f_x(\theta)$ is the scattering factor associated with X-rays. The differential cross section is then the modulus squared of the atomic scattering factor. It can also be useful to describe the atomic scattering factor using the Mott-Bethe formula,

$$f(q) = \frac{1}{2\pi^2 a_0} \left(\frac{Z - f_x(q)}{q^2}\right) , \tag{61}$$

where q is the magnitude of the scattering vector *i.e.*, $q = \sin(\theta)/\lambda$. If the scattering vector q has units reciprocal Angstroms, then the atomic scattering factor has units Angstroms. The X-ray scattering factor is an effective shielding term associated with the influence of the surrounding electron cloud.

The atomic-scattering factor is related to the projected atomic potential $\boldsymbol{V}(\boldsymbol{r}) \in \mathbb{R}$ by the Born Approximation [20] given as,

$$f(\theta) = -\frac{2me}{h^2} \int d^3\boldsymbol{r} \exp\left(j\boldsymbol{q} \cdot \boldsymbol{r}\right) \boldsymbol{V}(\boldsymbol{r}) \ , \tag{62}$$

which is significant since the atomic potential can be approximated through the inverse Fourier transform of the atomic-scattering factor. Following from equation 12, the exit wave function from plane wave-scattering is a linear phase shift of this incident wave function according to,

$$\begin{aligned} \psi_{\mathrm{o}} &= A \exp\left(jkz + \phi\right) \\ &= A \exp\left(jkz\right) \exp\left(\phi\right) \\ &= A \exp\left(jkz\right) [\cos\left(\phi\right) + j \sin\left(\phi\right)] \end{aligned} \tag{63}$$

where ϕ is small, such that $\cos\left(\phi\right) \approx 1$ and $\sin\left(\phi\right) \approx \phi$. This reduces the above equation to,

$$\begin{aligned} \psi_{\mathrm{o}} &= A \exp\left(jkz\right)[1 + j\phi] \\ &= A \exp\left(jkz\right) + \phi A \exp\left(jkz + \pi/2\right) \end{aligned} \tag{64}$$

which implies that all elastically scattered electrons undergo a phase shift of $\pi/2$.

The structure factor $F(\theta) \in \mathbb{C}$ is an extension of the atomic scattering factor to account for combinations of atoms, such as crystal structures. This now introduces what are known as Miller indices, characterising the orientation of a crystal with respect to main axes of the crystal. Ultimately, the orientation determines the diffraction pattern and images which are formed due to scattering. The structure factor is defined as,

$$F(hkl) = \sum_i f_i(hkl) \exp\left(2\pi j(\boldsymbol{u} \cdot \boldsymbol{r})\right) \ , \tag{65}$$

where $\boldsymbol{u} = [h, k, l]$. The structure factor also determines the allowed reflections which can be observed in a given diffraction pattern. As a simple example, assume a face centred cubic structure. The basis or primitive translation vectors are given as,

- $\boldsymbol{r}_0 = (0,0,0)$
- $\boldsymbol{r}_1 = (0,1/2,1/2)$
- $\boldsymbol{r}_2 = (1/2,0,1/2)$
- $\boldsymbol{r}_3 = (1/2,1/2,0)$

and $f_j(hkl) = f \; \forall \; j \in \mathbb{N}^{[0,3]}$. The reciprocal lattice vectors are given by the following,

- $\boldsymbol{u}_1 = \frac{\boldsymbol{r}_2 \times \boldsymbol{r}_3}{V}$
- $\boldsymbol{u}_2 = \frac{\boldsymbol{r}_3 \times \boldsymbol{r}_1}{V}$
- $\boldsymbol{u}_3 = \frac{\boldsymbol{r}_1 \times \boldsymbol{r}_2}{V}$

where $V = \boldsymbol{r}_1 \cdot (\boldsymbol{r}_2 \times \boldsymbol{r}_3)$.The resulting structure factor is therefore,

$$F = f\Big[1 + \exp\left[j\pi(h+k)\right] + \exp\left[j\pi(h+l)\right] + \exp\left[j\pi(l+k)\right]\Big] \; . \tag{66}$$

The permitted reflections stem from this result, indicating that there are only specific cases where the modulus square of the structure factor is non-zero. Euler's formula shows that for $\exp(j\theta) = 1$, then $\theta = 2n\pi$ for $n \in \mathbb{Z}$. Conversely, $\exp(j\theta) = -1$ for $\theta = (2n+1)\pi$ for $n \in \mathbb{Z}$. Suppose that $h + k = 2n$, in order for this to be satisfied, h and k must both be odd, or both be even since the right hand side is always even. Now let l be odd and assume h and k were both even. This implies that $h + l$ must also be odd, and $k + l$ is also odd. The resulting structure factor would have a value of zero, *i.e.*, this condition is not a permitted reflection. Since the sign is somewhat arbitrary, the same argument can be made if h and k were assumed odd and l assumed even. Consider then if $h,\ k,\ l$ are all even. The result would satisfy that for all combinations, the resulting exponent would be even, and hence the structure factor non-zero. If they are all odd value then summing any pair must be even also and the same result is determined. The permitted reflections for a face centred cubic and the first five reflections are therefore $(1,\ 1,\ 1), (2,\ 0,\ 0), (2,\ 2,\ 0), (3,\ 1,\ 1), (2,\ 2,\ 2)$.

This is demonstrated in the diffraction pattern shown in Fig. 2.7 determined for polycrystalline gold sample, a standard FCC structure. The radii of the rings are determined by,

$$d_{hkl} = \frac{a_0}{\sqrt{h^2 + k^2 + l^2}} \; , \tag{67}$$

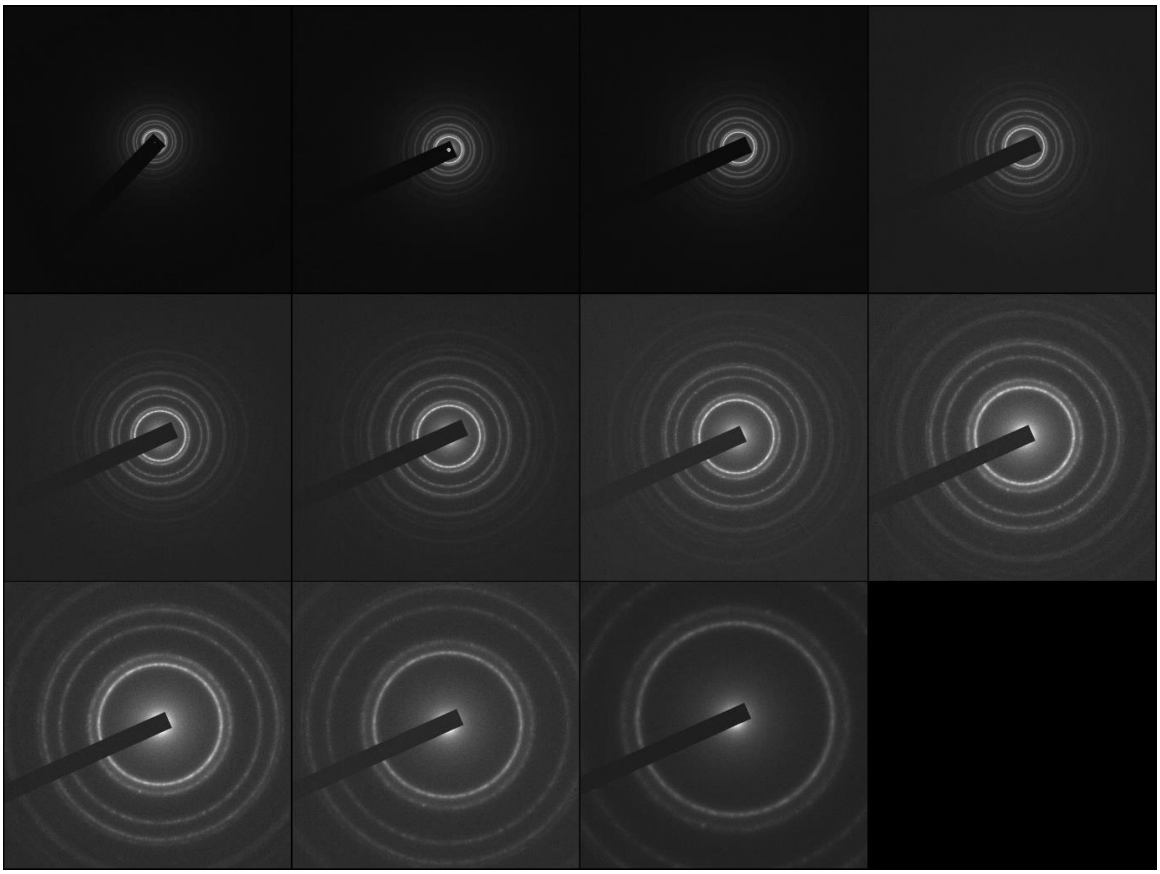

Figure 2.7: **Example diffraction patterns for polycrystalline gold sample demonstrating the permitted reflections.** Diffraction patterns acquired at different nominal camera lengths with increasing camera length left to right. 8cm, 10cm, 12cm and 15cm (top, left to right respectively), 20cm, 25cm, 30cm and 40cm (middle, left to right respectively), and 50cm, 60cm and 80cm (bottom, left to right respectively).

hence the ratios of the permitted reflection radii must satisfy this condition. The Bragg condition is given by,

$$\begin{aligned} n\lambda &= 2d_{hkl}sin\theta \\ &\approx 2\theta d_{hkl} \quad . \end{aligned} \tag{68}$$

where $n \in \mathbb{N}$ and θ is a Bragg angle. What this equation really says is that if the path difference between scattered and direct beams is some integer multiple of the wavelength, then they shall constructively interfere. Granted, this is diffraction basics, however the principle remains and owes to the quantum nature of electrons. To reiterate, the wavelength is related to the indeterminism on the electron position in space-time; electrons in phase are more likely to be found at a certain position (the diffraction spots) if the probability adds for that certain scattering vector.

Inelastic scattering

During an inelastic scattering process, the primary electron loses kinetic energy through interaction with the sample. This ultimately causes the energy state of the sample to increase, and the wavelength of the electron to increase. Given certain types of inelastic collisions, the excitation of the sample can lead to various other signals following de-excitation, or it can cause the sample to change unfavourably through beam damage. In this section, the various signals which can be collected through inelastic scattering shall be discussed.

Electron energy loss spectroscopy

Electron energy loss spectroscopy (EELS) is probably the most intuitive of these techniques as it is a direct measurement of how much energy an electron has lost as it passes through the sample. EELS is used to measure various properties such as the elemental composition of materials, the specimen thickness through energy filtered TEM (EFTEM), free electron density, valence states, band gap and nearest neighbour atomic structure [21–30]. EELS is especially useful for low Z-number materials [31] since the likelihood of inelastic scattering does not favor high Z-number elements. The scattering cross-section for EELS is related to the number of atoms per unit volume, and the energy loss function as described by Egerton [30]. Atomic resolution EELS has become a powerful tool in STEM for identifying chemical composition of these low Z-number atoms, which are weakly scattering in the Z-contrast regime.

Energy loss electrons can be described as having induced collective excitations (plasmon) or single excitations (low/core loss). Plasmons are the quantum pseudo-particle associated with the oscillation of valence electrons within a material [32]. Low loss single excitations of valence electrons generally occurs at less than 50eV through interband transition, *i.e.*, an electron in the valence or shallow core bands transitions to the conduction band. Core loss excitation corresponds to inner shell transitions, and these energy losses are characteristic as are the Auger electrons or x-ray photons emitted during the de-excitation.

A common analogy for understanding how the EELS spectra is produced is through how white light is separated into its components through a prism. Consider a thin sample, ignoring dynamical scattering effects, and assume that the incident electron is transmitted through the sample so that it can be collected by a spectrometer. Next, assume that the electron loses an amount of energy to the sample. As a result, the wavelength of this electron is increased,

its velocity decreased, and as such will be incoherent with respect to the direct beam. The transmitted electrons enter a magnetic prism, and given the Lorentz force, slower electrons (*i.e.*, less energetic) are dispersed more. The amount of dispersion is related directly to the energy, therefore by placing a detector after the prism, the number of electrons with a certain dispersion can be measured. A map can be formed within STEM mode, and reference spectra can be used to correlate the observed spectra to the position of the electron probe. This then forms a STEM-EELS map.

Energy dispersive x-ray spectroscopy

Energy dispersive x-ray spectroscopy (EDS) is another excellent demonstration of the practical implementation of quantum mechanics within experiment. Under the Bohr interpretation of atomic structure, electrons sit in discrete energy levels surrounding the nucleus of an atom. Each element has different permitted energy levels, and it is this uniqueness that allows spectral information to be interpreted as chemical composition [33]. Astronomers use the same principle to determine the composition of various objects in the universe, such as determining the weighting of elements within stars. Under certain conditions, an electron is able to gain sufficient energy that it can make a quantum leap to a higher energy level in the electronic shell structure. The atom itself is then excited, having an energy greater than the equilibrium of the ground state. After an amount of time, the atom will de-excite and the electron will transition back to a lower energy level. In the process, a photon is emitted which has the energy equivalent to the energy difference between the two states. In some cases where the energy transfer is sufficient, the electron will make several transitions known as a cascade. Regardless, the photon emitted has an energy which is unique to the atom which it came from and it is this that is used to characterise the chemical composition of the sample [34, 35].

Another x-ray signal, other than the one described, is known as bremsstrahlung x-rays [36]. When an electron changes momentum, and assuming the momentum change is sufficient, a photon within the x-ray band may be emitted [2]. This can happen through interaction of the electron with the Coulomb potential generated by the atomic nucleus. Bremsstrahlung x-rays are usually manifested as a noise which superimposes the spectra generated by energy dispersive x-rays.

Secondary electrons

[2]The photon can be of an arbitrary energy up to the energy of the incident electron.

When the electron beam interacts with the sample, the electrons within the conduction or valence bands may be sufficiently excited that they are ejected from the sample [37]. These electrons, typically with an energy less than 50eV, are then collected on a detector and an image formed [38]. Since these electrons are low energy, their mean free path is relatively short which makes secondary electron detection a surface sensitive imaging mode. Secondary electrons are mainly used as the primary detected signal in scanning electron microscopy (SEM) [39], although secondary electrons can be detected alongside transmitted signals using a low voltage STEM or SEM equipped with STEM detectors [40, 41].

2.3 Transmission Electron Microscopy

2.3.1 TEM

The TEM is a valuable tool for characterising complex materials, especially those contributing towards new technologies within electronics and medicine. In an era where computer chips are approaching manufacturing processes below a nanometre in scale, and the demand for high charge density battery materials is greater than ever before; analysis on the nanoscale is key to understand their properties.

Understanding these properties cannot be solved with one instrument alone, but rather a combination of nano-, micro-, and macro-scopic analyses are critical to give the scientist a fuller picture. The TEM is part of this characterisation chain, and its development from fundamental theory to present has perhaps been one of the catalysts for the modern world of today.

History of TEM

In the early 1930s, shortly after Louis de Broglie's PhD thesis was submitted, Ernst Ruska and Max Knoll developed the first TEM [42, 43] which was potentially motivated by the results of de Broglie's research, that electrons could exhibit wave-like properties [44, 45]. As has been addressed earlier, the electron is simply wave-like, and can exhibit the properties of a wave under specific conditions. This result meant that if an electron had a certain momentum, then it could conceivably have a wavelength on the order of (or shorter than) the atomic scale. This theoretical wavelength is given in equation 2, indicating that an electron accelerated by

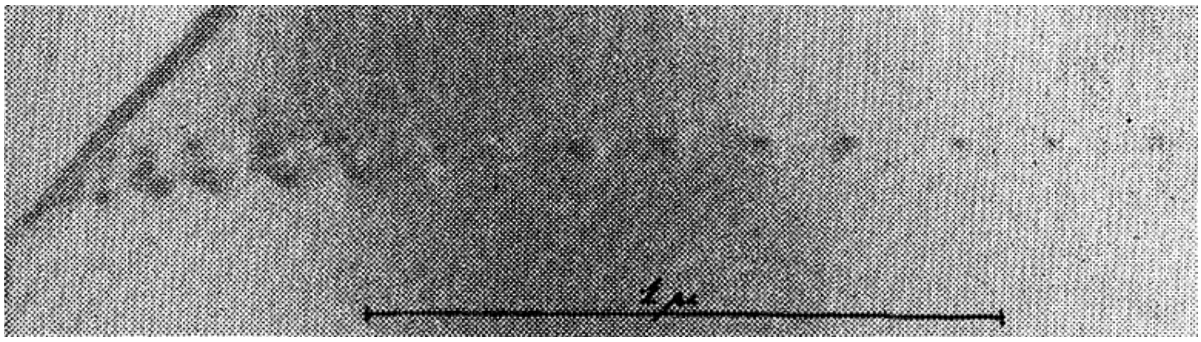

Figure 2.8: **Small angle grain boundary or slip band of a chrome-nickel steel thin film.** An early image taken from [1] showing how electron microscopy can image complex structures from thin films.

a 300kV has a theoretical wavelength of approximately 1.97pm, roughly twenty-five times smaller than the Bohr atomic radius [46–48].

A few years later, in 1936, the first commercially available TEM was built by Metropolitan-Vickers known as the *EM1* [49], and other manufacturers such as Siemens [50], Zeiss [51], and Hitachi [52], began producing their own TEMs around this time, and by 1944 the resolution was reduced to 2nm. By 1949, the Japan Electron Optics Laboratory (JEOL) produced the JEM-1, its first TEM [53], and is still producing some of the best electron microscopes available today.

Following on from this, around the 1950s, research focussed on the imaging of defect structures such as dislocations, twins, and stacking faults within thin films [1, 54–56]. Bollmann [1] investigated the small angle grain boundaries and slip bands within a chrome-nickel steel, and an example micrograph from this paper is shown in Fig. 2.8.

These early demonstrations paved the way towards modern materials characterisation and high-resolution TEM (HRTEM) was becoming a standard technique for analysing structures. The first sub-Å images of palladium and nickel were demonstrated in 1969 [57] using the axial illumination method. By isolating one of the diffracted beams through tilting the incident beam, a dark field image corresponding to this diffracted beam can be formed, exposing the lattice that gave rise to the specified diffracted beam [58].

Up to 1998, the limiting factor for HRTEM was aberrations. Simply put, aberrations cause the incident beam to have a different phase depending on the position of that incident wave-vector. As a result, the beam is non-uniform, giving rise to contrast which was not solely induced by the sample. In 1947, Otto Scherzer stated that spherical aberrations and chromatic

aberrations were unavoidable in TEM due to the electromagnetic lens design, but could be corrected for with hardware [59, 60]. The spherical aberration arises due to electrons crossing over at different depths along the optic axis, rather than at one single focal point. The lenses effectively deflect the electrons with varying strengths depending on the angle of incidence, and as such the image is distorted. Chromatic aberrations arise due to there being an energy spread amongst the electrons. Energy spreads imply a wavelength spread, leading to incoherence in the incident beam and varying cross over at the optic axis [61]. Hence, aberrations are simply a non-linear deflection of electrons, *i.e.*, the angle of incidence is not equal to the angle of deflection through the lens. At the initial cross-over, the source may be assumed to have a radius r_i, but the radius at the cross-over beyond the first lens has a minimum value r_c where $r_\mathrm{c} > r_\mathrm{i}$. In the ideal case, these two radii would be equivalent.

It is possible to quantify the coefficients of spherical and chromatic aberrations by $C_\mathrm{s} \in \mathbb{R}$ and $C_\mathrm{c} \in \mathbb{R}$ respectively, and the theoretical resolution which can be achieved according to each coefficient is estimated according to [62],

$$r(C_\mathrm{s}) = (0.12\lambda^3 C_\mathrm{s})^{1/4} \ , \tag{69}$$

$$r(C_\mathrm{c}) = \left(1.2\lambda \frac{\Delta E}{E} C_\mathrm{c}\right)^{1/2} \ . \tag{70}$$

As the accelerating voltage increases, spherical aberrations begin to dominate over chromatic aberrations. The ratio of spherical to chromatic aberration is approximately proportional to $\lambda^{-1/4}$, indicating that as the wavelength decreases (*i.e.*, accelerating voltage increases), the influence of spherical aberration will eventually outweigh the chromatic aberration. The first demonstration of spherical aberration correction for TEM was shown by Haider *et al.*in 1998 [63], where a hexapole corrector system was used to reduce the spherical aberration of the objective lens system. This ultimately improved the point resolution of the system from 0.28nm, down to less than 0.14nm.

A simple way to reduce the effects of chromatic aberration at low voltages is to use a source with a low energy spread. However, source coherency and cost are generally correlated, as highlighted in research by Quigley *et al.* [64].

Alternatively, a monochromator could be installed which filters dispersive electrons from the beam. Monochromators are relatively expensive, but can be beneficial especially for low-voltage (LV) imaging. This is highlighted in work by Bell *et al.* [65] which show thee first atomically resolved images at $40kV$. The work also highlights the benefits of LV imaging, such as reduced knock-on damage and improved contrast efficiency. This is especially useful for 2-D materials which are generally required to be imaged at LV to reduce the effects of knock-on damage, with the added benefit that they are inherently thin materials.

The TEM has developed significantly since its first construction, some 90 years ago. In more recent years, the advent of *in-situ* TEM, as well as gas-, liquid-, and cryo-stage TEM are being developed as methods for real-time analysis of complex dynamical structures [66–74].

TEM design

The design of a TEM can be broken into two key components, firstly the illumination system, followed by the imaging system. In this section, each shall be discussed as well as the importance of each for reliable image formation.

Illumination system

The illumination system is responsible for ensuring that the beam that interacts with the sample is as homogenous as possible, *i.e.*, flat. The beam should be approximately parallel (an extremely narrow convergence angle) and the electrons should be coherent (i.e., the electrons can interfere) and ideally *in phase* across the beam. In order to achieve this, electrons from a source must be manipulated so that they have these properties. The electrons are accelerated from the source through an acceleration tube, and then as is demonstrated in Fig. 2.9, cross-over prior to the first condenser lens, the C1 lens. The C1 lens strength is changed if the user wishes to change the electron flux through changing the probe width, typically known as *spot size*. By increasing the strength of this lens, the number of electrons which then pass into the second condenser lens (C2 lens) decreases. In typical low-dose TEM regimes, a large spot size is used to try and reduce beam induced damage.

Once the electrons have been manipulated by the C1 lens, they then pass into the C2 lens. The C2 lens and twin lens system typically work in tandem to control the convergence angle of the beam. This is typically done using the *brightness* control, and the cross-over of the beam nearest to the sample can be moved below or above the sample. During alignment, the beam

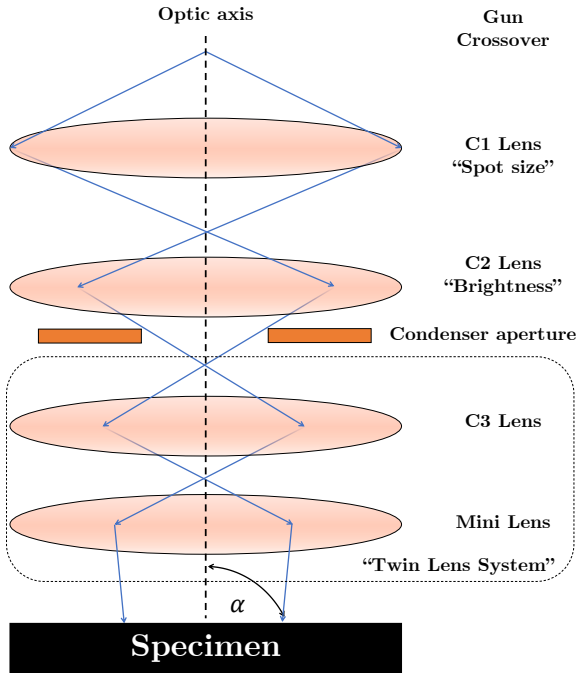

Figure 2.9: **Schematic for the illumination system of a TEM column.** The illumination system consists of a series of condenser lenses which aim to form an approximate parallel beam on the sample with a small convergence angle α.

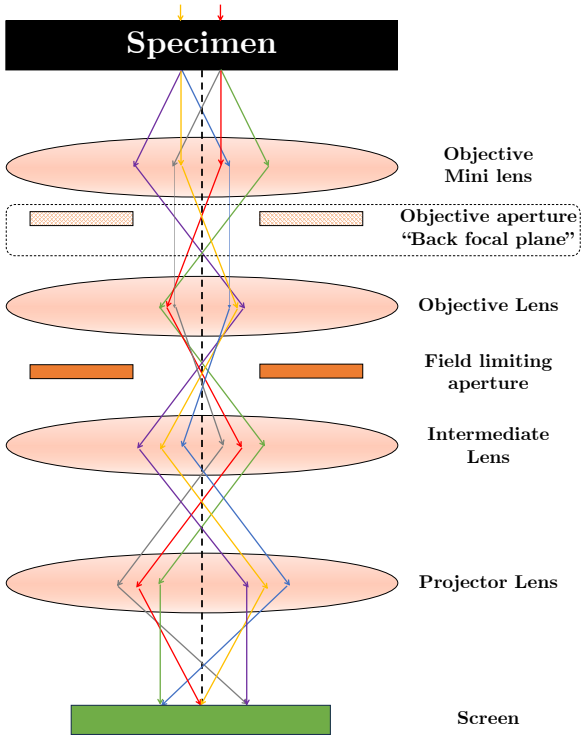

Figure 2.10: **Schematic for the imaging system of a TEM column.** The imaging system is used to project either an image or diffraction pattern to the screen or camera by use of objective, intermediate, and projector lenses. In this example, a diffraction pattern is formed and the objective aperture is assumed removed and there for demonstration only.

is typically condensed to a small spot and then centred using the shift X-Y controls, therefore it is vital that the C2 and twin-lens systems are functioning properly to prevent misalignment. Note also that a condenser aperture can be inserted prior to thee twin-lens system. In typical operation, this aperture is usually set to "open", which is the largest available aperture. This aperture size can be reduced to limit exposure of the beam onto the sample and to reduce the convergence angle of the beam.

In Fig. 2.9, the optic axis is indicated by a black dashed line. The optic axis is defined by the centre of the objective lens and shouldn't change, and the goal of the alignment is to ensure that the beam is parallel, symmetric (*i.e.*, free of condenser and objective lens astigmatism) and centred on this optic axis.

Imaging system

After the specimen, the imaging system is effectively responsible for the projected signal that arises on the camera or phosphorus screen. This can be an image of the sample or a diffraction pattern depending on the lens/aperture configuration. How does this image or diffraction pattern form and how is the imaging system responsible? Firstly, assume that the user has well aligned the column and the beam is centred on the optic axis. The electrons pass through an objective lens and then at the back focal plane a diffraction pattern forms. The electrons can pass through an objective aperture for imaging which can improve resolution by reducing the collection angle of electrons. Typically electrons scattered to higher angles are the most likely to be influenced by lens aberrations, therefore by blocking these signals the resolution can be improved.

The electrons then pass through further objective lenses which form the first "image" of the specimen at the image plane. The objective lenses are responsible for forming and focussing this first image. That is, when defocus is changed, the strength of the objective lens system is changed depending on the desired magnification. The electrons can then pass through field limiting aperture, effectively reducing the width of the image which is projected from the sample. This is typically removed for imaging but inserted for diffraction to gather the signal from a specified region of interest. The electrons enter an intermediate lens system which magnifies the image from the objective lens, and by changing the strength of this lens either the image or diffraction pattern are projected, focussing the diffraction pattern if required. Finally, a series of projector lenses are responsible for magnifying the signal onto the screen or camera.

In essence, imaging system projects either the diffraction pattern from the back focal plane to the screen or camera, or it projects the image from the image plane to the screen or camera. The simplest set up for this is depicted in Fig. 2.10 showing a ray diagram for forming a diffraction pattern.

Other considerations

The output signal quality is highly dependent upon alignment, as is expected. As mentioned, it comes down to the beam being spherically symmetric and centred on the optic axis. Other key hardware include shift and tilt coil sets which allow the beam to be moved without changing the strength of the lenses. For example, a projected image can be shifted on using the projector lens alignment. Furthermore, the incident beam can be shifted using the beam

shift alignment and this is used a lot in TEM alignment to ensure the beam is being aligned along the correct direction. This is important as the beam can always be manipulated in such a way that it appears centred along the axis but if the beam is positioned off axis then the lens strength will not be at the correct values.

Aperture alignment is also important since a misaligned aperture will cause a non-spherical beam to be incident on the sample. In most modern TEMs, the apertures are controlled through a motor, but in the majority of TEMs, aperture control is done by hand. It is possible to drop an aperture into the column if the aperture becomes unscrewed during adjustment, which is more likely to happen if the aperture is far from the optic axis and cannot be seen. If this happens, lowering the magnification can help, or begin by centring a larger aperture.

Aligning the stage to the eucentric height is also an important task. The eucentric height is the plane normal to the optic axis which satisfies a reference focus condition *i.e.*, objective lens strength. A point on the optic will not move laterally if the sample is tilted.

Contrast transfer function in TEM

Aberrations plague electron microscopy due to the challenges associated with designing the lens system. An ideal lens system would ensure that all electrons crossover at the same points through the column, however this ideal circumstance is not possible according to work by O. Scherzer [59]. The path difference between the ideal case (a spherical wavefront) and the actual wavefront defines the aberration of the incident wave. The contrast transfer function defines which scattering vectors $\boldsymbol{u}$ contribute to the final contrast in the image.

The aberration function for TEM $B(\boldsymbol{u})$ is given as,

$$B(\boldsymbol{u}) = \exp\left[j\chi(\boldsymbol{u})\right] \tag{71}$$

where the term $\chi(\boldsymbol{u})$ is given as,

$$\chi(\boldsymbol{u}) = \pi\Delta f\lambda u^2 + \frac{1}{2}\pi C_\mathrm{s}\lambda^3 u^4 \ , \tag{72}$$

where Δf is the defocus value, λ is the wavelength, C_s is the spherical aberration coefficient, and u is the magnitude of the scattering vector. The scattering frequency is then cut off at a

maximum frequency u_a according to an aperture function $A(\boldsymbol{u})$ where $A(\boldsymbol{u}) = 1$ for $|\boldsymbol{u}| < u_a$ and $A(\boldsymbol{u}) = 0$ otherwise. The final component is related to attenuation of the beam due to limited spatial and temporal coherence, known as the envelope function $E(\boldsymbol{u})$. This function essentially forms a virtual aperture at the back focal plane of the objective lens, therefore the objective aperture should be selected no larger that the virtual aperture [75].

The contrast transfer function (CTF) $H(\boldsymbol{u})$ is then given as [76],

$$H(\boldsymbol{u}) = A(\boldsymbol{u})E(\boldsymbol{u})B(\boldsymbol{u}) \ . \tag{73}$$

The envelope function is a product of the spatial, temporal and gaussian envelope functions. The spatial envelope function arises due to the fact that the source is not point-like; it is approximated a series of point-like sources emitting electrons from various initial starting points. This gives rise to an angular spread in the initial emitted electrons and this is quantified by the spatial envelope function.

The temporal envelope function quantifies energy spread in the transmitted beam. This energy spread arises due to an non-monochromatic source with an energy spread of ΔE, an energy spread due to instability in the objective lens current ΔI, and instability in the acceleration voltage ΔV. These quantities add in quadrature to form a defocus spread δ given by [77],

$$\delta = C_c \sqrt{4\left(\frac{\Delta I}{I}\right)^2 + \left(\frac{\Delta V}{V}\right)^2 + \left(\frac{\Delta E}{V}\right)^2} \ . \tag{74}$$

The gaussian envelope function accounts for deflections, instability such as stage drift, and noise induced by changing fields within the column. This envelope function is assumed to be gaussian, and can cause blurring in the final image. Note that all of the envelope functions are gaussian in form, with different parameters characterising their strength with respect to the scattering vector. Fig. 2.11 is a demonstration of the CTF in TEM with each component separated.

The point resolution is given by the first crossover of the CTF, and the larger the scattering vector where this cross-over occurs, the smaller the point resolution. Otto Scherzer determined an optimal defocus value given a spherical aberration coefficient [78]. The basis of this comes from *(i)* finding where the CTF is most flat (*i.e.*, where the gradient is zero) and *(ii)* where the

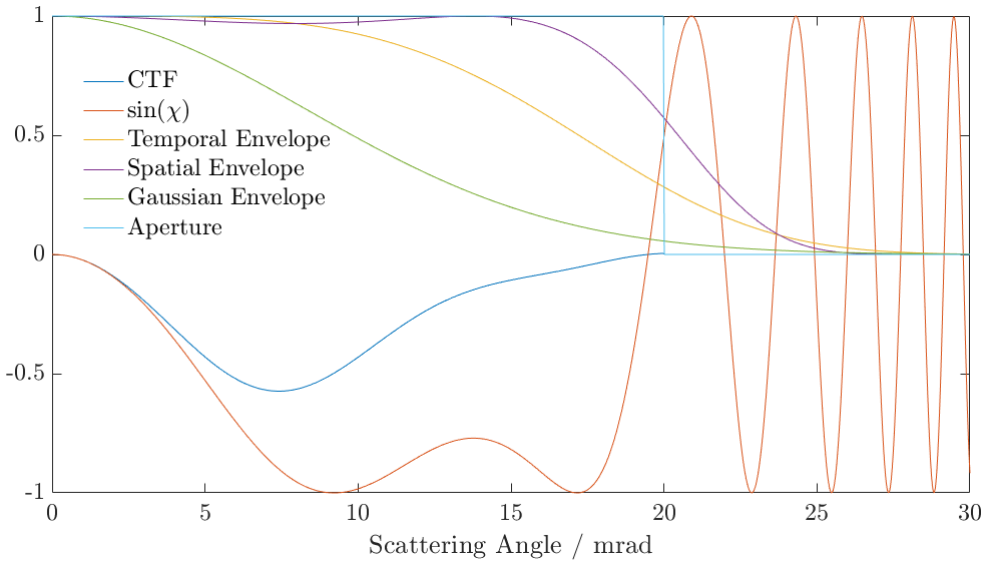

Figure 2.11: **Example CTF for TEM.** The CTF for a TEM beam at Scherzer defocus with an acceleration voltage of 200kV and spherical aberration coefficient of 1mm. Envelopes are set with realistic parameters.

CTF has the same sign and non-zero for the largest range of scattering vectors. The intensity transfer function $T(\boldsymbol{u})$ is given as,

$$T(\boldsymbol{u}) = A(\boldsymbol{u})E(\boldsymbol{u})2\sin\left(\chi(\boldsymbol{u})\right) \ , \tag{75}$$

where the factor 2 arises due to multiplication of the wave-function by its conjugate. The curve is flat when the gradient of $\chi(\boldsymbol{u})$ is zero, *i.e.*,

$$\frac{d\chi(\boldsymbol{u})}{du} = 2\pi\lambda u[\Delta f + C_s\lambda^2 u^2] = 0 \ , \tag{76}$$

implying that

$$\Delta f + C_s\lambda^2 u^2 = 0 \ . \tag{77}$$

Secondly, setting the value of $\chi = -2\pi/3$ since in this region $\sin(\chi)$ will be approximately equal to -1 generates a new function given as,

$$-\frac{2\pi}{3} = \pi\Delta f\lambda u^2 + \frac{1}{2}\pi C_s\lambda^3 u^4 \ . \tag{78}$$

Rearranging equation 77 such that u^2 is isolated and substituting into equation 78 yields,

$$\Delta f_{\text{Sch}} = -\left(\frac{4}{3}C_{\text{s}}\lambda\right)^{1/2} , \tag{79}$$

where Δf_{Sch} is the Scherzer defocus which optimises the defocus of the objective lens given a certain spherical aberration coefficient and wavelength. At this defocus value, the first cross-over is at the largest scattering vector possible and defines the point resolution of the microscope given as,

$$d_{\text{Sch}} = \left(\frac{3}{16}C_{\text{s}}\lambda^3\right)^{1/2} . \tag{80}$$

assuming the sample satisfies the weak phase object approximation. A sample is a potential which interacts with the wave-function of the electron to modify its amplitude and phase. For phase objects, only the phase is modified and not the amplitude, assuming that the sample is thin. Weak phase objects are a special case of phase objects whereby the phase is only slightly modified. For non-phase objects, the amplitude is also modified, examples include objects containing heavy atoms, or thick samples. This is important to consider for appropriate simulation of specimen, as well as for analysing the contrast in S/TEM images. This will be explained in more detail throughout the remaining chapters.

In the context of contrast transfer, the specimen is often defined by an object function $o(\boldsymbol{r})$. Assuming the specimen is a phase object, the object function can be written as,

$$o(\boldsymbol{r}) = \exp\left(-j\sigma V_{\text{t}}(\boldsymbol{r}) - \mu(\boldsymbol{r})\right) , \tag{81}$$

where $V_{\text{t}}(\boldsymbol{r})$ is the projected potential from the specimen, σ is an interaction constant, and $\mu(\boldsymbol{r})$ is an absorption function. The projected potential is approximated as the integral of the 3-D potential through discrete depths, although the most accurate approximation would integrate the potential through infinitesimally small depths. The weak phase object approximation neglects absorption and the projected potential is assumed to be small, leading to the weak phase object approximation,

$$o(\boldsymbol{r}) = 1 - j\sigma V_{\text{t}}(\boldsymbol{r}) , \tag{82}$$

which holds if the sample is very thin, such that the amplitude is unitary and only the phase is slightly modified, with the above following from [79]. The resulting wave-function in real

space is a convolution of the transfer function in real space $h(\boldsymbol{r})$ with the object function. The resulting wave-function in reciprocal space is the product of the aberration function given in equation 73 with the object function in reciprocal space (*i.e.*, the Fourier transform of $o(\boldsymbol{r})$). Depending on the set-up, the resulting image/diffraction pattern is the modulus square of the wave-function in real/reciprocal space respectively- this is what the camera or screen measures.

The CTF can be manipulated to form *passbands* which form flat regions within the CTF at higher spatial frequencies which can contribute to the image [80]. Passbands are achieved by setting the defocus value according to,

$$\Delta f_{\mathrm{p}}^{n} = -\left[\frac{8n+3}{2}(C_{\mathrm{s}}\lambda)\right]^{1/2} \ . \tag{83}$$

A final remark on TEM is on resolution. Up to this point, the only consideration is on the maximum spatial frequency limited by the CTF. In the absense of aberrations, the Rayleigh criterion defines the diffraction limited resolution as [81],

$$r_{\mathrm{th}} = 1.22\frac{\lambda}{\beta} \ , \tag{84}$$

where r_{th} is the theoretical resolution and β is the collection semi-angle [82]. In practice, the actual resolution is limited by the aberrations which add in quadrature given by,

$$r = \sqrt{(r_{\mathrm{th}}^2 + r_{\mathrm{s}}^2 + r_{\mathrm{c}}^2 + r_{\mathrm{com}}^2 + r_{\mathrm{ast}}^2)} \ , \tag{85}$$

where r_{s}, r_{c}, r_{com}, r_{ast} are the resolution limits associated with spherical, chromatic, comatic and astigmatic aberrations respectively. In practice, the latter two aberrations are minimised during alignment using bright tilt and stigmators, and the chromatic aberration as discussed is corrected through monochromators. The spherical aberration can be corrected for thanks to spherical aberration correctors, but typical TEMs aren't generally equipped with a C_{s} corrector. As such, it is typical to reduce equation 85 down to just the theoretical and spherical resolution limits such that,

$$r \approx \sqrt{(r_{\mathrm{th}}^2 + r_{\mathrm{s}}^2)} \ . \tag{86}$$

An approximate function can be derived in terms of the collection semi angle,

$$r(\beta) = \left[\left(\frac{\lambda}{\beta}\right)^2 + (C_s\beta^3)^2\right]^{1/2} , \tag{87}$$

where the function should be minimised by taking the derivative and setting to zero, which results in an optimal collection semi-angle β_{opt},

$$\beta_{opt} = 0.77\left(\frac{\lambda}{C_s}\right)^{1/4} , \tag{88}$$

and the minimum value of equation 87 is then,

$$r_{min} \approx 0.91(C_s\lambda^3)^{1/4} . \tag{89}$$

It's worth nothing that this number is significantly higher than the wavelength of the electron and the Rayleigh criterion. It's easy to imagine the excitement in the 1930s as physicists and engineers postulated the potential resolution of electron microscopes. So far, this chapter has been a story about how fundamental physics can be used to generate a machine capable of remarkable tasks. The TEM is a statement of science and engineering, combining the elegance of quantum mechanics with the perseverance of engineering.

2.3.2 STEM

The scanning transmission electron microscope (STEM) is regarded as the state-of-the-art instrument for atomic scale imaging, with the instrument being used to acquire the image which holds the record for highest resolution [83]. The STEM is an adaptation of the TEM, where the incident beam is now converged to form a focussed probe with a diameter typically smaller than 1Å for the majority of spherical aberration corrected instruments. For this reason, atoms can be individually excited if the probe is situated upon it, and therefore a direct atom measurement can be recorded.

In this section, the history of the STEM shall be presented as well as the design of the column. The probe forming system shall be extended into the mathematics which describes the aberrations present within a STEM probe, then finally a discussion on the signal acquisition modalities which are common within STEM.

History of STEM

The history of the STEM is well documented in literature, such as in the well-known textbook *Scanning Transmission Electron Microscopy* by S. J. Pennycook and P. D. Nellist [84]. In 1937-38, Baron Manfred von Ardenne successfully designed and built the first STEM, demonstrating a resolution of 40nm [85], and soon after a resolution of 10nm [86]. von Ardenne was operating his STEM with an accelerating voltage of 60kV, however he wanted to increase this to *extra-high-voltage i.e.*, up to 300kV, and then 1MeV. The machine was successfully capable of almost 300kV, however it was destroyed during a bombing raid in 1944 [87]. The machine was designed to be mounted to a 1MeV discharge tube [88], and it is rather upsetting to see that the price of war prevented a different chain of events which could have accelerated STEM development to modern day.

It took until 1966 and the groundbreaking work by Albert Crewe at Argonne National Lab before the STEM was further developed, this time including a field-emission electron source to increase brightness and reduce the energy spread of the probe forming electrons [89, 90]. Using this design with further advancements [91], Crewe was able to demonstrate resolution below 5Åusing the STEM through the imaging of individual thorium and uranium atoms [92, 93]. This demonstrated the potential application for the STEM to image atomic structure, as well as the potential for atomic scale spectroscopy through EELS and EDS.

STEM-EELS was demonstrated initially for nucleic acid bases by Crewe *et al.* [94], with Isaacson later using STEM-EELS to calculate the minimal dose requirements for STEM imaging of biological specimens [95]. By inference, it would appear that these specimen were chosen since the STEM was able to resolve these nanoscale structure, but it wasn't until a few years later in 1974 when Wall *et al.*demonstrated the use of elastic dark field imaging to image thorium crystallites at approximately 3Å resolution [96], as well silver atoms using a 43kV accelerating voltage. This paper concludes with a fitting statement that although the resolution of the STEM and conventional TEM (CTEM) were equivalent, the STEM was a multi-dimensional imaging tool, being able to collect multiple signals from different scattering events. Wall concludes that this is important for improving dose efficiency, and in turn minimising the beam induced damage.

The first commercially available STEM was the Vacuum Generators (VG) HB5, which was

first installed in March 1974 at Queen Elizabeth College with a 100kV electron source. By 1976, the HB5 had been installed at MIT, with Siemens also now manufacturing their Elmiskop ST100F STEM [97]. Other manufacturers began developing their STEM instruments in the 90s and early 2000s, and with the advent of probe correctors developed by Krivanek *et al.* around this time, 2Å resolution was demonstrated [6, 98] on a modified VG HB5. Haider *et al.* similarly introduced an aberration corrector for TEM, showing a resolution of 1.4Å a year later [99].

High angle annular dark field (HAADF) imaging was now a standard imaging technique used for high resolution imaging of hard materials such as the distribution of dopants within silicon [100] and platinum nanoparticles. HAADF imaging was powerful since the image contrast was directly interpretable as it was based on the square of the atomic number. This scattering is approximately Rutherford scattering to high angles (between 2 and 5 times the convergence semi-angle), as described in section 2.2.4. It was the work of S. J. Pennycook [101] and others that made HAADF STEM imaging the go-to method for imaging high-Z number elements at atomic resolution with works throughout the 90s [102–112].

STEM design

As discussed, the STEM fundamentally differs from the TEM by way of a convergent electron probe which rasters over the sample, with the scattering collected at each probe location being used to form the image. In order to achieve a convergent probe, a series of lenses must be aligned and adjusted to correct for low-order aberrations such as defocus. Higher order aberrations such as spherical aberration are corrected for using a probe corrector.

Since the shape of the electron probe ultimately determines the quality of the final image, this section will focus on the design of the probe forming system, then extending into the principle of reciprocity.

Probe forming system

The key to STEM is the probe. For high resolution imaging say using HAADF scattering, the probe should be as small as possible with a circular symmetry. There are techniques for correcting residual aberrations in post processing such as ptychography, however here the prior is considered.

The electrons are emitted from a source such as a cold FEG with some energy spread ΔE

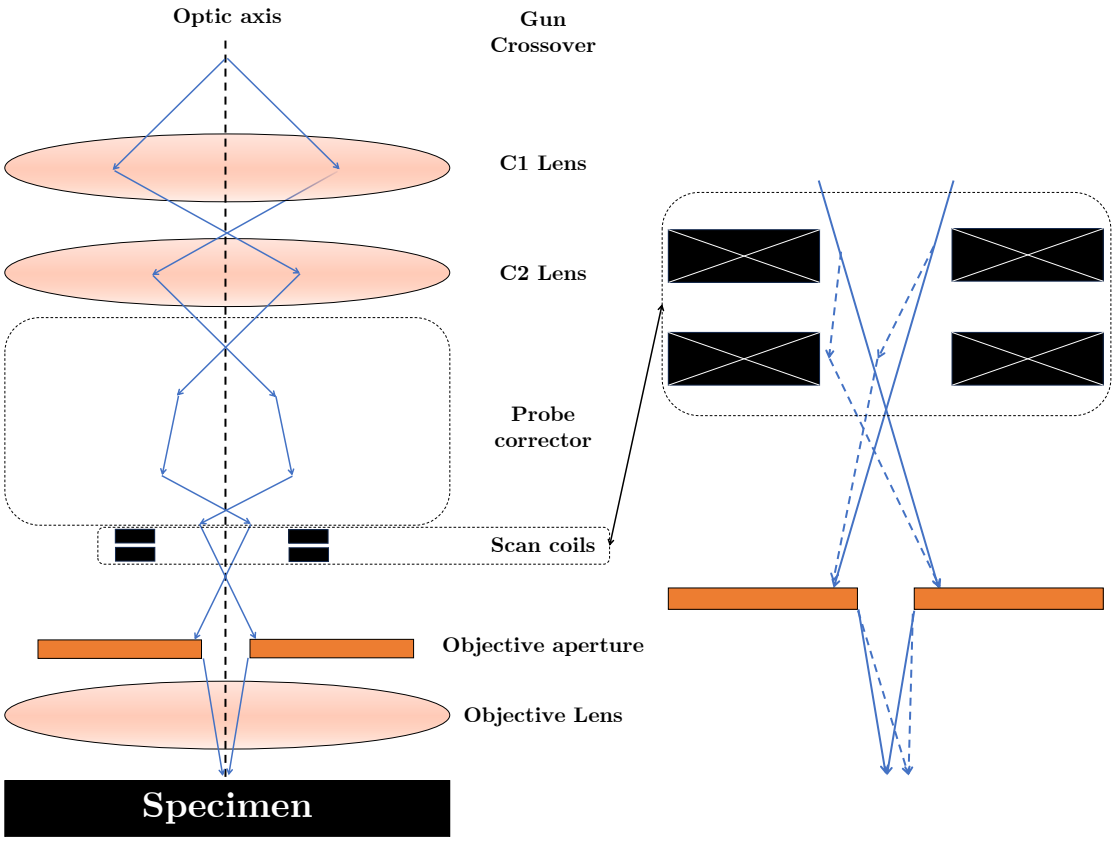

Figure 2.12: **A schematic for the probe forming system in a STEM.** The probe forming system consists of series of lenses and a probe corrector to correct residual aberrations. There is also a scan coil system in order to raster the probe over the sample.

and then collected by a first condenser lens (C1 lens). There is then a crossover before the second condenser lens (C2 lens) collects the electrons. The electrons then enter the probe correction system (if equipped), followed by the scan coil system, an objective lens, then finally striking the sample as shown in Fig. 2.12.

As in TEM, the probe should be aligned along the optic axis, only deviating according to the scan position. In order to align a STEM probe, the Ronchigram is used as a diagnostic tool. The Ronchigram (named after Enrico Ronchi [113]) is a projected of the sample due to the probe onto the Fruanhofer diffraction plane [114]. The Ronchigram contains a mixture of real and reciprocal space information, making it an *interesting imaging mode* in the words of Andrew Lupini [115]. The Ronchigram changes significantly for misalignments in the probe, coming away from circular/n-fold star symmetry depending on whether a n-pole (typically hexapole *i.e.*, n = 6) corrector is equipped. This allows for a manner of useful applications, not least alignment, but also sample tilt for example.

In order to align a STEM probe, the objective aperture is typically removed (or set to its largest value depending on the instrument) and the probe is set to stationary at a thin amorphous part of the specimen. Following this, eucentric height can be achieved by changing the position of the stage such that the Ronchigram is at Gaussian focus. From here, the Ronchigram can be under-focussed or over-focussed and symmetries observed. At large defocus values, an image of the specimen is projected onto the viewing screen since the beam is broad at the specimen. For simplicity, assume a spherical aberration corrector is present but the comatic and stigmatic aberrations require alignment.

In order to correct from here, the bright tilt (condenser alignment coils) and condenser stigmators, and defocus controls are iteratively adjusted to get onto the coma-free axis, astigmatism-free axis, and at Gaussian focus. To get to the coma-free axis, the Ronchigram should have the 6-fold symmetry caused by the corrector with the Ronchigram centre being stationary as the defocus is varied. If the Ronchigram contains striations in a particular direction, this is indicative of stigmatic aberrations, and they can be adjusted until the Ronchigram is homogenous. Note that the defocus must also be adjusted as typically changing the astigmatism will affect the defocus of the probe. These three steps are repeated until a flat, smooth, and homogenous central region of the Ronchigram is seen on the viewing screen. An objective aperture is then inserted (typically 20-30 mrad) and that region contained in the shadow from the aperture should be completely homogenous. The Ronchigram will be discussed more later on in this chapter when discussing the contrast transfer function for STEM 2.3.2.

Following the specimen, a series of detectors can be used to collect certain scattered signals or above the sample to collect X-rays or possibly secondary electrons. Common detectors are the radial HAADF and circular BF monolithic detectors which measure the induced electric current from the incident electron flux and assign this as intensity. In addition, a camera can be inserted to observe the Ronchigram. These cameras are typically charge-coupled devices (CCD) due to their relative inexpensive cost compared to direct electron detectors (DED) for viewing, although depending on the experiment, a DED should be used to increase sensitivity and frame rates. Acquisition modes are discussed in section 2.3.3.

Contrast transfer function in Z-contrast STEM

A Ronchigram or in-line hologram is formed through the convolution of object transfer function $O(\boldsymbol{k})$ and probe function $P(\boldsymbol{k})$ at a given probe location. For a given probe location, the intensity measured on the Ronchigram at a given scattering vector is given as,

$$\mathcal{I}(\boldsymbol{r}_p, \boldsymbol{k}_d) = |P(\boldsymbol{r}_p, \boldsymbol{k}) \otimes O(\boldsymbol{k})|^2 \ . \tag{90}$$

The probe function is given as,

$$P(\boldsymbol{r}_p, \boldsymbol{k}) = A(\boldsymbol{k}) \exp\left[-j(\chi(\boldsymbol{k}) - 2\pi \boldsymbol{r}_p \cdot \boldsymbol{k})\right] \ , \tag{91}$$

where $A(\boldsymbol{k})$ is an aperture function as in section 2.3.1, and $\chi(\boldsymbol{k})$ is the aberration function given by,

$$\chi(\boldsymbol{k}, \phi) = \frac{2\pi}{\lambda} \sum_{n,m} \frac{1}{n+1} C_{n,m} (k\lambda)^{n+1} \cos\left[m(\phi - \phi_{n,m})\right] \ , \tag{92}$$

where ϕ is the azimuthal angle, and the indices m, n of the aberration follow the notation of Krivanek [6]. This is the same definition which was derived in section 2.3.1 for HRTEM imaging. For Z-contrast STEM where high scattering angles are considered, the dominant signal arising comes from the electron-phonon interactions *i.e.*, thermal diffuse scattering and nuclear scattering. This signal is incoherent and the annular detectors are assumed to collect all the scattering at these high angles, which is a critical assumption to derive the OTF and image intensity.

The image intensity for Z-contrast is given as the the following [116],

$$z(\boldsymbol{r}_\mathrm{p}) = |p(\boldsymbol{r}_\mathrm{p})|^2 \otimes o(\boldsymbol{r} - \boldsymbol{r}_\mathbf{p}) \ , \tag{93}$$

where $P(\boldsymbol{r}_\mathrm{p})$ is the complex probe amplitude and $o(\boldsymbol{r})$ is the object function in real space. This is the definition of incoherent imaging [2]. The image is then a convolution of the object function with a real-positive intensity point spread function, and the Fourier transform of this arrives at the OTF for STEM. Importantly, the OTF for Z-contrast has no contrast reversal, decaying

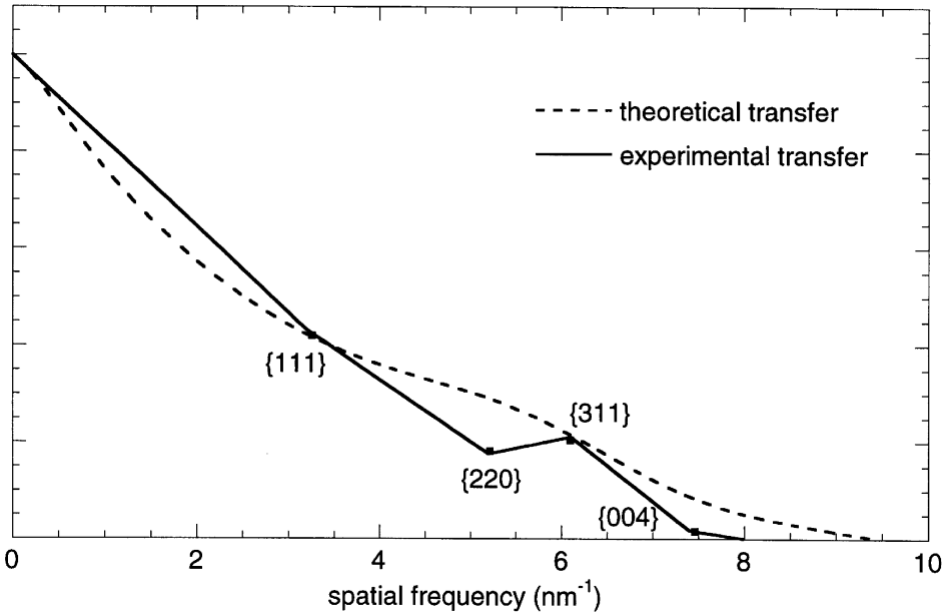

Figure 2.13: **Experimentally derived and theoretical OTF for Z-contrast STEM.** The estimated OTF for a JEOL JEM 2010F taken with permission from [2]. The OTF is estimated from the power-spectrum of experimentally acquired silicon dumbbells and interpolated between reflections.

to zero as the spatial frequency increases. This is demonstrated in Fig. 2.13.

Principle of reciprocity

The principle of reciprocity describes the relationship between a CTEM and a STEM. Consider Fig. 2.14 where a simple schematic for the CTEM and STEM are drawn side-by-side. The principle of reciprocity essentially states that if the signal is considered elastic, forward electron paths in a CTEM (*i.e.*, starting at the source and finishing at a screen) are equivalent to reverse paths in a STEM (*i.e.*, staring at the detector and ending at the source) due to the symmetry in the optical system [3, 117]. The optics of the CTEM system after the sample are equivalent to the optics of the STEM prior to the sample. In STEM alignment, the quality of the final signal is determined by the condition of the probe, which is of course above the sample. On the other hand, the quality of a CTEM image is determined by the alignment of the signal *after* the sample.

2.3.3 Contemporary data acquisition modes in STEM

Just as the TEM can acquire various signals, be them global due to a parallel beam, the STEM can also acquire multiple different signals with spatial and temporal resolution. As has been

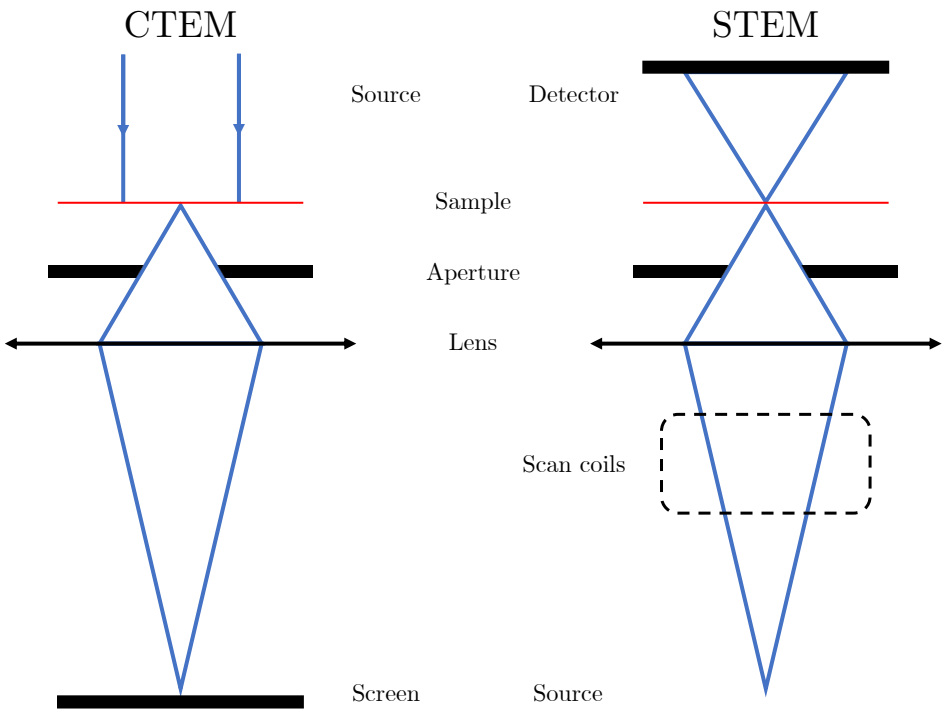

Figure 2.14: **Diagram showing the principle of reciprocity for CTEM and STEM.** CTEM (left) and STEM (right) schematics showing the reciprocal nature of CTEM and STEM, where the source and screen/detector are inverted. Figure replicated from [3], Fig. 1.

seen in the earlier sections, certain signals are better than others for characterising certain materials. For example, if a user wanted to image a weak phase object, such as a biological specimen, collecting the high angle elastic signal would be inefficient as the elements within that sample would be of low Z-number, and the scattering cross section would be low. Therefore other data acquisition modalities have been formed which aim to overcome these issues by optimising the flux efficiency of the acquisition; maximise the useful signal out for every electron that goes in.

In this section, contemporary STEM data acquisition modes are discussed, highlighting where each is useful and why each should be carefully considered when characterising certain samples.

Bright field and dark field STEM imaging

The typical imaging modes within STEM are built on radial detectors. Effectively, these detectors are designed to collect scattered electrons in some angular range for a given probe location, and then this number is assigned to that probe location. The intensity $z_{\boldsymbol{r}_\mathrm{p}}$ measured given a Ronchigram $\mathcal{I}(\boldsymbol{r}_p, \boldsymbol{k})$ at probe location $\boldsymbol{r}_\mathrm{p}$ is given as,

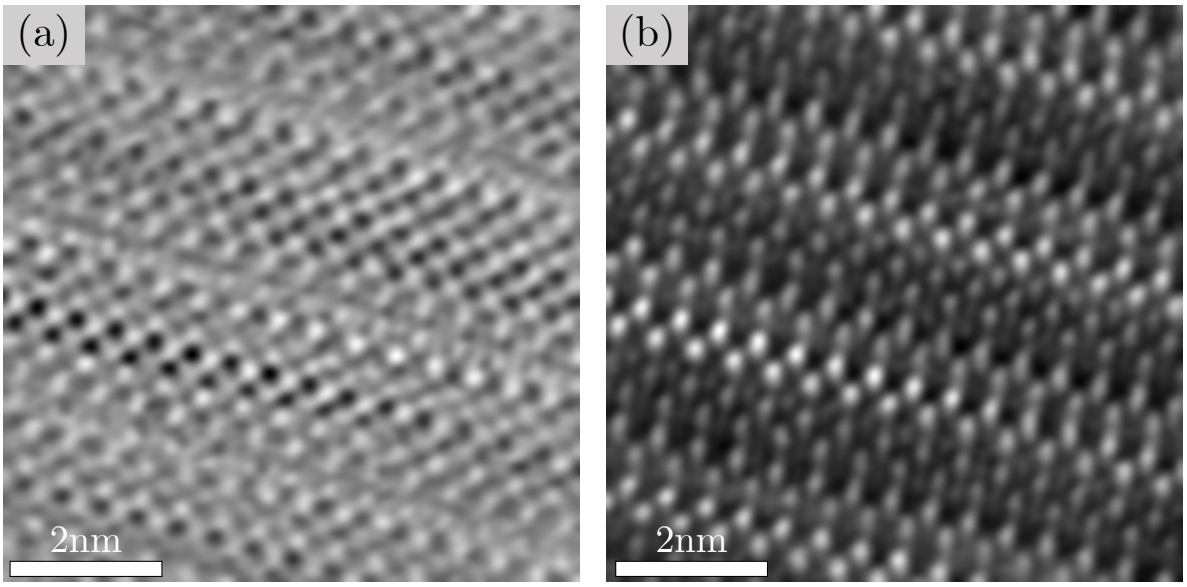

Figure 2.15: **Examples of atomic resolution bright field and annular dark field STEM images.** (a) Bright field image and (b) dark field image of a layered bismuth structure. The bright field image shows phase contrast, whereas the dark field image is correlated to the Z-number of the elements present within the sample as well as the thickness.

$$z_{\boldsymbol{r}_\mathrm{p}} = \int_{\boldsymbol{k}_\mathrm{i}}^{\boldsymbol{k}_\mathrm{o}} \mathcal{I}(\boldsymbol{r}_p, \boldsymbol{k}) d^2\boldsymbol{k} \ , \tag{94}$$

where the values $\boldsymbol{k}_\mathrm{i}$ and $\boldsymbol{k}_\mathrm{o}$ denote the inner and outer reciprocal space vectors of the detector. The reciprocal space vector is related to the scattering angle through $k = \theta/\lambda$.

Spectroscopy

Typical spectroscopic methods within STEM are EELS and EDS which can reveal quantitative chemical information about the sample. Both methods are the result of inelastic scattering processes, as discussed in section 2.2.4. Given the scattering cross-section of each of these methods, the signal-to-noise is generally low unless a higher beam current is used or a longer dwell time to increase electron fluence at the sample. For these reasons, EELS and EDS spectral images (or maps) are generally rather noisy and/or take a long time to acquire.

An EELS map is generated by raster scanning the probe over a region of interest. At each probe located, the transmitted electrons enter an electromagnetic prism and are deflected according to their energy which is recorded as a 2-dimensional distribution on a detector. This distribution is then integrated perpendicular to the axis corresponding to the energy loss and the data stored. This results in a 3-dimensional dataset $\mathcal{X} \in \mathbb{R}^{H_\mathrm{P} \times W_\mathrm{P} \times N_\mathrm{E}}$ where N_E is the num-

ber of energy channels which are recorded, and $H_\mathrm{p}, W_\mathrm{p}$ are the scan grid height and width respectively.

Similarly, an EDS spectral image is produced by rastering the electron probe over the region of interest and then the detection of x-rays with energy hc/λ onto a scintillator for that given probe location. This results in a 3-dimensional dataset $\mathcal{X} \in \mathbb{R}^{H_\mathrm{p} \times W_\mathrm{p} \times N_\alpha}$ where N_α is the number of x-ray energies which are recorded.

Four-dimensional STEM

Four-dimensional STEM (4-D STEM) has become a popular tool in STEM by virtue of its multimodal imaging, *i.e.*, various analyses can be performed from a single dataset. A 4-D STEM data is acquired through collecting a convergent beam electron diffraction (CBED) pattern at the far-field for each probe location in a raster scan. The 4-D STEM data is then collected to form a 4-D data array given as $\mathcal{X} \in \mathbb{R}^{H_\mathrm{p} \times W_\mathrm{p} \times H_\mathrm{d} \times W_\mathrm{d}}$ where $H_\mathrm{d}, W_\mathrm{d}$ are the height and width of the detector collecting the CBEDs, respectively.

The 4-D STEM acquisition can be broken into two distinct forms; focused probe and defocused probe 4-D STEM. Focussed probe 4-D STEM is useful for characterising electrical properties such as the projected electric field and projected charge density of the sample. It can also be used to form *virtual detector images* where the CBED is integrated over a custom angular range to mimic a fixed detector. Another analysis method is focused probe electron ptychography, where the probe and object functions can be deconvolved through a closed form or iterative solver.

Defocused probe 4-D STEM is most powerful for pytchographic phase image recovery using iterative solvers. By defocusing the probe, a larger region of the sample is exposed to the probe. By taking advantage of the overlap between neighbouring probe locations, the object and probe can be iteratively updated to minimise some cost function, deconvolving the two in the process. 4-D STEM and ptychography are discussed in more detail throughout sections 6 and 7.

2.4 Limitations of electron microscopy

Despite all the benefits of electron microscopy, there are inherent drawbacks which must be considered and accounted for. In this work, the main focus is on STEM and the possible improvements that can be made, but prior to that the limitations must be understood to provide a motivation. In this section, the drawbacks which underpin the purpose of this research are explored in more detail.

2.4.1 Beam damage mechanisms

As previously discussed, because of the developments made in STEM over recent decades, probes have become smaller, brighter, and more coherent. Despite this being a benefit for samples which can remain stable under this illumination, there are a host of samples which cannot. This sample instability/degradation is commonly known as beam damage (or simply damage) [118, 119]. Beam damage can be considered as electron-specimen interactions that change the structural properties of the sample being looked at. This is something that prevents an accurate representation of the sample, which makes any analysis derived near redundant if not especially accounted for. Whenever a STEM is used, careful consideration for beam damage potential must be given, and an understanding of the possible damage mechanisms is crucial. We will now categorise the common mechanisms which give rise to damage.

Knock-on damage

As discussed in section 2.2.4, electrons can interact with a sample which give rise to both elastic and inelastic collisions. Consider the case where the electron passes very close to the nucleus of an atom with a small impact factor, such that the scattering angle is high. In this case of elastic scattering, sufficient energy can be transferred from the electron to the nucleus such that the atom becomes displaced from its equilibrium position. The amount of energy transferred from the electron to the nucleus is given by [120],

$$E = E_{\max} \sin^2(\theta/2) \ , \tag{95}$$

where θ is the scattering angle, and $E_{\max}$ is the maximum energy that can be transferred from the electron to the nucleus corresponding to $\theta = \pi$ rad. The value of $E_{\max}$ is given without

proof as [120],

$$E_{\max} \approx 2E_0 \frac{E_0 + 2m_e c^2}{Mc^2} \ , \tag{96}$$

where M is the nuclear mass of the target nucleus and other terms have their usual meaning described throughout this work. Equation 96 implies that lighter elements are more susceptible to higher energy transfer from the incident electron.

If the atom which has been displaced continues on a particular trajectory with sufficient momentum, it can cause further displacements of other atoms within the sample- this is known as cascading. Cascading can lead to the formation of defects, vacancies, or holes within the sample, making analysis of pristine samples difficult- if not impossible.

For knock-on damage to occur, the accelerating voltage needs to be sufficiently high such that the electron has enough momentum to displace an atom, *i.e.*, beyond the knock-on damage threshold [121]. Therefore, if a sample is susceptible to knock-on damage, then one can reduce the accelerating voltage to mitigate. This is commonly done for the analysis of so-called 2-dimensional materials [122].

Radiolysis

Radiolysis arises from inelastic scattering of electrons and refers to the cleavage of chemical bonds in the sample due to the interaction. When electrons penetrate the sample, they can ionize molecules and break chemical bonds, leading to the formation of radicals and the release of gas species. This process can alter the sample's chemical composition, introduce artefacts, and possibly induce structural changes.

Radiolysis becomes more pronounced in materials containing organic, insulating or semiconducting materials. As a thought experiment, consider an electron incident upon an arbitrary material. Next, assume that this electron interacts with an atomic electron such that it is displaced from its initial state and a hole is created. For conducting materials, this hole is quickly filled by an electron within the material, and as such there is not sufficient time for the atom to dissociate. However, now suppose that the material is insulating. If an electron is displaced from either the valence band or inner atomic shell and a hole is created, it is likely that there would be sufficient time for the atom to dissociate, leading to the cleavage of chemi-

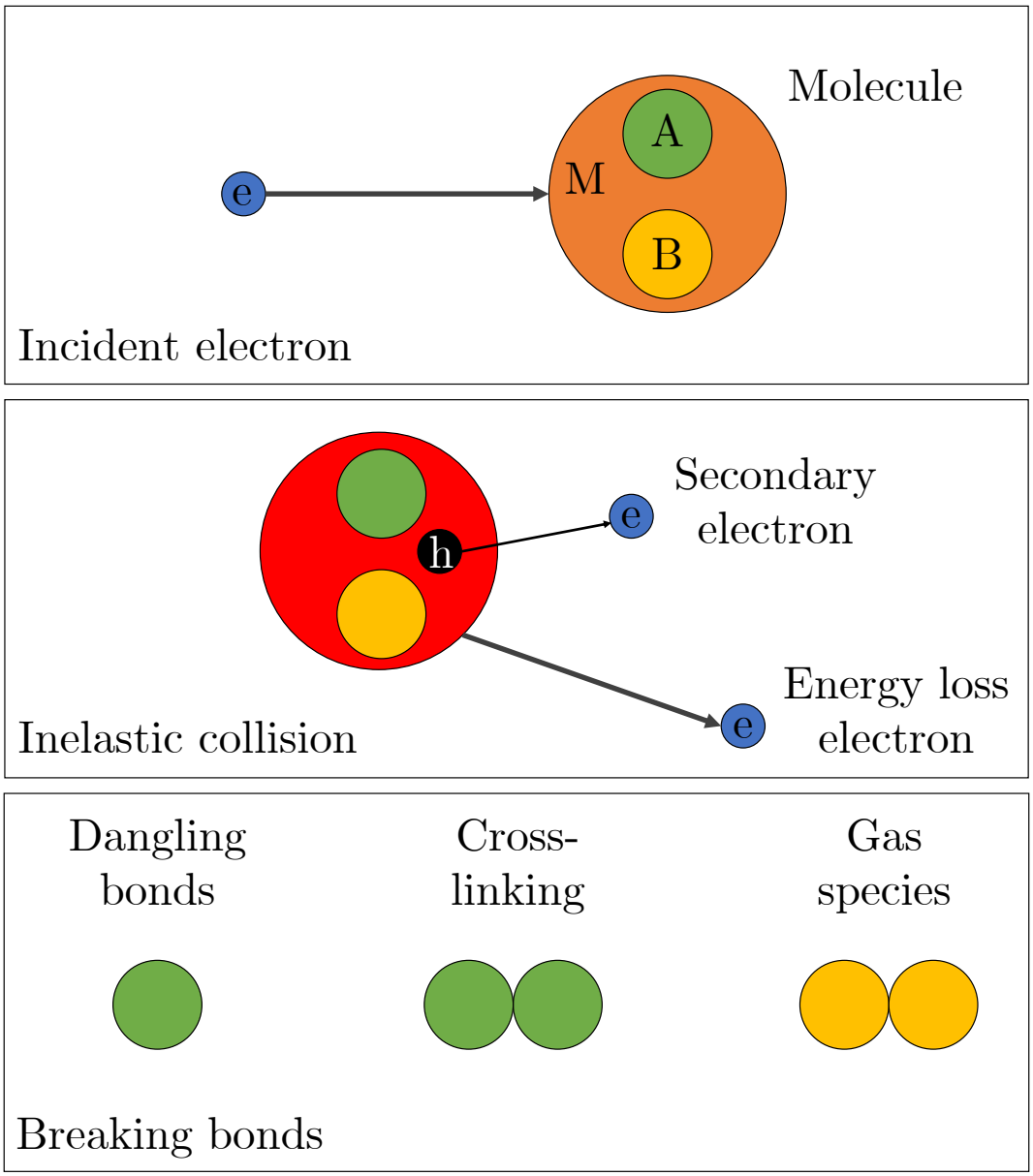

Figure 2.16: **Workflow of the radiolysis process.** A high energy electron collides with a molecule composed of light or organic material. The inelastic collision causing a secondary electron to be ejected, leaving behind a hole. This then causes the bond to break, leading to dangling bonds, cross-linking, and potentially the formation of gaseous species.

cal bonds. For semi-conducting materials, this process is dependent upon the band-gap of the material.

Following on from work by R. F. Egerton [120], for organic materials such as biological specimen, radiolysis can induce the formation of various chemical species and cross-linking. As shown in Fig. 2.16, the incident electron causes a secondary electron to be ejected leaving behind a hole. For hydrocarbons, the mobility of hydrogen may cause hydrogen diffusion, ultimately preventing the bond from reforming. As such, this can leave dangling bonds, and

therefore cross-linking. Ultimately, this results in a change of chemical structure from the pristine state leading to irreversible beam damage.

Sample heating

Inelastic electron-electron scattering between the beam and the sample can cause local temperature within the sample to increase through the energy transfer. This can lead to damage within the sample and can induce phase transformations, sintering, and evaporation of volatile species [118]. For the majority of samples, heating is not a big issue. On the other hand, polymers can degrade substantially since they have poor thermal conductivity [120]. To mitigate heating effects, lower beam currents and shorter exposure times are often used. Heating is also related to the probe diameter, but this does not scale linearly and the increase temperature change is typically negligible, even with the same beam current.

Another solution to reduce heating is to use cryogenic stages. As has been discussed, cryostages can be used to decrease the rate of radiolysis, due to the decreased mobility of radicals (*i.e.*, short range order destruction [123]). Although, it may also be that the temperature gradient is more important than the absolute temperature for a given sample damage due to heating.

Charging

When the incident electron beam interacts with an electrically insulating material, it can cause the accumulation of charges on the sample surface. This charging effect arises due to the imbalance between the rate of electron injection and the rate of charge dissipation from the sample [118, 124]. The accumulated charges can deflect the incident electron beam, distort imaging, and potentially induce sample damage [118]. Charging is typically seen during SEM image acquisition of insulators such as biological specimen [125] and polymers [126]. Kim *et al.* [127] suggest an osmium coating on the specimen to allow surface charges to dissipate.

2.4.2 Contamination

Carbon contamination, or simply contamination, is one of the most infuriating aspects of STEM as it can obscure the intended region of interest, cause charging, and inhibit the resolution [128]. It fundamentally results from excessive hydrocarbons present within the STEM

column and contaminants on the sample surface, and as a result, a thick hydrocarbon layer can build up onto the surface of the sample [129–132].

There are several reasons why there might be excessive hydrocarbons. Firstly, some samples are more susceptible to hydrocarbon adsorption, and as a result the sample itself is inserted into the column whilst being coated in a thin layer of hydrocarbons. Ultimately, the user is expected to be aware of how likely this is to occur with their sample so that exposure can be minimized. This includes preparing samples in a glove box, transporting the sample in a vacuum or neutral gas containing box or holder, or using specialist sample grids.

Secondly, another reason why the column may contain excessive hydrocarbons is due to poorly maintained sample holders, so ensuring that the sample holder is cleaned regularly is paramount. This can be done through a plasma cleaner, and making sure that the user is wearing nitrile gloves. Another useful technique is to use silicon wipes to clean the o-rings and the rod.

Thirdly, it is important to keep the column at as close to vacuum as possible. If the vacuum level is poor, then the column will contain residual gas molecules. These gas molecules interact with the electron beam, and as such they are ionized. When the electron beam interacts with the sample, due to its velocity through the material, a slightly positive electric field is generated at the surface, since the mobility of the electrons is greater than that of the ions. This causes more electrons to be incident around that region so that when the electron probe moves, the ions are then deposited at the surface where the probe had just visited, as shown in Fig. 2.17

Work by Hugenschmidt *et al.* [129] shows that the contamination increases with beam current, but saturates at high beam currents. The reason for this is due to the strength of the electric field induced (see Fig. 2.17) by the incident electron beam. The higher the current, the stronger that electric field will be according to,

$$-\nabla^2\phi = \frac{\rho}{\epsilon_0} \tag{97}$$

so the migrating electrons within the specimen also drive migrating contaminants on the surface of the specimen. However, this understanding does not account for the role of the beam in dispersing already migrated and deposited contamination during exposure. The beam itself

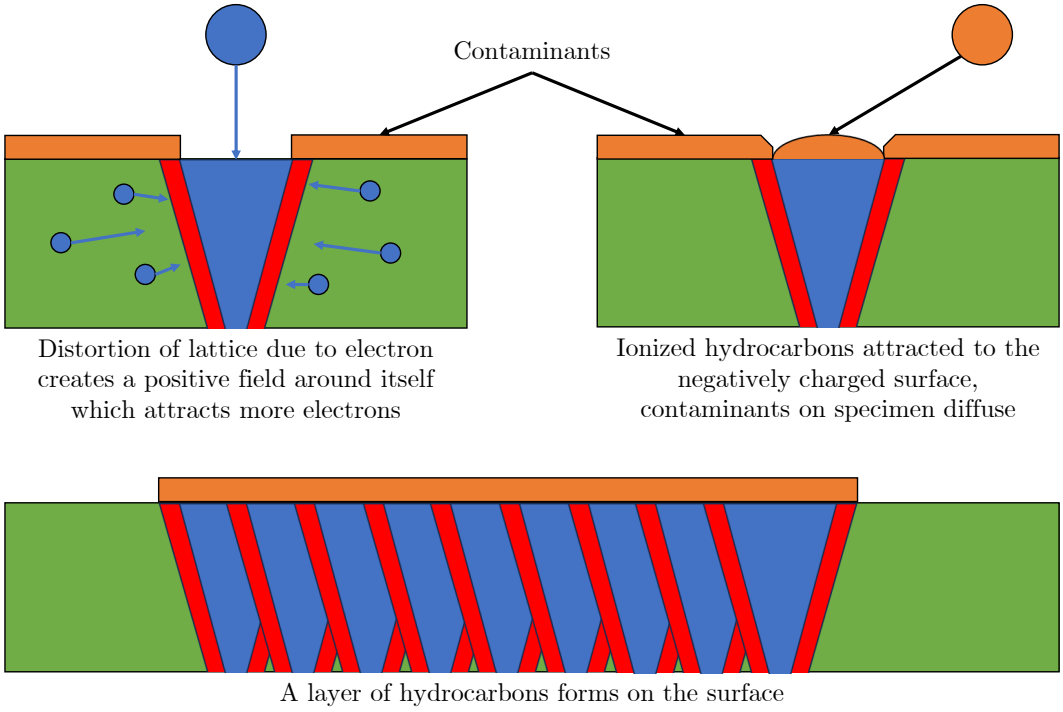

Figure 2.17: **Contamination formation mechanism in STEM.** The incident electron beam induces polymerization of surface contaminants and the adsorption of ionized residual gaseous hydrocarbons within the column. A layer of contamination can also form on the bottom of the sample if the sample is sufficiently thin.

can also reduce contamination if the beam current is sufficiently high, and in observation, it is often under low beam currents that contamination is most resistant to disperse. Typically, the contamination will take longer to form at lower beam currents, until it reaches saturation. By this point, the beam is unable to effectively redistribute the contamination. At high beam currents, the rate of contamination growth versus dispersion of contamination is in favour of the latter, as argued by various works [131, 133, 134].

If contamination forms during an experiment, then there are several proposed solutions. The first is to ensure than the liquid nitrogen dewar is filled so that the cold finger is able to condense the hydrocarbons before they migrate to the specimen. A second solution is to use what is known as a *beam shower*. A beam shower acts to redistribute the contaminating hydrocarbon layer which has formed on the surface, or to pin down the existing surface contaminants which is effectively forming a thin layer of contamination. This is commonly used in most STEM experiments, and has been cited as one of the most effective techniques [129]. A third solution is to create a contamination barrier, effective against surface contaminant migration. By etching a boundary of contamination surrounding a region of interest, the rate

of contaminant migration can be reduced sufficiently that the region can be investigated at higher magnification with reduced contamination formation. The latter was demonstrated in the Ph.D thesis of Yoshie Murooka [135].

Although contamination is a drawback of STEM, it can be managed by strict maintenance of the microscope or by applying the discussed mitigations during experiment. It is postulated that probe sub-sampling could help with reducing contamination build up, however this has not been demonstrated here nor in other works so far.

2.4.3 Scanning systems

Electron probe scanning systems have not changed dramatically since their initial design. The scanning system is composed of sets of scanning coils which aim to shift and tilt the probe to the desired probe location, followed by a third condenser lens to condense the probe. The scan coils are designed to ensure that the probe is parallel to the optical axis during the scan [136].

Scan coils are inherently unreliable due to hysteresis. For STEM, hysteresis typically refers to the mismatch of the beam position, as determined by the electromagnetic scan coils, from its target position, as determined by the scanning electronics. This mismatch is largely attributed to the inductance of the scan coils; as the current in the coils changes, which determines the field strength, and thus the probe location, this change in current is resisted. The result is that, rather than changing instantaneously as desired, the electromagnetic field moves smoothly between one state and another. When scanning quickly, as is often done in STEM, this change in current is slower than the scan speed, which results in observed hysteresis. For raster scanning, where the scan pattern is fixed and predictable, correcting this after-the-fact is trivial and often taken for granted. Moving away from raster scanning, however, poses significant restrictions on your choices due to hysteresis. As each electron microscope manufacturer has their own proprietary scan coil design, hysteresis may present differently in different microscopes [137].

For standard raster scanning, increasing the flyback time allows the beam to settle and become more stable at the beginning of each line scan, and the resulting edge distortions are not measured. However, this slows down the speed of the scan, making flyback time a source of inefficiency. A more favourable solution would be to directly the map the position of the beam to the counts measured by the detector, this way there is no wasted signal. Using an appro-

priate model, such as those proposed in works cited here [137–140], can be used to account for image distortions. A more robust solution would be the use of an electrostatic scan coil system as postulated by Kovarik *et al.* [141]. In theory, the delay on the voltage change across the coils would be sufficiently low that the beam position would change significantly faster.

2.4.4 Noise

Electron micrographs are corrupted by noise which is essentially undesirable counts that can often *hide* the expected contrast within an image. Noise can arise from various sources. The detectors themselves can induce noise artefacts through their own circuitry, and the electron source can contribute through thermal or electrical instabilities. Thermal instability within the sample can also cause noise, since the electrons and nuclei within the sample are never stationary. These vibrations, known as phonon excitations, lead to scattering of electrons away from the atomic equilibrium position. Furthermore, charging effects can give rise to noise, since the atomic potential is shielded by a residual electric field.

Even if all of the above is omitted by having the column at perfect vacuum, the microscope suspended in a vacuum free of external influence, and the sample at close to zero Kelvin, there is one source of noise which cannot be ignored- Poisson noise. Fundamentally, electrons are discrete packets of energy which, when detected, are counted by a detector. An image is simply an approximation of a probability distribution, which would in ideally be the amplitude of an exit wave-function within real space. In order to get the most accurate solution, the exposure time would have to be *infinite*. In practice, however, the exposure time is typically seconds, or in the case of STEM it is typically microseconds per probe location. This means that imaging is a balancing act between counts and exposure time, which for most materials is usually not a compromise but a requirement that must be fulfilled before the sample is irreversibly damaged.

To better describe this noise, it is helpful to consider the meaning from an electrons perspective. It can help to think as an electron would think as it is emitted from a source. As it leaves, it has no idea where it is meant to go, nor how long it has to do it. In fact, one could describe that emitted electron as lazy, simply caring about minimising its energy expenditure along any possible trajectory. In order to figure this out, omitting time constraints, the electron decides to consider all manner of trajectories that it can take. It might choose to go to the moon

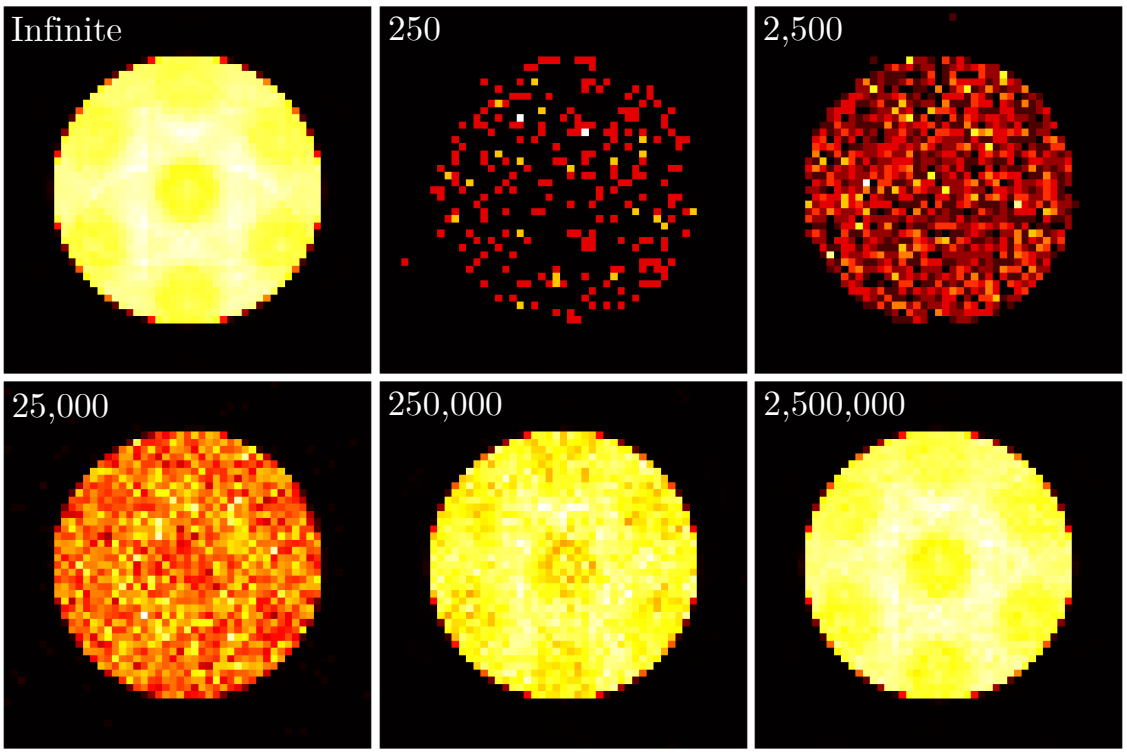

Figure 2.18: **Example of Poisson noise corrupted convergent beam electron diffraction pattern.** Increasing the electron fluence reduces the Poisson noise in the measured data. The number of electrons permitted per probe is indicated in the top left corner of each CBED pattern.

and back, it might oscillate within the column, or perhaps it might consider attempting to pass through the sample. In fact, it doesn't *consider* all of the trajectories, it takes *all* of them. But then, why is it only possible to count one electron?

This is the crux of quantum mechanics, indeterminism. The indeterminism gives rise to the Poisson noise since the observations are discrete. In the Copenhagen interpretation, the act of measuring forces the electron to *decide* which trajectory to take, which could be measured on the detector. Hence, the continuity of the wave-function then becomes a count on the detector and the electron resumes its particle like existence.

This counting regime is relatively simply to model. An exit wave-function can be calculated then taking the modulus squared will form a probability distribution. Determine an integrated fluence over the field-of-view and then distribute this over the sample using the exit wave-function as a bias. Each pixel will have it's own Poisson distribution, where the mean number of counts is determined by the fluence and the bias, and the variance is also determined by that same value. Examples are demonstrated in Fig. 2.18.

3 | Compressive Sensing and Image Inpainting

The field of signal processing concerns itself with addressing the conversion of continuous signals into discrete electronic signals such that the continuous signal can be recovered or estimated from the measurements. Images, specifically micrographs, are an example of such signals which are ideally continuous but are measured in a discrete form *i.e.*, pixels. In this chapter, the fundamentals of signal processing are considered such as Shannon-Nyquist theorem. This then leads into the theory of compressive sensing and ultimately image inpainting, whereby it is possible to recover an approximate representation of a signal from sparse direct measurements (subsampling in this case), and how this can be applied to STEM. Image recovery methods are presented and results are given, highlighting the limitations in all cases and specifically outlining the cases where subsampling may not be applicable.

3.1 Overview of Signal Processing, Compressive Sensing and Image Inpainting

To begin with, consider a simple sinusoidal function $f : \mathbb{R}_t \rightarrow [-A, A]$ which is a function of time $t \in \mathbb{R}_+$ given as,

$$f(t) = A \sin(2\pi B t) \ , \tag{1}$$

where $A \in \mathbb{R}$ is an amplitude and $B \in \mathbb{R}_+$ is a frequency. This signal is shown in Fig. 3.1. This signal is then to be collected using a detector which can only readout at a fixed frequency of B Hz. The question is then, what should the readout frequency be to fully recover the signal

from a few measurements as possible?

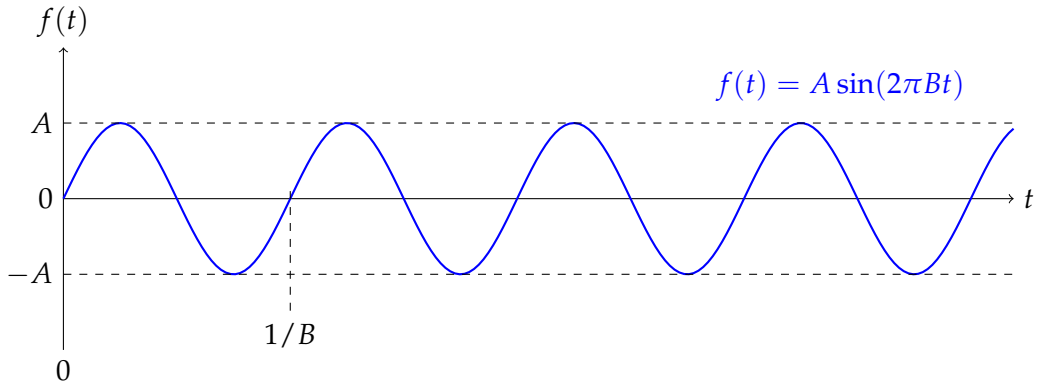

Figure 3.1: **Sinusoidal function.** The signal has a frequency B and an amplitude A given by Equation. 1.

The Nyquist-Shannon sampling theorem[1] (given here without proof) [142–144] states that a signal, say $f(t)$, which contains no frequency higher than B Hz, can be completely reconstructed from measurements which are equispaced at fewer than $1/2B$ seconds apart. The Nyquist-rate is therefore defined as $2B$. This is shown in Fig. 3.2, where the sampling frequency greater than the Nyquist-rate must be used to recover the signal.

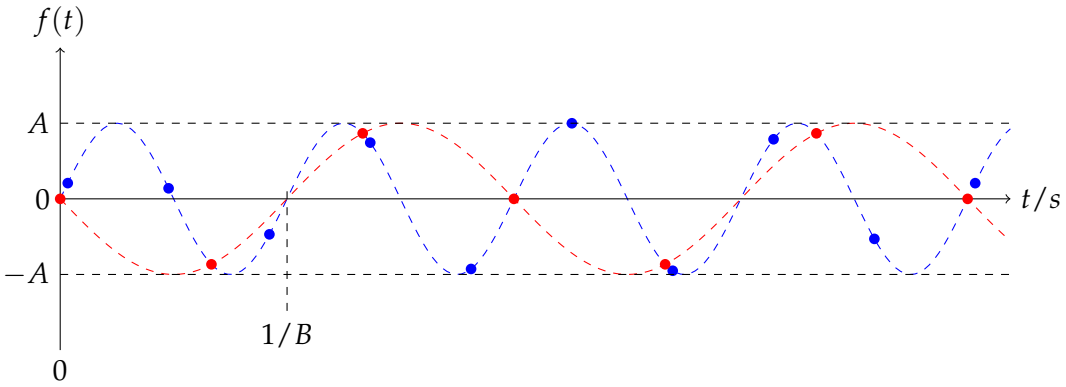

Figure 3.2: **Sinusoidal function sampled at various frequencies.** The function sampled and recovered at $1.1\times$ Nyquist-rate (blue) and the same function sampled and recovered at $0.75\times$ the Nyquist-rate (red). The lower sampling frequency exhibits aliasing, whereas the higher sampling frequency recovers the true signal. In each case, an interpolation with a sinc kernel is used to reover the signal.

For completeness, assume that the signal is acquired at a frequency of β Hz where $2B < \beta$. The Nyquist-rate is $2B$Hz, whereas the Nyquist frequency is defined by the sampling rate and is given as $\beta/2$; this determines the upper bound for the frequency which can be recovered given a selected sampling rate.

[1]Often also known as the Whittaker-Nyquist-Shannon sampling theorem given that the theorem was previously discovered by E.T.Whittaker in 1915.

Up until recently the mid 2000's, it was generally accepted that the Nyquist-rate was the limit which defined the minimum sampling required to recover a signal. However, work by David L. Donoho and Emmanuel J. Candés showed that it was possible to recover approximate signals from a set of incomplete (*i.e.*, below Nyquist-rate) measurements. This theory, known as Compressive Sensing (CS) [145–147], has been implemented in signal acquisition methods, such as magnetic resonance imaging (MRI) [148], radio interferometry [149], and infrared imaging [150]. In order for a signal to be acquired in a CS framework, it must meet certain criteria. The first of these is that the signal is sparse or compressible within some basis; their definitions and that of a dense signal are given below as,

- **Sparse**: A signal is sparse in the sparsity basis Ψ, if the majority of its components are zero, and only few components are non-zero valued.
- **Compressible**: A signal is compressible in the sparsity basis Ψ, if the majority of its components are approximately zero, and only a few measurements have a significant weighting.
- **Dense**: A signal is dense in the sparsity basis Ψ, if the majority of the components of that basis are non-zero valued and the above conditions are not-satisfied.

The second criteria is that the sampling basis must be incoherent with respect to the sparsity basis. That is, if a signal contains a few dominant frequencies then the sampling should be incoherent such that it does not match the dominant frequencies of the signal. Fig. 3.3 demonstrates that a signal is poorly estimated if the selected sampling mask matches the frequency of the signal it is trying to recover; this violates the criteria described above.

When an image is acquired (*i.e.*, the signal) considerations must be made as to whether it is compressible. The first step in this is to consider the Nyquist-rate for images, which describes the minimum pixel size used in order to avoid aliasing and incorrect upsampling. To do this, the information limit is considered for the sensing method or tool which performs the image acquisition. All imaging systems define their theoretical resolution based upon their contrast transfer function. The information limit is set by the spatial frequency at which no further non-zero contrast transfer can occur, which is defined as $k_{\max}$. In the case of STEM imaging, and following from section 2.3.2, this limit is set by the ratio between the convergence semi-angle (α) and electron wavelength (λ),

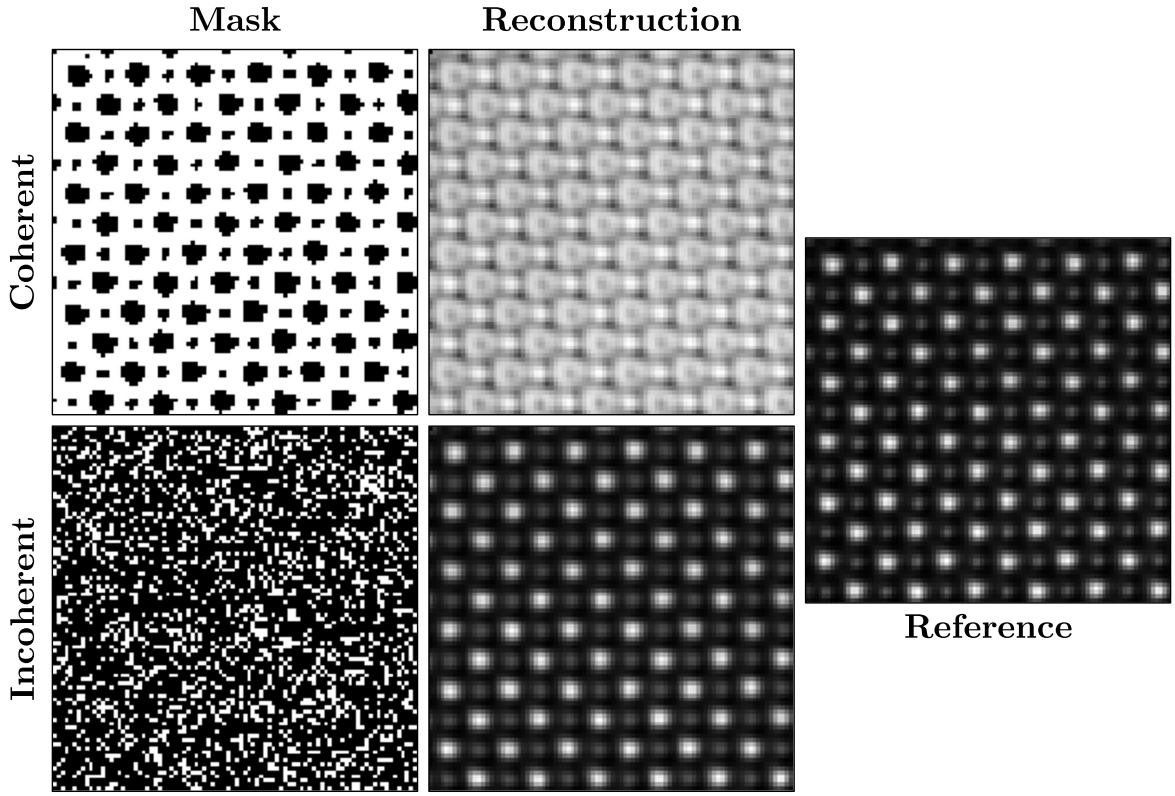

Figure 3.3: **Demonstration of correct mask selection based on incoherence property.** When a mask is selected which is coherent with the sparsity basis of the signal that it is concerned with, the recovery is poor, even if more measurements are taken. In the case of a random sampling the mask is incoherent with respect to the sparsity basis, and recovery is well estimated.

$$k_{\max} = \frac{2\alpha}{\lambda} \ . \tag{2}$$

Using this, the optimal scan-step Δ_p can be defined for high resolution STEM imaging which provides a result that can always be upsampled to give a desired pixel size. The result follows from the Nyquist-rate, which by taking $k_{\max}$ as the maximum frequency yields,

$$\Delta_p < \frac{1}{2k_{\max}} = \kappa_s \in \mathbb{R}^+ \ . \tag{3}$$

This limit κ_s states that for the highest resolution attainable in high-resolution HAADF STEM, the scan-step must be sufficiently small. Assuming an oversampled image with $\Delta_p << \kappa_s$, by performing an equispaced row-column down-sampling to increase the effective Δ_p, it is possible to see the effect of being *(i)* above, *(ii)* close to and *(iii)* below the Nyquist-rate. This

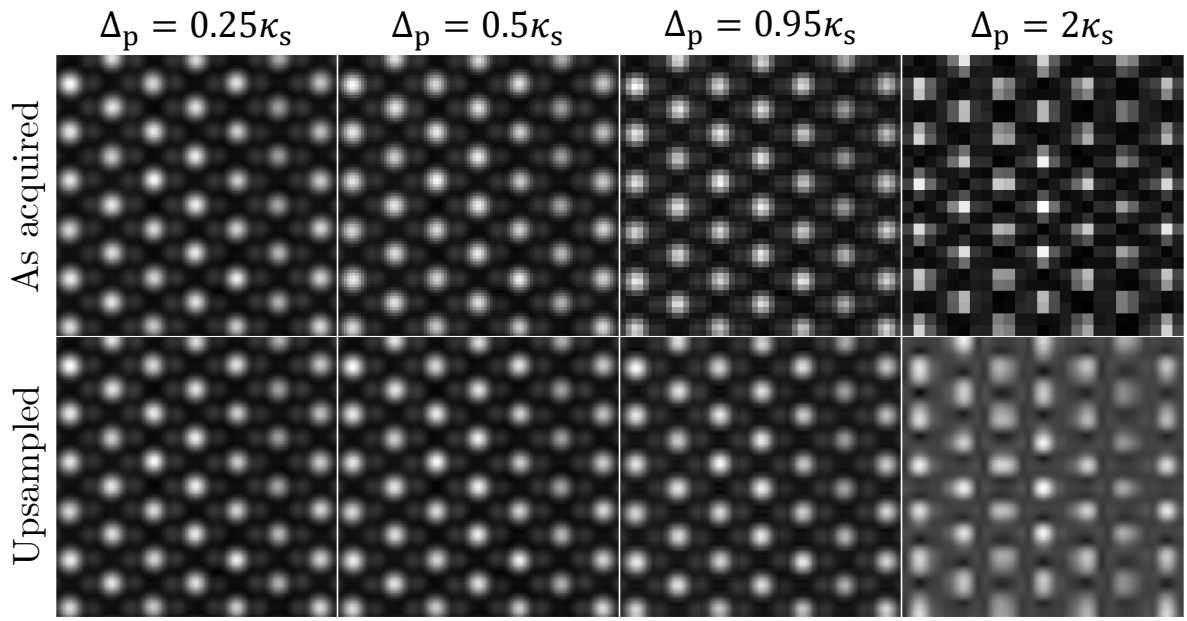

Figure 3.4: **Comparison of image upscaling with respect to scan-step at acquisition.** When the scan-step, Δ_p is sufficiently small, the image can be upscaled to arbitrary size without loss of information. However, if the scan-step is too large, the image cannot be upscaled since the sampling rate is less than the Nyquist-rate. The convergence semi-angle used for simulations here was 30mrad, an acceleration voltage of 300kV, and Scherzer defocus was also used.

exercise is given in Fig. 3.4.

In the context of this thesis, sub-sampling is employed as the branch of compressive sensing techniques. A random sub-set of measurements are acquired, and the data is then recovered using an inpainting algorithm which will promote sparsity in a sparsity basis.

3.2 Inpainting methods

Inpainting is a branch of signal processing, which aims to recover approximations of datasets from sparse measurements. In the context of images, this implies sub-sampled data where select pixels are missing, forming an incomplete dataset. There are several existing methods which can be used to recover an approximation of the fully-sampled dataset. These methods may include sparse coding steps such as an Orthogonal Matching Pursuit (OMP) [151, 152] using a predetermined dictionary, or a deep learning methodology where similar datasets are used to train a generative neural network [153, 154]. Each case has its benefits, but for a general solution, which requires no prior knowledge, two solutions are considered. The first is the Beta Process Factor Analysis with Expectation Maximisation (BPFA-EM, or here simply BPFA)

algorithm which is able to infer a dictionary from a sub-sampled acquisition, whilst using that dictionary to inpaint the missing data through an expectation maximisation inference step, however other inference techniques can be employed such as Gibbs sampling [155]. Secondly, a custom algorithm known as the (Regularised-) Local Means Inpainting (R-LMI), which uses a kernel interpolation strategy combined with a sparsity promoting regulariser. In this section, each technique is discussed, parameters described, and results for each method are presented.

3.2.1 Beta Process Factor Analysis

The BPFA algorithm is a powerful tool for inferring the missing data within a dataset, such as an image or a multi-dimensional dataset. At the core of the BPFA are dictionary learning and a sparse coding processes, however they are inherently connected to one another. In this section, details of the BPFA algorithm are discussed and justification for why it is suitable for inpainting sub-sampled EM data is given.

Lay description of the BPFA algorithm

The BPFA process, like most complex algorithms, is generally difficult to comprehend in terms of notation. During the past three years, it has been beneficial and ultimately helpful for discussions, to describe the process by considering an analogy. Here, a lay description is presented for the BPFA process and how the dictionary and missing pixels are inferred based upon sub-sampled data.

Consider a work of art, uncompleted by the original artist for whatever reason is appropriate. This artist had a unique style, one that no single artist alive could hope to replicate. Strangely, the original artist decided to paint the artwork one brush stroke at a time, analogous to a pixel in an image. The original artist also decided to paint in some random fashion, applying a sparse distribution of brush strokes over the canvas, forming a sub-sampled piece of artwork.

An art dealer comes along and knows that the original artist's work is well sought after, and decides that something must be done to complete the masterpiece so that they can make a nice profit. However, they cannot find the original artist, but the dealer knows a thing or two about Bayesian inference problems. They decide that by combining the abilities of multiple different artists, it may be possible to *inpaint* the artwork and approximately recover the

intended masterpiece, sufficiently so that it would convince most people into believing it was completed by the original artist.

The dealer then drafts in a group of artists from around the world, each with their own style. The dealer then decides to break the original artwork into a series of patches, ensuring that the patches overlap over certain regions of the artwork. Each of the patches are then stacked up, and one by one the artists get to work to solving what each of the patches should look like. Each of the artists represent a dictionary column, and each small patch of art represents an overlapping patch in an image.

When the first artist receives the first patch, they look at it and say, "oh, no this is not something I can help with, it is far too different to how I paint! I will have to change my style slightly to get more in tune with the original artists intentions". The second artist looks at the patch and says, "this is not too different to what I would do, I will try and have a go at guessing what bit of my style can be added, and measure the difference between what I do, and what the original artist painted. I will also slightly change my style based on how different my work is". The third artist then takes the patch from the second artist after they are complete and makes an appropriate judgement, similar to the first two artists. Each of the remaining artists do the same thing, and when they have all made their changes, the patch is put back onto the original artwork. This step is repeated for all the overlapping patches in the image, and overtime each of the artists start to converge towards a steady style for inpainting each of the patches. Eventually, after all the patches are inpainted, a weighted average is taken over the overlapping regions, and an approximation of the masterpiece is complete. A full cycle of all the patches through the different artists is known as an epoch, and this can be repeated multiple times, with the idea that each of the artists should get slightly better at inpainting over each epoch.

Once the masterpiece has gone through enough epochs, the dealer is satisfied with the results. The dealer looks at other artwork by the original artist and thinks that the styles are just about close enough. The dealer then goes to an auction, and the buyers are just as convinced. The inpainted artwork is now considered a masterpiece, although it is only an approximation of what it should have been.

This analogy describes how the BPFA algorithm completes an inpainting task, albeit somewhat reduced in complexity. The obvious caveats are associated with how much a patch

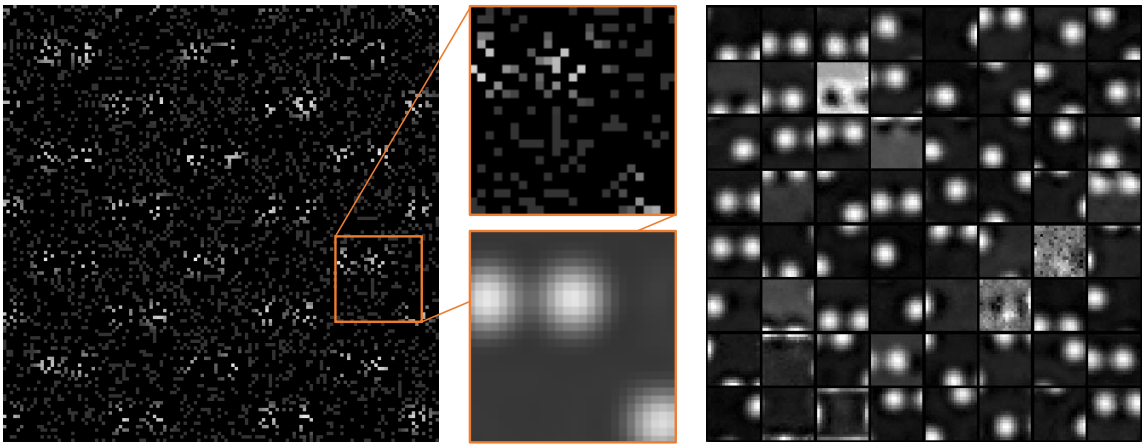

Figure 3.5: **Example of the inpainting process for one overlapping patch.** Demonstration of a subsampled image (left), one of the overlapping patches (top, middle), and the same patch reconstructed (middle, bottom) using the dictionary learned using the BPFA (right).

should be changed, and how much a dictionary atom should be updated. It is useful, however, to consider this kind of analogy when considering how to optimise a reconstruction. It could be that there are too few artists, too few patches, or too few brush strokes contained within each patch.

Mathematical basis of BPFA

Each of the parameters in the BPFA algorithm play a role in determining the final reconstruction. The meaning of each parameter is also important for understanding how it can be changed or optimised for the best reconstruction attainable. Each parameter for reconstructing an image of size $M \times N$ can be described according to,

- **Patch size**, $b \in \mathbb{N}$: The dimensions which defines the size by which the input should be broken into for dictionary learning and sparse coding. The patch size is typically square, but does not generally have to be. For a reconstruction, $2 \leq b < \min[M \times N]$.
- **Number of patches**, $N_p \in \mathbb{N}$: The total number of overlapping patches in the data given as $N_p = (M - b + 1) \times (N - b + 1)$. This number scales approximately with the image size.
- **Number of dictionary columns**, $K \in \mathbb{N}$: The number of basis atoms which can be learnt and then summed with corresponding weightings $\boldsymbol{\alpha} \in \mathbb{R}^K$ to reconstruct one of the (overlapping) patches from the image.
- **Sparsity limit**, $s \in \mathbb{N}$: The maximum number of dictionary atoms which can contribute

to the reconstruction of any given patch.

- **Number of patches per batch, or batch size**, $N_b \in \mathbb{N}$: Given that there are generally a large number of overlapping patches in the data, it is beneficial to randomly shuffle the patches into batches of size N_b. The dictionary is updated after a batch is processed.
- **Number of epochs**, $N_{\text{e}} \in \mathbb{N}$: This is the total number of passes over the full data set (*i.e.*, all patches).
- **Number of iterations**, $N_{\text{iter}} \in \mathbb{N}$: The total number of dictionary updates (or iterations) is $N_{\text{iter}} = N_{\text{e}} \times \lceil N_p / N_b \rceil$.
- **Learning rate**, $LR \in \mathbb{R}^{(0,1]}$: This number controls the step size at each update of the hyperparameters. This number is effectively the speed at which the dictionary atoms learn features of the image. If it is too small, convergence will be too slow. If it is too large then the convergence may not reach the correct minimum value.

The remaining parameters are omitted because they are not generally modified.

Given a sub-sampled measurement $\boldsymbol{y}$, first partition it into N_p overlapping patches $\{\boldsymbol{y}_i\}_{i=1}^{N_p}$, with each patch $\boldsymbol{y}_i \in \mathbb{R}^{b^2}$; hence, resulting in $N_p = (M - b + 1) \times (N - b + 1)$ total number of patches. Similarly, partition the sample image, mask operator, and noise as $\{\boldsymbol{x}_i\}_{i=1}^{N_p}$, $\{\boldsymbol{P}_{\Omega_i}\}_{i=1}^{N_p}$, and $\{\boldsymbol{n}_i\}_{i=1}^{N_p}$ respectively, such that for each patch $i \in \{1, \cdots, N_p\}$,

$$\boldsymbol{y}_i = \boldsymbol{P}_{\Omega_i} \boldsymbol{x}_i + \boldsymbol{n}_i \in \mathbb{R}^{b^2} \quad . \tag{4}$$

Furthermore, assume that each image patch is sparse in a shared dictionary, *i.e.*, $\boldsymbol{x}_i = \boldsymbol{D}\boldsymbol{\alpha}_i$, where $\boldsymbol{D} \in \mathbb{R}^{b^2 \times K}$ denotes the dictionary with K atoms and $\boldsymbol{\alpha}_i \in \mathbb{R}^K$ is a sparse vector of weights or coefficients for the i'th patch. Unlike traditional sparse coding approaches, which require a pre-defined dictionary or at least the number of dictionary atoms, here the desire is to jointly learn the shared dictionary and weights, given the subsampled measurements. The BPFA approach allows for inference of $\boldsymbol{D}$, $\{\boldsymbol{\alpha}_i\}_{i=1}^{N_{\text{p}}}$, K, and the noise statistics and in turn reconstruction of the sample image.

BPFA assumes that *(i)* the dictionary atoms $\{\boldsymbol{d}_k\}_{k=1}^{K}$ are drawn from a zero-mean multivariate Gaussian distribution; *(ii)* both the components of the noise vectors $\boldsymbol{n}_i$ and the non-zero components of the weight vectors $\boldsymbol{\alpha}_i$ are drawn *i.i.d.* from zero-mean Gaussian distributions;

Parameter	Value
b	16
N_{p}	12769
K	36
s	6
N_{b}	4096
N_{e}	7
LR	0.95

Table 3.1: **Parameters for simulation of BPFA with sub-sampling and realistic noise.** The values for each parameter corresponding to Fig. 3.7 for the test of the BPFA algorithm applied to data.

(iii) the sparsity prior on the weights is promoted by the Beta-Bernoulli process [156]. Mathematically, for all $i \in \{1, \cdots, N_p\}$ and $k \in \{1, \cdots, K\}$,

$$\boldsymbol{y}_i = \boldsymbol{P}_{\Omega_i} \boldsymbol{D} \boldsymbol{\alpha}_i + \boldsymbol{n}_i, \qquad \boldsymbol{\alpha}_i = \boldsymbol{z}_i \circ \boldsymbol{w}_i \in \mathbb{R}^K, \tag{5a}$$

$$\boldsymbol{D} = [\boldsymbol{d}_1^\top, \cdots, \boldsymbol{d}_K^\top]^\top, \qquad \boldsymbol{d}_k \sim \mathcal{N}(0, B^{-2} \boldsymbol{I}_{B^2}), \tag{5b}$$

$$\boldsymbol{w}_i \sim \mathcal{N}(0, \gamma_w^{-1} \boldsymbol{I}_K), \qquad \boldsymbol{n}_i \sim \mathcal{N}(0, \gamma_n^{-1} \boldsymbol{I}_{B^2}), \tag{5c}$$

$$\boldsymbol{z}_i \sim \prod_{k=1}^K \text{Bernoulli}(\pi_k), \qquad \pi_k \sim \text{Beta}(\tfrac{a}{K}, \tfrac{b(K-1)}{K}), \tag{5d}$$

where $\boldsymbol{I}_K$ is the identity matrix of dimension K, operator $\circ$ denotes the Hadamard product, and a and b are the parameters of the Beta process. The binary vector $\boldsymbol{z}_i$ in (5d) determines which dictionary atoms to be used to represent $\boldsymbol{y}_i$ or $\boldsymbol{x}_i$; and π_k is the probability of using a dictionary atom $\boldsymbol{d}_k$. In (5c), γ_w and γ_n are the (to-be-inferred) precision or inverse variance parameters. It is common to place a non-informative, *i.e.*, flat, gamma hyper-priors on γ_w and γ_n, by fixing them to small values [157]. The sparsity level of the weight vectors, *i.e.*, $\{\|\boldsymbol{\alpha}_i\|_0\}_{i=1}^{N_p}$ is controlled by the parameters a and b in (5d). However, as discussed in [158], those parameters tend to be non-informative and the sparsity level of the weight vectors is inferred by the data itself.

Unknown parameters in the model above can be inferred using Gibbs sampling [158], variational inference [156], or (as in this thesis) Expectation Maximisation (EM) [159, 160]. In short, EM involves an expectation step to form an estimation of the latent variables, *i.e.*, $\{\|\boldsymbol{\alpha}_i\|_0\}_{i=1}^{N_p}$, and a maximisation step to perform a maximum likelihood estimation to update other parameters. Since the number of patches N_p may be large, a stochastic (or mini-batch) EM approach is implemented, where the N_p patches are (randomly) partitioned into batches of size N_b and those batches are processed sequentially. Similar ideas have been used in [160, 161].

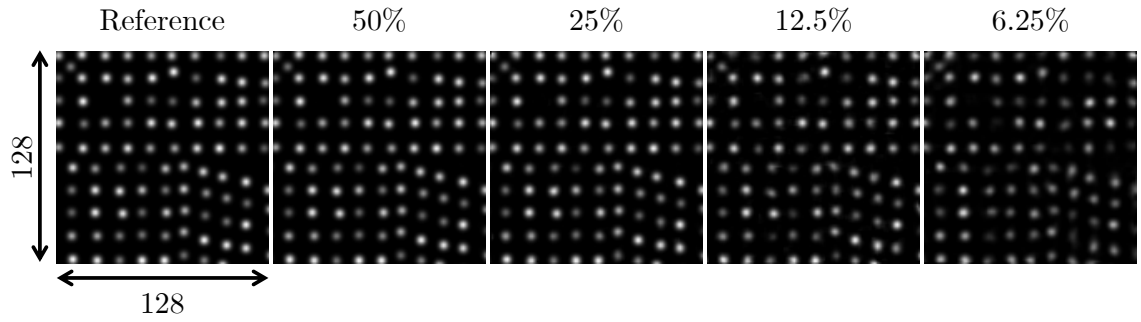

Figure 3.6: **Testing the BPFA algorithm on a complex structure.** Inpainting results at various sampling ratios (above each reconstruction) for the complex structure containing various defects such as an interstitial dopant, a vacancy, a lattice distortion, a grain boundary, and a screw dislocation. The radii of the atoms in the structure are approximately 3 - 3.5 pixels, which is equivalent to roughly a 0.25Å -0.35Å scan step.

Applying BPFA to data

In order to test the suitability of the BPFA algorithm for inpainting STEM data, a reconstruction series is performed for high resolution simulated images of silicon dumbbells. Different dose levels are also applied to test the algorithms robustness to noise. The noise model is Poisson, whereby a simulated reference provides the likelihood of electron detection at a certain pixel. As the dose increases, the Poisson noise approximates Gaussian noise as the BPFA expects. The total number of electrons permitted is the integral of the electron fluence across the field of view. There is also a bias which permits detection at any region to mimic spurious counts. The model follows the simplified noise model from section 2.4.4. The electron fluence was varied from 500eÅ^{-2} up to $\sim 10^6$eÅ^{-2}. The parameters for the BPFA were the same for all experiments and are given in Table 3.1.

As the results in Fig. 3.7 show, it is possible to recover the data from far fewer measurements than initially acquired, even with low signal. In fact, the BPFA is able to denoise the data significantly, increasing the perceived signal-to-noise ratio. Although this is a periodic structure, which is relatively simple to inpaint, the results indicate that distributing the dose in a sub-sampled regime can yield visually identical results to the raw data at higher doses.

There are various works which demonstrate the BPFA algorithm (and its variants) recovery quality for different image types, as well as electron microscopy data [158, 162–164].

To further demonstrate the BPFA algorithm's robustness to complex structures, an image is constructed containing various types of defects. These defects include an interstitial dopant,

a vacancy, a screw dislocation, a lattice distortion, and a grain boundary. Furthermore, each of the 'atoms' have randomly assigned intensity to further complicate the structure. The model is taken from [165] and reconstructions demonstrated in Fig. 3.6. As can be seen, the BPFA does not care if the sample is periodic or not, since each patch of the image is operated on independently. This demonstrates that the BPFA (with careful parameter selection) is robust to the inpainting of complex nanoscale structures.

3.2.2 Regularised Local Means Inpainting

Another approach to image inpainting is through a technique referred to as Regularized Local Means Inpainting (R-LMI). This method follows a common class of inpainting techniques known as interpolation, however with the added functionality of sparsity promotion though regularization in the Fourier or discrete cosine transform (DCT) domain. Furthermore, it combines kernel convolution to accurately estimate pixel values based on local (i.e., nearby) pixel values.

Inpainting process of R-LMI

Assume a 2-D signal $\mathbf{Y} \in \mathbb{R}^{M\times N}$ which only contains $\mu \in \mathbb{N}$ measurements where $\mu \ll \eta = M \times N$. The set of pixels which are not sampled are given as $\hat{\Omega} \subset \Omega := \{1, \cdots, M\} \times \{1, \cdots, N\}$ such that a mask $\boldsymbol{\mathcal{M}}_\Omega \in \{0,1\}^{M\times N}$ is defined where $\boldsymbol{\mathcal{M}}_{\Omega_p} = 0$ if $\boldsymbol{p} \in \hat{\Omega}$ and $\boldsymbol{\mathcal{M}}_{\Omega_p} = 1$ otherwise.

Secondly, assume a Gaussian kernel $\boldsymbol{K} \in \mathbb{R}^{H_\mathrm{K}\times W_\mathrm{K}}$ where $H_\mathrm{K} \in \mathbb{N}$ and $W_\mathrm{K} \in \mathbb{N}$ are the height and width of the kernel, respectively. The kernel is then defined as,

$$\boldsymbol{K}(\boldsymbol{r}, \sigma) = \exp\left(-\frac{||\boldsymbol{r}_0 - \boldsymbol{r}||^2}{\sigma^2}\right) , \tag{6}$$

where $\boldsymbol{r}_0$ is the centre of the kernel, and $\boldsymbol{r} \in \{1, \cdots, M\} \times \{1, \cdots, N\}$ is a pixel location within the kernel. The next step is to consider a non-sampled pixel location $\boldsymbol{p}$ and the region of the image $\boldsymbol{y}_p \in \mathbb{R}^{H_\mathrm{K}\times W_\mathrm{K}}$ which is contained where the centre is at $\boldsymbol{p}$ and will be referred to as a patch of the image. Following this, consider the mask which is also contained within that region $\boldsymbol{m}_p \subset \boldsymbol{\mathcal{M}}_\Omega$ *i.e.*, the mask-patch with the same size as the kernel. The value of the pixel at location $\boldsymbol{p}$ is then given as,

$$\hat{\boldsymbol{X}}(i,j) = \begin{cases} \boldsymbol{Y}(i,j), & \text{if } (i,j) \text{ is sampled.} \\ \frac{\sum_{j'=1}^{W_K}\sum_{i'=1}^{H_K}\boldsymbol{K}(i',j')\boldsymbol{y_p}(i-i',j-j')}{\sum_{j'=1}^{W_K}\sum_{i'=1}^{H_K}\boldsymbol{K}(i',j')\boldsymbol{m_p}(i-i',j-j')} & \text{if } (i,j) \text{ is not sampled.} \end{cases}, \tag{7}$$

and Eq. 7 is then repeated for all $\boldsymbol{p}$ to form the full reconstruction $\hat{\boldsymbol{X}}$. Once this is completed, a regularization step is performed to promote sparsity in some orthonormal basis A such as the Discrete Cosine Transform (DCT). The regularization can be either hard-thresholding or soft-thresholding, or commonly known as l_0-norm or l_1-norm regularization respectively. For l_0-norm regularization, the result is obtained by,

$$\hat{\boldsymbol{X}} = A^{-1}\Big[H_\kappa\big[A[\hat{\boldsymbol{X}}]\big]\Big] \tag{8}$$

where H_κ is a hard threshold keeping only κ largest components from the basis A. For l_1-norm regularization, the result is obtained by,

$$\hat{\boldsymbol{X}} = A^{-1}\Big[S_{\kappa,\tau}\big[A[\hat{\boldsymbol{X}}]\big]\Big] \tag{9}$$

where S_κ is a soft threshold keeping only κ largest components from the basis A and shrinking them by a factor of τ which is the smallest value of the remaining components which are non-zero.

Applying R-LMI to data

Now that R-LMI has been defined mathematically, it is natural to consider where it may be applied whilst still returning functionally identical results. Based on the underlying nature of R-LMI, it is safe to assume that it is best suited to smooth data which is dominated by low spatial frequency signals. For demonstration, high-resolution STEM simulations of silicon dumbbells are considered to find suitability, the same images which are considered in section 3.2.1.

As Fig. 3.8 shows, the R-LMI algorithm is not robust to high noise levels, which is expected even with regularisation. This is due to the fact that the noise is not inferred during the inpainting step, and as such can be amplified during inpainting. This can lead to artefacts

and ultimately poor recovery. However, when the dose is higher, the quality is significantly improved. At these higher doses, the noise approximates a Gaussian noise, which is far easier to discard in the regularisation step. Furthermore, at the higher doses and sampling rates, the R-LMI is somewhat comparable to the BPFA results, indicating that R-LMI is most suited for fast recovery of sufficiently sampled, high signal datasets.

3.3 The importance of patch size and kernel size for the BPFA and R-LMI algorithms

In the above sections, there has been little mention of the importance of selecting the correct parameters for reconstruction. The work by Nicholls *et al.* [164] neatly outlines the various parameters of the BPFA algorithm, outlining the most important parameters which require some form of tuning. The most important parameter, as determined by this work, is the *patch size*. The patch size is also analogous to the kernel size as in the R-LMI algorithm, and ensuring that a correct size is used is important to optimise both *(i)* speed and *(ii)* quality of reconstruction. Essentially, the patch size should be in the Goldilocks zone- not too small, not too big, but just right for the problem at hand. It is observed that when the patch size is too small, the image reconstruction is speckled, and when it is too big (without changing other parameters), the image is blurred or inconsistent with the ground truth.

The first step to determining the correct patch size based on the provided input is to consider the simplest case. Imagine if the input data was simply a matrix of ones, and the mask was sampled purely at random at a specific sampling rate, which will be referred to as the global sampling rate $g \in \mathbb{R}^{(0,1]}$. Now, imagine one of the many overlapping patches in that dataset, where the dimension of that patch is $b \times b$. The next question to ask is, what are the chances that one of the overlapping patches contains no sampled data? This is the first step to identifying a lower-bound, since it is assumed that a patch must contain data if it is to be inpainted.

Consider the p^{th} overlapping patch of size $b \times b$, where $b \in \mathbb{N}$, from a 2-D data where $M \in \mathbb{N}$ measurements are taken from a possible set of $N \in \mathbb{N}$ such that $M \leq N$. Let $r_i \in \{0, 1\}$ be a binary variable for $i \in \{1, \cdots, N\}$ such that,

$$r_i = \begin{cases} 1 & \text{if sampled.} \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

The number of measurements within the p^{th} patch, $\hat{M}_p$, is therefore,

$$\hat{M}_p = \sum_i r_i \ , \tag{11}$$

which has an expectation value $\mathbb{E} \in \mathbb{R}$ given as,

$$\mathbb{E}[\hat{M}_p] = b^2 \cdot \frac{M}{N} \ . \tag{12}$$

For each pixel in the patch, the likelihood that it will be sampled is $M/N = g$, and the likelihood that it won't be sampled is $1 - M/N = 1 - g$. For the total patch, there are b^2 pixels, which can be thought of as trials. Therefore this follows a binomial distribution where the variance on the number of measurements observed within the overlapping patch is,

$$\text{Var}(\hat{M}_p) = b^2 \cdot g \cdot (1 - g) \ . \tag{13}$$

The local sampling ratio, *i.e.*, the sampling ratio of the p^{th} is therefore the number of measurements within the p^{th} patch $\hat{M}_p$ divided by b^2 with expectation value (proof in the appendix A2.1, Eq. 3),

$$\mathbb{E}\left[\frac{\hat{M}_p}{b^2}\right] = g \ , \tag{14}$$

and variance,

$$\text{Var}\left[\frac{\hat{M}_p}{b^2}\right] = \frac{g(1-g)}{b^2} \ . \tag{15}$$

For sufficiently large samples, this binomial distribution will approximate a normal distribution [166] such that the local sampling ratio is approximately distributed as,

$$l \sim \mathcal{N}\left(g, \ \frac{g(1-g)}{b^2}\right) \ , \tag{16}$$

where $b \in \mathbb{N}$ is the patch size. The result implies that the likelihood of an overlapping patch containing no sampled data is minimised if *(i)* a higher sampling ratio is used or *(ii)* a larger patch size is used, since the mean should be maximised and the standard deviation minimised in this case.

To verify this through observation, a series of Monte-Carlo simulations are performed with various sampling rates and patch sizes. The local sampling ratio, *i.e.*, the sampling ratio within any one given overlapping patch $l_{\mathrm{p}} \in \mathbb{R}^{[0,1]}$ is determined. The standard deviation of this local sampling ratio is then calculated, and Fig. 3.10 is an example visualisation of the standard deviation of the data with respect to the global sampling ratio for a patch size of 16.

This empirical finding supports the conclusions drawn in Eq. 16. To get a minimum value for the patch size, consider the case where the mean is equal to the standard deviation,

$$\begin{aligned} g &= \frac{\sqrt{g(1-g)}}{b} \\ \implies b &= \sqrt{\frac{1-g}{g}} \ . \end{aligned} \tag{17}$$

By plotting the patch size as a function of the global sampling ratio (as seen in Fig. 3.11), if the value is positive then it implies a 86.4% probability that none of the overlapping patches will contain zero sampling. To increase this likelihood, to greater than 97.5%, then it follows that,

$$b \geq 2\sqrt{\frac{1-g}{g}} \ , \tag{18}$$

and if the likelihood is to be greater than 99.8%,

$$b \geq 3\sqrt{\frac{1-g}{g}} \ , \tag{19}$$

which implies that the minimum patch size $b_{\mathrm{min}} \in \mathbb{N}$ for UDS should be,

$$b_{\mathrm{min}} = 3\sqrt{\frac{1-g}{g}} \ . \tag{20}$$

Therefore, a minimum bound on the patch size for any uniform density sampling mask can

be determined based on the users desired confidence that all patches contain sampled data. The problem now, however, is that a line-hop mask is not an ideal uniform density sampling mask. The 'randomness' of line-hop is set somewhere between random row-wise sampling and uniform density sampling, therefore consider the case of applying a row-wise mask. In the case of row-wise sampling, only one of the axis has a random selection. By following the same logic as for UDS, the local sampling distribution is approximately normally distributed according to,

$$l \sim \mathcal{N}\left(g,\ \frac{g(1-g)}{b}\right)\ , \tag{21}$$

i.e., the variance is increased by a factor of b with respect to equation 16.

Now consider how a line-hop mask is constructed. The sampling ratio is given as $1/(r_\mathrm{h} + r_\mathrm{p})$ where $r_\mathrm{h} \in \mathbb{N}$ and $r_\mathrm{p} \in \mathbb{N}_0$ are the row height and row padding, respectively. If the patch size is selected such that $b \geq 2(r_\mathrm{h} + r_\mathrm{p})$, then this guarantees that every overlapping patch will contain data. If $b < 2(r_\mathrm{h} + r_\mathrm{p})$, then the problem becomes more complex, however by equation 21, the hard constraint on the lower bound $b_\mathrm{min} \in \mathbb{N}$ can be given as,

$$3\sqrt{\frac{1-g}{g}} \leq b_\mathrm{min} \leq 2(r_\mathrm{h} + r_\mathrm{p}) \leq 3\left(\frac{1-g}{g}\right)\ , \tag{22}$$

which is satisfied for $\forall g \in (0, \frac{1}{3}]$, and

$$3\sqrt{\frac{1-g}{g}} \leq b_\mathrm{min} \leq 2(r_\mathrm{h} + r_\mathrm{p})\ , \tag{23}$$

is satisfied for $\forall g \in [\frac{1}{3}, \frac{1}{2}]$ and,

$$b_\mathrm{min} \leq 2(r_\mathrm{h} + r_\mathrm{p})\ , \tag{24}$$

is satisfied for $\forall g \in [\frac{1}{2}, 1]$.

Empirically, a slightly tighter lower bound for $> 99.8\%$ confidence that all patches are sampled is given without proof as,

$$\frac{3}{2}(r_\mathrm{h} + r_\mathrm{p}) \leq b_\mathrm{min} \leq 2(r_\mathrm{h} + r_\mathrm{p})\ , \tag{25}$$

which is satisfied for $\forall g \in (0, 1]$. Note, that for all of the cases given above, the patch size used should be changed according to $b_{\min} = \lceil b_{\min} \rceil$.

One other consideration for line-hop is the use of non-square patch shapes. The above condition applies for the row size of the patch, assuming that the line-hop is row-wise. This means a smaller column size of the patch can be used to increase the speed of the algorithm and to reduce blurring effects. That case is omitted here due to algorithm support. It is also noted that the same arguments for minimum kernel size can be used, *i.e.*, kernel size and patch size are interchanged.

The next logical step is to estimate an upper bound for the patch size. This is a difficult task since various other parameters can also influence the quality of reconstruction once the patch size is sufficiently large. By making assumptions about the data, it is possible to estimate this value. For example, assuming the data contains perfectly periodic structure, such as that for a pristine atomic resolution image, the patch size can theoretically be any size greater than the minimum value. However, a similar image, now containing a vacancy may not also have the same benefit. To test this, a set of reconstructions are performed on two simulated images of molybdenum disulfide- one is pristine, and one contains various sulfur vacancies.

As Fig. 3.12 shows, when the data contains aperiodic features such as vacancies, the patch size must be selected carefully to avoid incorrect inpainting of artefacts. In the case of periodic structures, the patch size is less important since the same patch can be used to represent multiple regions of the image. On the other hand, the number of dictionary elements and sparsity limit can be adjusted to improve the performance at larger patch sizes. This is demonstrated in Fig. 3.13.

The reason for this can be thought of as the ability for the dictionary to learn more localised features given that there are more dictionary atoms available to populate. According to Beal [167], the number of dictionary columns K should have an upper bound given by,

$$K = \left\lceil b^2 + \frac{1}{2}\left[1 - \sqrt{1 + 8b^2}\right] \right\rceil \ , \tag{26}$$

however there is no lower bound to indicate how small the dictionary can be, and the investigation of this is left to future work. For now, parameters should be selected based on efficiency for live use of the BPFA algorithm, and tuning can be done offline for more thorough analysis,

as Nicholls *et al.* described in [164].

3.4 Conclusions

This chapter has outlined the basic principles of compressive sensing theory, as well as the fundamental properties of sampling and sufficient sampling. Ultimately, for a signal to be compressed, it must be compressible or sparse within a basis set. Given the forward sensing model of some electron imaging modes, those images are inherently compressible. In those cases where there is an image which is not compressible, the image may still be oversampled.

Two inpainting techniques are presented, firstly the BPFA, which is suitable for approximate recovery from compressed measurements (*i.e.*, the sampling is below the Nyquist-rate) of noisy data, then secondly the R-LMI which is suitable for sufficiently high signal datasets with sufficient sampling. It is also important to note that the sampling limit, which determines whether the R-LMI can be used is signal dependent, with further investigation left for future work.

At this stage, a motivation for subsampling has been presented, as well as a method to approximately recover the signal. The next step is to consider how this can be done in practice, and how it may be extended to multi-dimensional datasets.

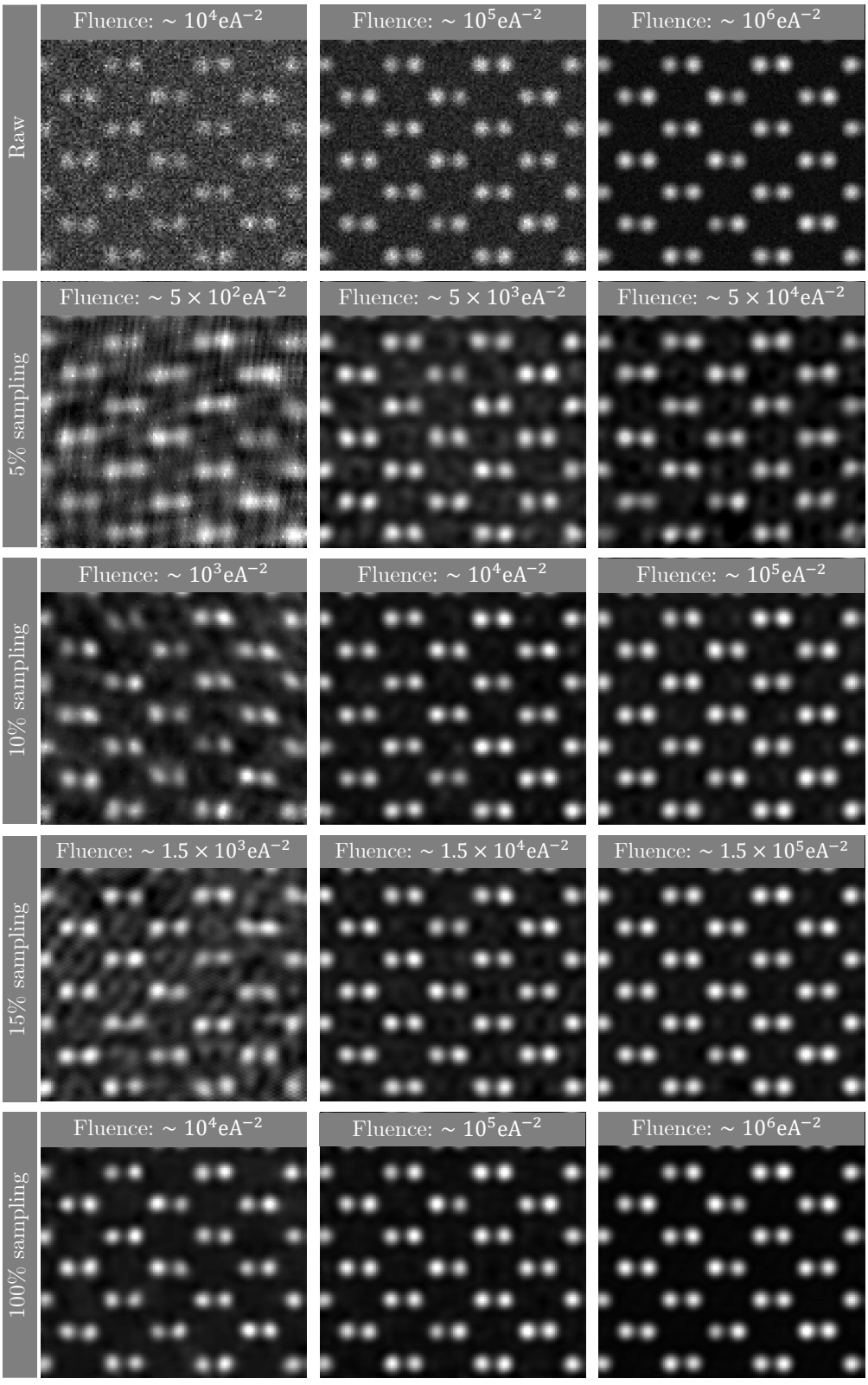

Figure 3.7: **Testing the BPFA algorithm at different fluences and sampling rates.** The top row contains the raw data as acquired with different electron fluence, and the remaining rows are reconstructions through the BPFA algorithm at 5%, 10%, 15% and 100% respectively from top to bottom. Each column corresponds to the raw data in the top row. At the top of each image is the indicated electron fluence. The convergence semi-angle used for simulations here was 30mrad, an acceleration voltage of 300kV, and Scherzer defocus was also used. The scan-step is 0.125Å, which is finer than the Nyquist sampling rate of 0.1642Å.

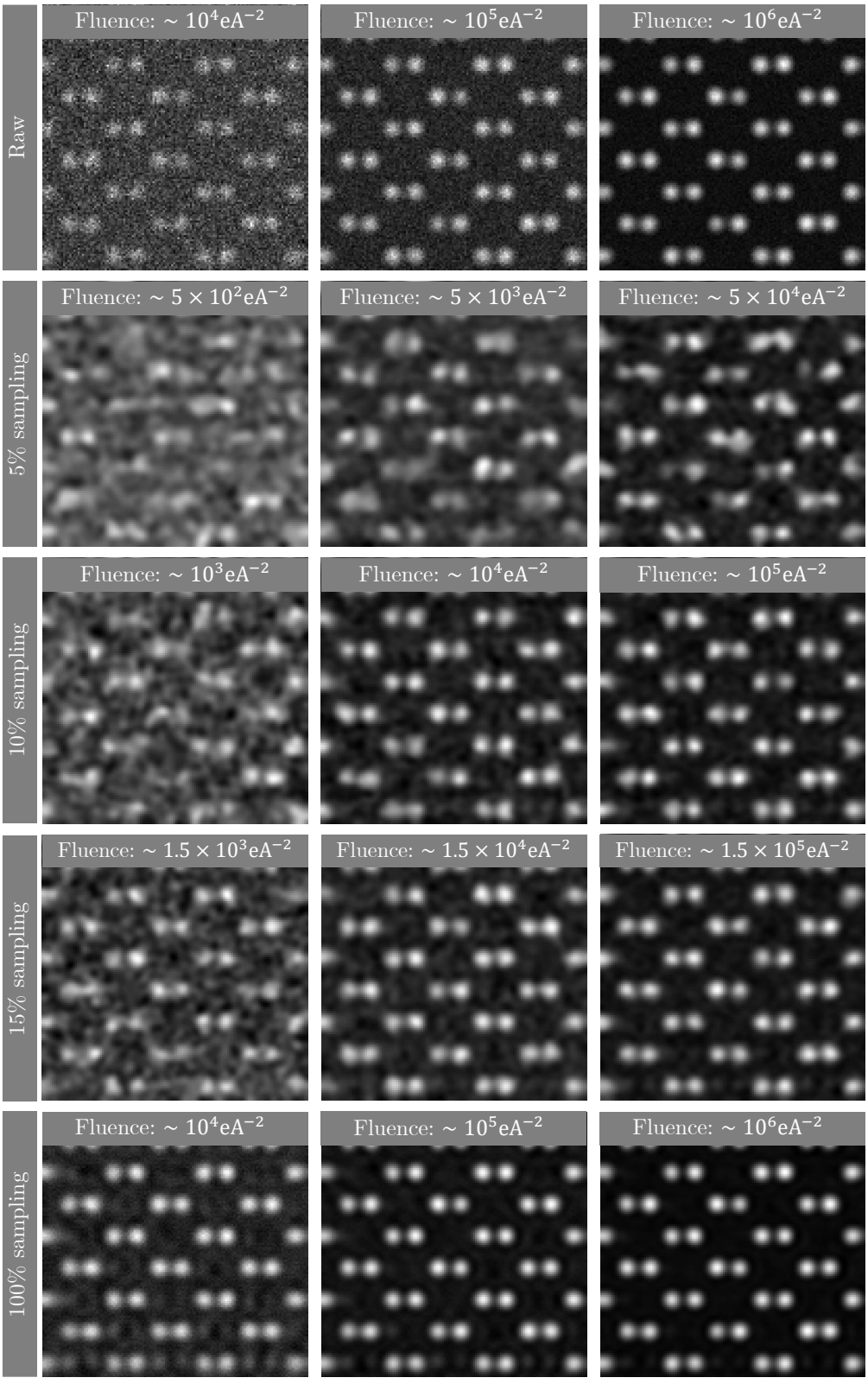

Figure 3.8: **Testing the R-LMI algorithm at different electron fluences and sampling rates.** The top row contains the raw data as acquired with different electron fluence, and the remaining rows are reconstructions through the R-LMI algorithm at 5%, 10%, 15% and 100% respectively from top to bottom. Each column corresponds to the raw data in the top row. At the top of each image is the indicated electron fluence. The convergence semi-angle used for simulations here was 30mrad, an acceleration voltage of 300kV, and Scherzer defocus was also used. The scan-step is 0.125Å, which is finer than the Nyquist sampling rate of 0.1642Å.

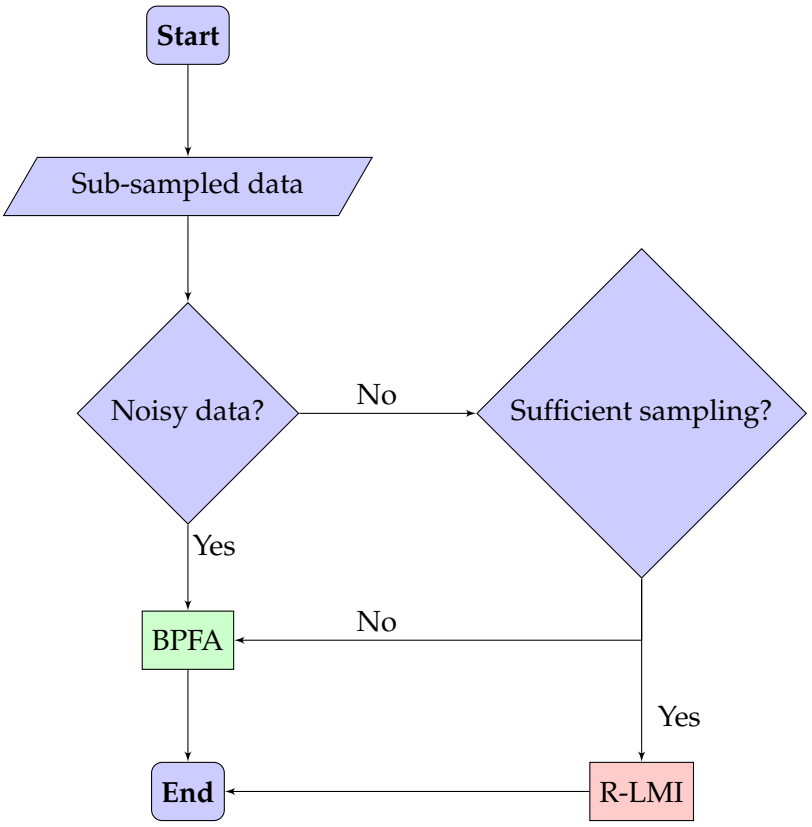

Figure 3.9: **A workflow for deciding which of the algorithms to use for speed and simplicity.** If the motivation of recovery is to arrive at the best solution, then an optimised BPFA should provide this best solution. However by considering the properties of the input data, the most efficient recovery *i.e.*, the one which generates a sufficient solution in the shortest amount of time, may be found using the R-LMI or the BPFA.

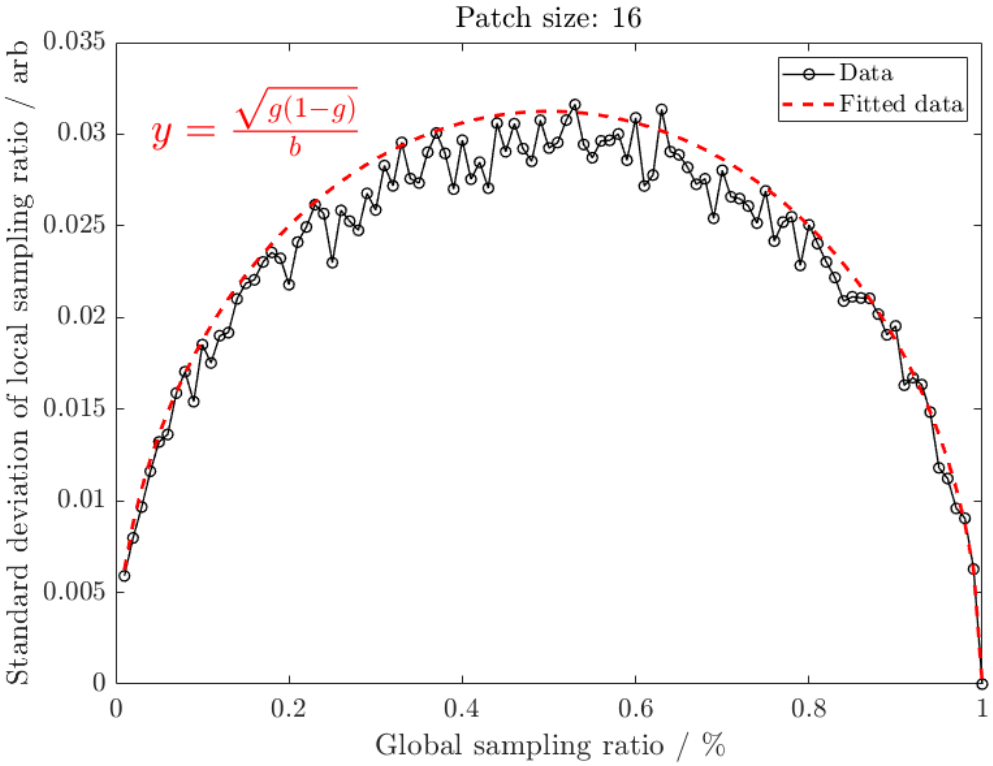

Figure 3.10: **Local sampling standard deviation as function of global sampling ratio for patch size of 16.** The empirical results match the findings deduced from Eq. 16

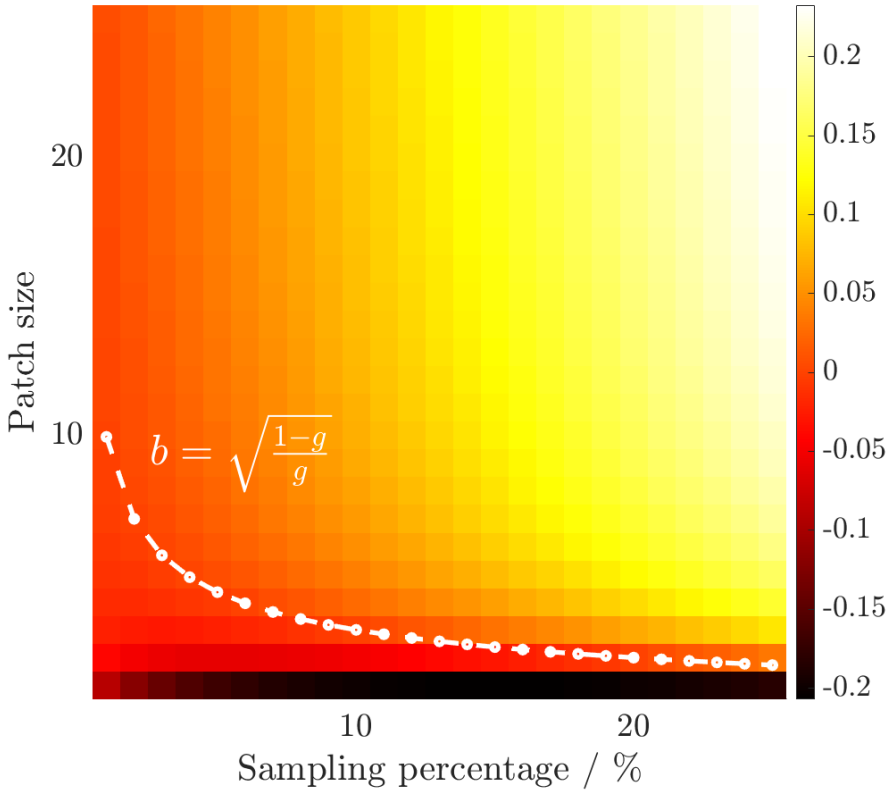

Figure 3.11: **Difference between the mean and standard deviation according to equation 16 as a function of global sampling ratio and patch size.**

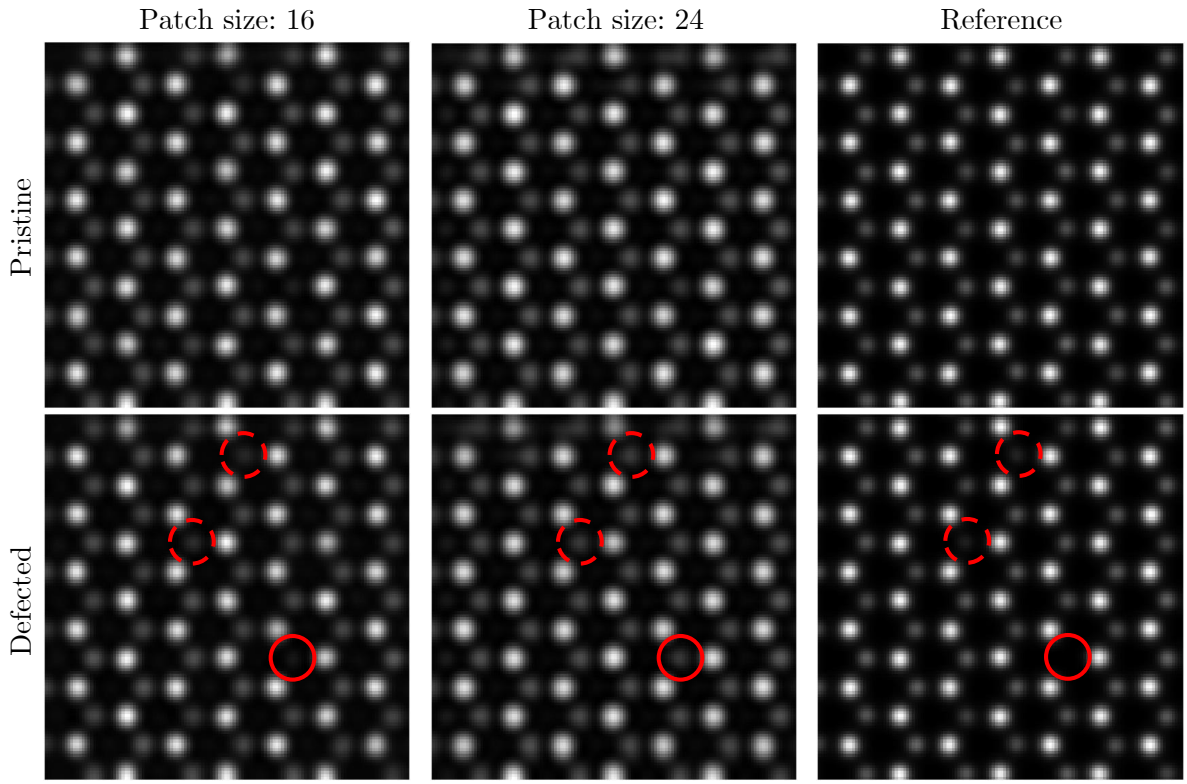

Figure 3.12: **Results of using different patch sizes to inpaint a MoS_2 simulated HAADF images; pristine and containing vacancies.** Using the incorrect patch size can lead to incorrect inpainting, especially if it is too large, or too small. The dashed red lines indicate single sulfur vacancies, and the solid red line indicates a double sulfur vacancy. When the sample is pristine, the choice of patch size is less important since the periodicity does not change. The convergence semi-angle used for simulations here was 39.1mrad, an acceleration voltage of 60kV, and Scherzer defocus was also used. The scan-step is 0.1575Å, which is finer than the Nyquist sampling rate of 0.3111Å.

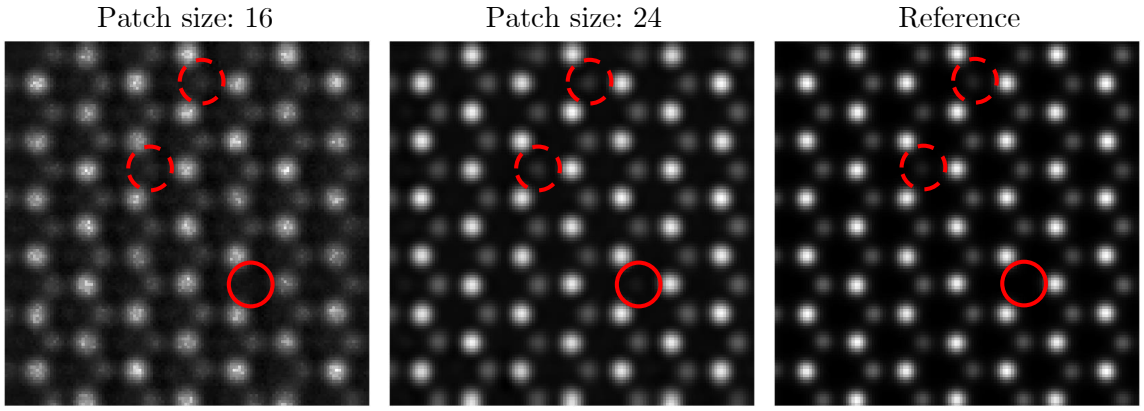

Figure 3.13: **Results of increasing the number of dictionary atoms and sparsity limit for inpainting a MoS_2 simulated HAADF image containing vacancies.** By increasing the number of dictionary atoms, as well as increasing the sparsity limit, there is a significant improvement in the reconstruction using larger patch sizes to recover the contrast of the vacancies.

4 | Applying Compressed Sensing Methods to STEM

4.1 Overview

As highlighted in this work, despite the progress made in STEM over recent decades, certain drawbacks have arisen that have ultimately limited the applicability of STEM analysis to a large number of samples. With the emergence of increasingly efficient and brighter electron sources, coupled with the advancement of aberration correction technology capable of focusing the electron probe to sub-angstrom dimensions, and the aspiration to achieve high-precision imaging of individual atoms, the electron beam probes have become remarkably intense. STEMs function based on an electron probe (typically smaller than 0.1nm) that scans across a material, and the resulting interaction between the electrons and the sample is captured at each position within the specimen using tailored detectors. The resulting interaction can be visualized through the acquisition of transmitted or scattered electrons using fixed monolithic radial detectors and pixelated detectors, or involving spectroscopy by gathering X-rays (Energy-Dispersive X-ray Spectroscopy or EDS), or measuring the energy loss of the transmitted electron beam (Electron Energy-Loss Spectroscopy or EELS).

For beam-stable materials which can undergo exposure to the intense electron probes, this is all well and good. However, for the wide range of samples which cannot, other considerations must be made which take in account the beam-influence. Beam damage mechanisms were discussed in detail in section 2.4.1, with radiolysis perhaps being the most dominant mechanism which plagues the analysis of low-Z number materials, organic matter, and hybrid materials such as metal-organic frameworks.

Another consideration is *speed*. Typically, a STEM will operate with a dwell-time (the time spent at each probe location within a raster scan) on the order of $10 - 20\mu$s. Therefore, for a megapixel image, the time-to-acquire is on the order of $10 - 20$s, without accounting for flyback[1]. This makes in-situ STEM scan-size limited (reduce the number of pixels) and signal limited (reduce dwell-time). Furthermore, if a sample is drifting due to thermal instability or charging, the image will distort during the scan which means a user will have to spend more time acquiring data or wait until the microscope is stable, which for some reason always seems to be at 5pm.

Although several methods have been employed to overcome beam damage through low-dose techniques, as well as including cryogenic-stage STEM methods, it may be favourable to simply focus on a simpler objective: acquire the minimum amount of signal necessary in the shortest amount of time possible. Regardless of the sample, what has been described there is efficiency and the requirement to eliminate redundancy.

A candidate solution, which can be applied under any existing solution to overcome the aforementioned limitations, is the inclusion of a compressive sensing (CS) or sub-sampling approach. In this chapter, details on how a CS-STEM experiment is performed and examples of where it has been applied are presented. Furthermore, the chapter concludes with the inclusion of theory as a potential catalyst for improving existing STEMs through a method known as *dictionary transfer*.

4.2 Methodology of experimental CS-STEM

This section outlines the methodology for performing a CS-STEM experiment from how to control the probe, what the best strategies are for sampling, and how the data is fed into the inpainting algorithm. There is now a streamlined approach that is used within the Liverpool group, and has been implemented on other microscopes such as the Grand ARM2 "Ruska" at the Rosalind Franklin Institute, and more recently at CNR-IMM in Catania, Italy.

Although the technique which is now used is far more straightforward, prior to this, there was a lot of time spent considering how to go about a live CS-STEM experiment. This section will also present these limitations, and learning outcomes which have ultimately led to the

[1]Flyback time is the amount of time allowed for the beam to return to the start of the next row after completing the previous row acquisition, typically on the order of 300μs

current method.

4.2.1 Controlling the probe

To begin with, it is important to outline the naive assumptions which could be made regarding this method. The first naive assumption is that the probe will go where it is told to go as soon as it is told to do so. As many electron microscopists will know, hysteresis can cause a delay which restricts the response time of the probe to a change in the voltage across the scan coils. As a result, the probe generally lags behind its expected location. For reference, hysteresis is discussed in more detail during section 2.4.3. Following on from this, it is also important to recognise that it is less important to care about the average deviation between the actual and intended probe locations, but more important to consider how the deviation changes across the scan. This is addressed in work by Nicholls *et al.* [137], where different scan trajectories have different standard deviations with respect to the expected probe location.

In order to have control over the probe locations, a scan generator is required. At the Albert Crewe Centre in Liverpool, the JEOL 2100F is equipped with Direct Electron FreeScan system, and more recently a Quantum Detectors scan generator. The role of the scan generator is relatively straightforward; load in a file containing X and Y coordinates for a probe location and apply a stepping voltage to the scan coils which changes the strength of the electromagnetic field at the scan coil plane. This then alters the position of the probe, and the process is continued for that set scan which has been designed.

However, due to the hysteresis effects, the probe is not free to follow an arbitrary path unless the dwell time is sufficiently high. For this reason, designing suitable scanning patterns is vital to reduce the effects.

4.2.2 Designing a suitable scanning pattern

In order to effectively sub-sample scan measurements, there are considerations that must be made which have been highlighted throughout this work. A suitable scanning pattern would consider the effects of scan coil hysteresis, mask incoherence, stability when the dwell time is short (*i.e.*, speed), and the beam damage as a result of the scan being used in order to maximise efficiency. There are certain assumptions which are made when balancing these considerations. For example, consider the average pixel-wise distance between each succesive probe

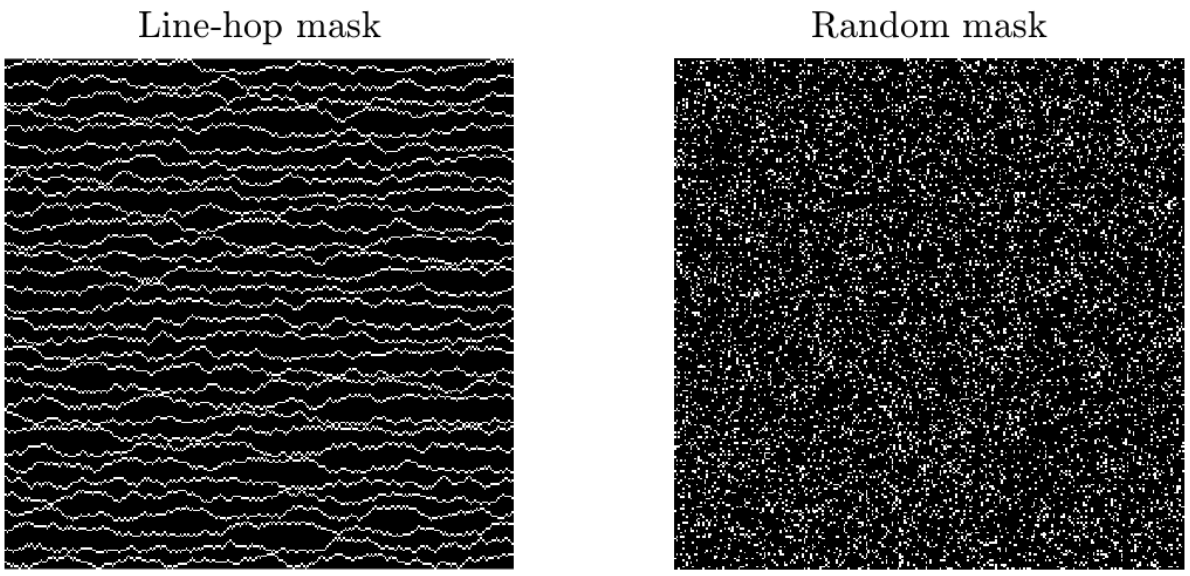

Figure 4.1: **Examples of line-hop and random (UDS) masks.** The line-hop mask provides a pseudo random sampling regime which reduces hysteresis at short dwell times, as well as providing a suitable degree of incoherence.

position. If this is maximised, then the electron flux is distributed as optimally as possible over the field of view, but the hysteresis induced will be high if the dwell time is short. Similarly, a random scan which is highly incoherent will induce hysteresis if the dwell time is relatively short.

There are various types of masks which have been considered for real data acquisition. The ideal mask is random or uniform density sampling (UDS), where each probe has a probability $s_p \in \mathbb{R}^{[0,1]}$ of being sampled. The key is to minimise the distance that the probe travels between successive scan points. As discussed in section 3.3, the line-hop mask is an alternative solution which can compromise the factors above. Fig. 4.1 is a comparison between a line-hop and a UDS mask at equivalent sampling ratios.

A line-hop sampling mask is formed by allowing the probe *hop* either up or down (at random) perpendicular to its forward trajectory (supporting code is given in the appendix A3.1). This reduces the distance between successive scan points whilst also giving sufficient incoherence in the mask. To compare the resulting image quality, a series of simulated CS experiments and reconstructions were performed using both UDS and line-hop masks for an experimental HAADF image of silicon dumbbells with a scan-step of 0.13Å. In this test, the sampling rate is varied from 10% to 50% and the results are averaged over 5 runs. Each test image was reconstructed using a patch size of 12×12. The results are plotted and shown in Fig. 4.2.

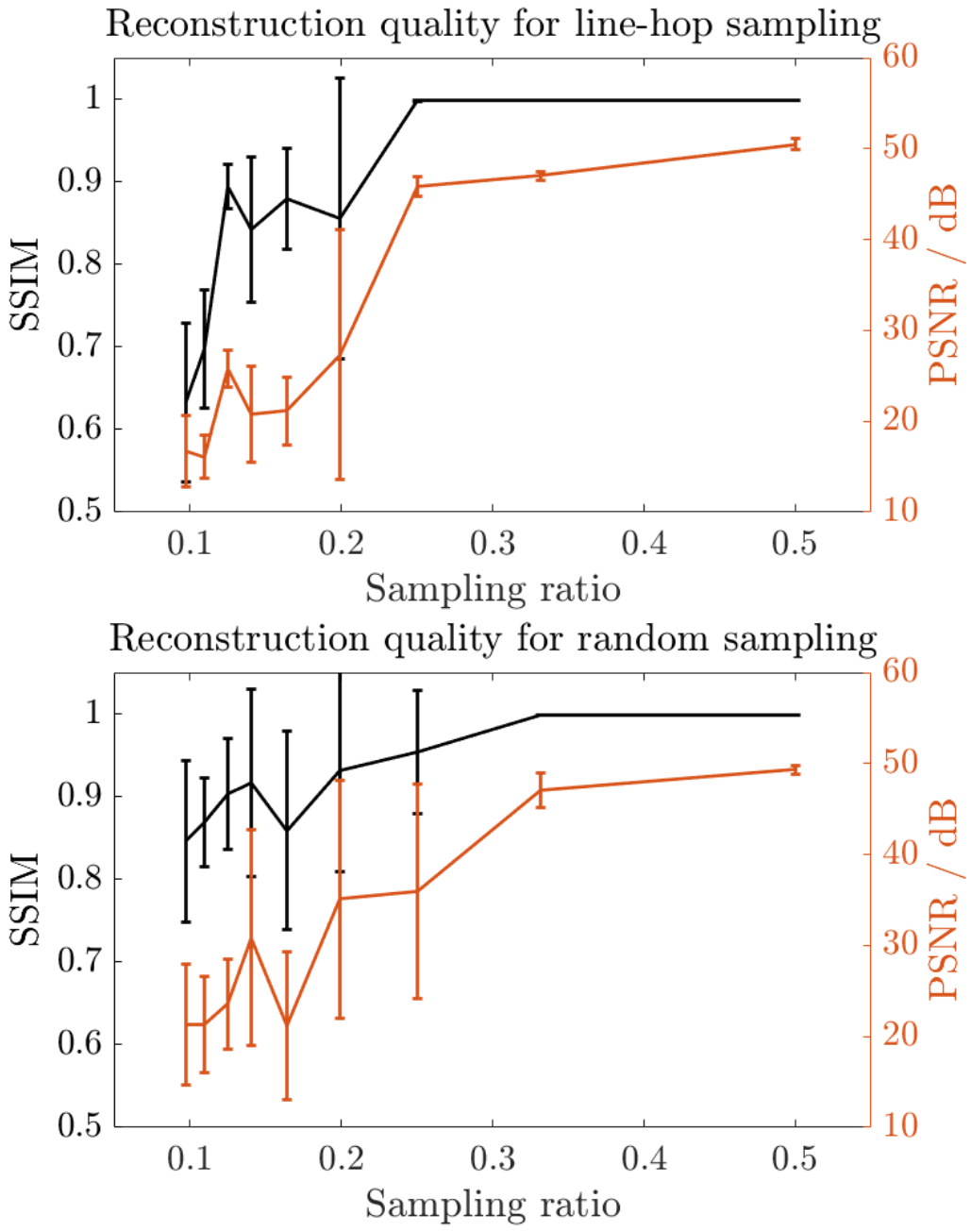

Figure 4.2: **Comparing the reconstruction quality of line-hop versus UDS as a function of sampling rate.** The top figure shows the reconstruction quality for a line-hop mask, which performs well down to 12% sampling. On the other hand, UDS performs much better below 12% indicating that line-hop may not be suitable when lower sampling rates are required.

An important result from this test is that line-hop is limited to a minimum sampling ratio greater than that for random sampling. The results can be improved by increasing patch size for the lower sampling rates, however as was seen in the previous section, increasing patch size could also lead to inconsistent results or results containing artefacts. The other hyper-parameters can be adjusted too to reduce errors, however line-hop is a sufficient solution for CS-STEM with a short dwell time.

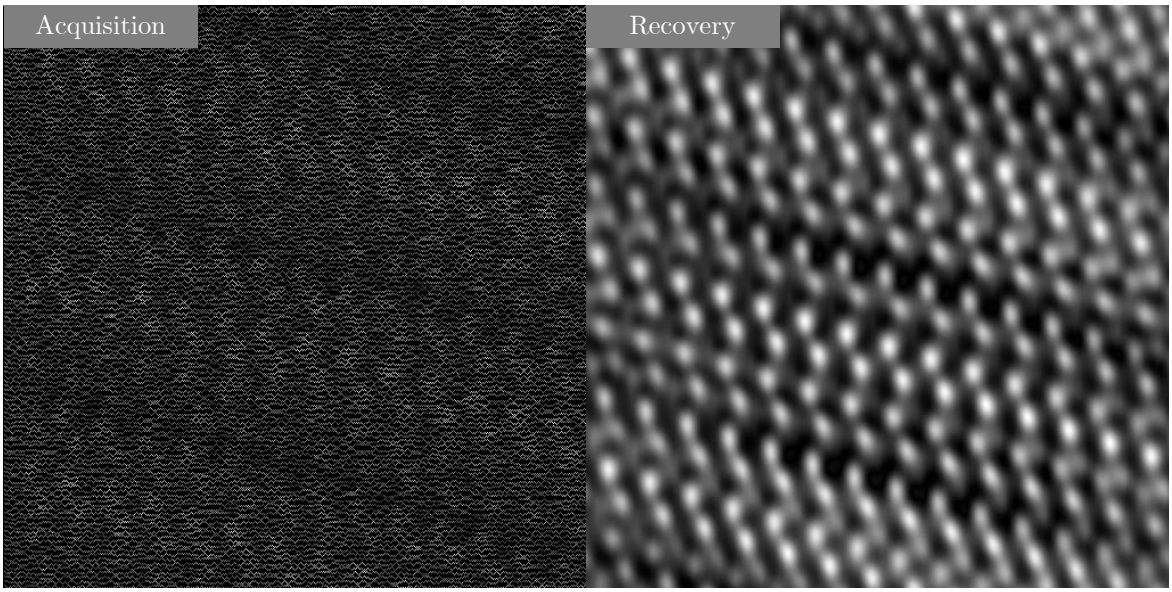

Figure 4.3: **Example of a sub-sampled HAADF image as acquired and inpainted using the Direct Electron system.** The sub-sampled data (left) is acquired by providing the Direct Electron system with a set of X-Y probe coordinates, and then the sub-sampled image is passed through the BPFA to generate a reconstructed image (right). In this case, a 25% line-hop sampling mask is used. The convergence semi-angle used for the experiment here was 25mrad, an acceleration voltage of 200kV, and Scherzer defocus was also used. The scan-step is 0.109Å, which is finer than the Nyquist sampling rate of 0.2508Å.

4.2.3 Extracting the data and inpainting

Once the sub-sampled data has been acquired, it must be extracted and then inpainted using the algorithm of choice. In the case of the Direct Electron FreeScan system, the sub-sampled data must be saved as an image and then loaded in for inpainting. This is a multiple step process; the acquisition is separate from the inpainting. This method is fine for single frame acquisitions, but for live image inpainting during alignment for example, the inpainting software would require direct access to the scan generator. Fig. 4.3 is an example of using the FreeScan system to acquire sub-sampled STEM data, in this case applied to a layered bismuth sample for testing.

The Quantum Detectors (QD) scan generator has since been directly integrated with the SenseAI[2] inpainting and acqusition software to do real-time image inpainting of two-dimensional STEM signals such as HAADF and BF imaging. Through the SenseAI software, the scanning probe can be positioned arbitrarily and the corresponding signal attributed to that probe loca-

[2]SenseAI is a spin-out company from the University of Liverpool formed by myself, Nigel Browning, Daniel Nicholls, and Jack Wells. The SenseAI software was developed primarily by Jack Wells, however the methods are equivalently developed by the four of us.

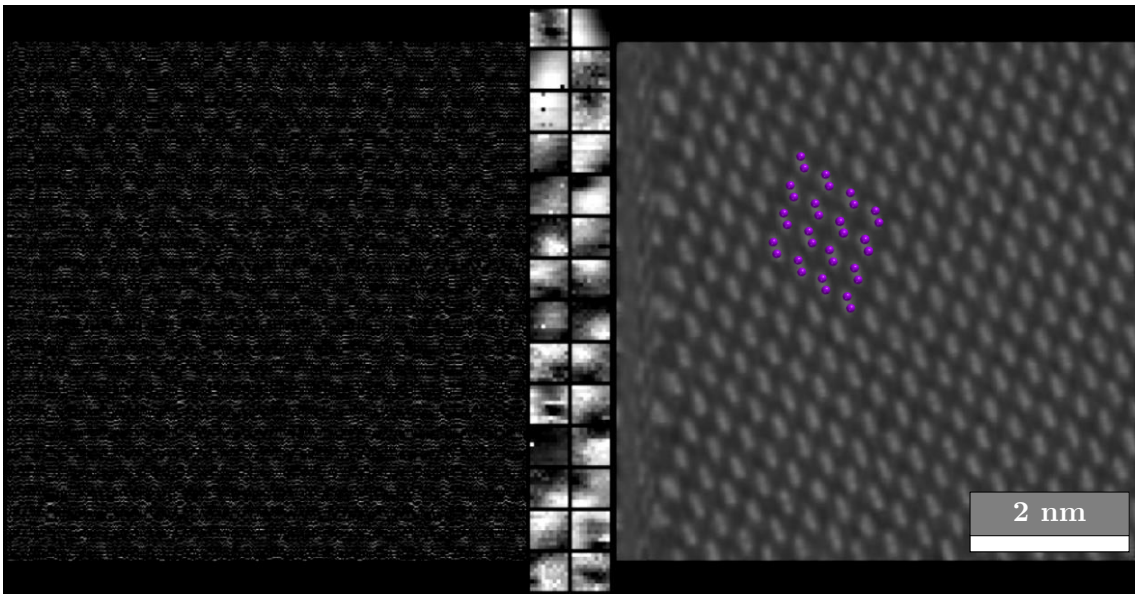

Figure 4.4: **One of the frames from live CS-STEM acquisition and inpainting using the SenseAI software.** The sub-sampled acquisition (left) is acquired using the SenseAI software and a QD scan generator. The data is then inpainted using the BPFA algorithm as shown on the right side of the figure. The dictionary learned from the sub-sampled data is shown in the middle. Credit to Jack Wells for implementation and the RFI for providing the sample. The convergence semi-angle used for the experiment here was 30.8mrad, an acceleration voltage of 300kV, and Scherzer defocus was also used. The scan-step is 0.125Å, which is finer than the Nyquist sampling rate of 0.16Å.

tion. This has allowed for real-time image inpainting of STEM data.

The first experiment using live inpainting of sub-sampled STEM data was performed using the SenseAI software and a QD scan generator in a collaboration with the Rosalind Franklin Institute (RFI). Here, a silicon dataset was imaged to test whether the inpainting can recover the dumbbells with reliability during live acquisition.

As can be seen in Fig. 4.4, the live inpainting successfully recovers the silicon dumbbells. Furthermore, it was possible to focus the image, correct astigmatism, change magnification and find regions of interest using the reconstructed data. This is significant, since the microscope can be used in the same way regardless of whether the acquisition is sub-sampled or fully sampled. This can allow for lower electron fluence during alignment, perhaps allowing for improved image quality of beam sensitive samples.

The same experiment was also performed at CNR-IMM in Catania using a JEOL JEM 200F. Silicon was also used as a test sample to verify that the result was consisted with that observed at RFI. In these experiments, a higher magnification was used to test the stability of inpainting

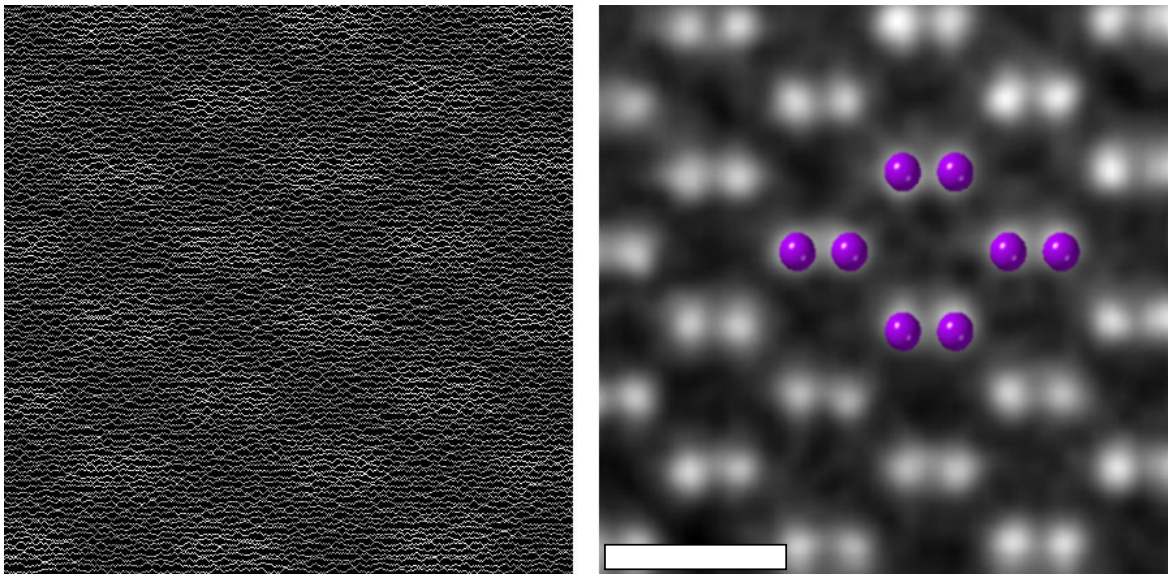

Figure 4.5: **Video frame of high resolution reconstruction of silicon using SenseAI.** The sub-sampled acquisition (left) is acquired using the SenseAI software and a QD scan generator. The data is then inpainted using the BPFA algorithm as shown on the right side of the figure. Scale bar indicates 5Å. Credit to CNR-IMM for providing the sample. The convergence semi-angle used for the experiment here was 30mrad, an acceleration voltage of 200kV, and Scherzer defocus was also used. The scan-step is 0.031Å, which is finer than the Nyquist sampling rate of 0.21Å.

to fluctuations, although this was for observation only.

The results presented in Fig. 4.5 show a high resolution frame taken from a live reconstruction feed. The image shows the Si {004} dummbells resolved at 0.136nm. This result is in agreement with that observed at the RFI and shows that the deployment of sub-sampling is robust across different (JEOL) microscopes. It also shows that the stability of the system, regardless of sub-sampling.

4.3 Improving resolution through dictionary transfer

One important question which is often asked by those enquiring about how the BPFA algorithm generates a dictionary is whether the learned dictionary is the optimal dictionary for the inpainting task. As the algorithm learns a new dictionary or updates the dictionary for each instance, this generally comes at the cost of speed. In high-resolution Z-contrast imaging, the images are effectively just white balls on a black background, perhaps containing defects or something more exotic. In many cases, all the images of atoms should produce near identical dictionaries if the patch shape is to encompass just one atom.

In this section, a method known as *dictionary transfer* is discussed, where an optimal seeding image is used to train a dictionary which is then transferred to inpaint experimental STEM data.

4.3.1 Finding the right seed

As the background chapter presented, the theory of STEM is well understood. STEM simulations are also a very well researched topic within the field, and there are various algorithms available which can calculate expected contrast in STEM imaging modes [168–171]. In many cases, these simulations are used in analysis of STEM images to verify observations, however could these simulations be used *during* experimental acquisition to seed recovery? Chapter 5 will go into more detail regarding the specifics of STEM simulation, here it is assumed that they are sufficiently accurate at estimating contrast in STEM imaging.

The benefit of using a simulated image is that they are inherently noise-free (*i.e.*, infinite dose), can be constructed to an arbitrary scan-step, and the parameters can be set such that the image is free of aberrations. Now although this may not be the case in experimental acquisition since residual aberrations generally cannot be avoided, it is not unreasonable to expect that simulations could effectively correct these aberrations simply by providing the inpainting algorithm with the dictionary generated from an aberration free image.

For example, consider the two dictionaries generated from an experimental image, and the other from a simulated image in Fig. 4.6. It is clear that the dictionary generated from the simulated data would be free of noise, whereas the noise in the experimental data has to be modelled by the dictionary atoms. In a similar way, an experimental image (perhaps subsampled) may contain aberrations, drift artefacts, or even be slightly off axis. The dictionary learned from that experimental data would itself contain similar artefacts found within the image, and would inpaint those as expected. Now this isn't such a bad thing and should be encouraged to ensure that the operator can make the required adjustments. However, it would also be useful to use the simulated dictionary to drive recovery as well, since that recovery could be the optimal experiment.

In order to test if this is effective for inpainting experimental data, an experimentally acquired Z-contrast image of silicon dumbbells was used as a reference image, and a simulated image of the same structure with the same scan step as a transfer source. The reference data

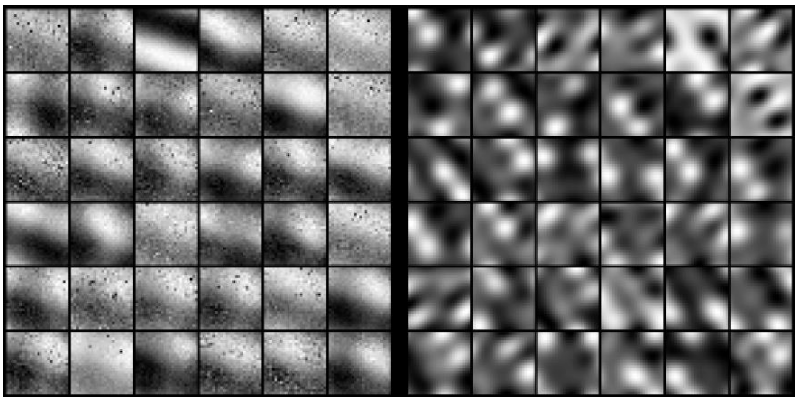

Figure 4.6: **Comparison between a dictionary learned from an experimental image and one learned from a simulated image.** As can be seen, the experimental image dictionary (left) is slightly nosier than the dictionary learned from the simulated image (right). This implies that the reconstruction using the simulated image should be noiseless if the dictionary is appropriate for recovery.

was used to simulate experimental CS by sub-sampling to 3%, and then inpainted using a self-learned dictionary as well as a dictionary learned from the transfer source. The results are presented in Fig. 4.7.

As Fig. 4.7 shows, the transfer of the dictionary from the simulated image provides a much better reconstruction than the self-learned dictionary. The reason for this is that the dictionary learned from the transfer source is easier to learn than the self-learned dictionary, providing a much better basis for the input to be reconstructed. It is important that the simulated image source matches the orientation of the input if the patch size contains more than just one atom. This could lead to artefacts if the matching is not done appropriately for inpainting larger structures in the image.

4.3.2 Applying the method

The method was tested live in the inpainting of fully sampled silicon dumbbells, the same data which is shown in Fig. 4.4 without sub-sampling. As can be seen in Fig. 4.8, the use of a dictionary learned from a simulated image improves the resolution of the result. The reason for this is due to the fact that the simulated data is clean, on-axis, and free of astigmatism. It could be that the resolution in the self-learned case is hampered due to contamination, but the transfer from the simulated data is able to correlate the correct features.

Another option is that the result from transfer is simply the result of an optimal experiment, however given the properties of the algorithm used, the solution given is the one with the least

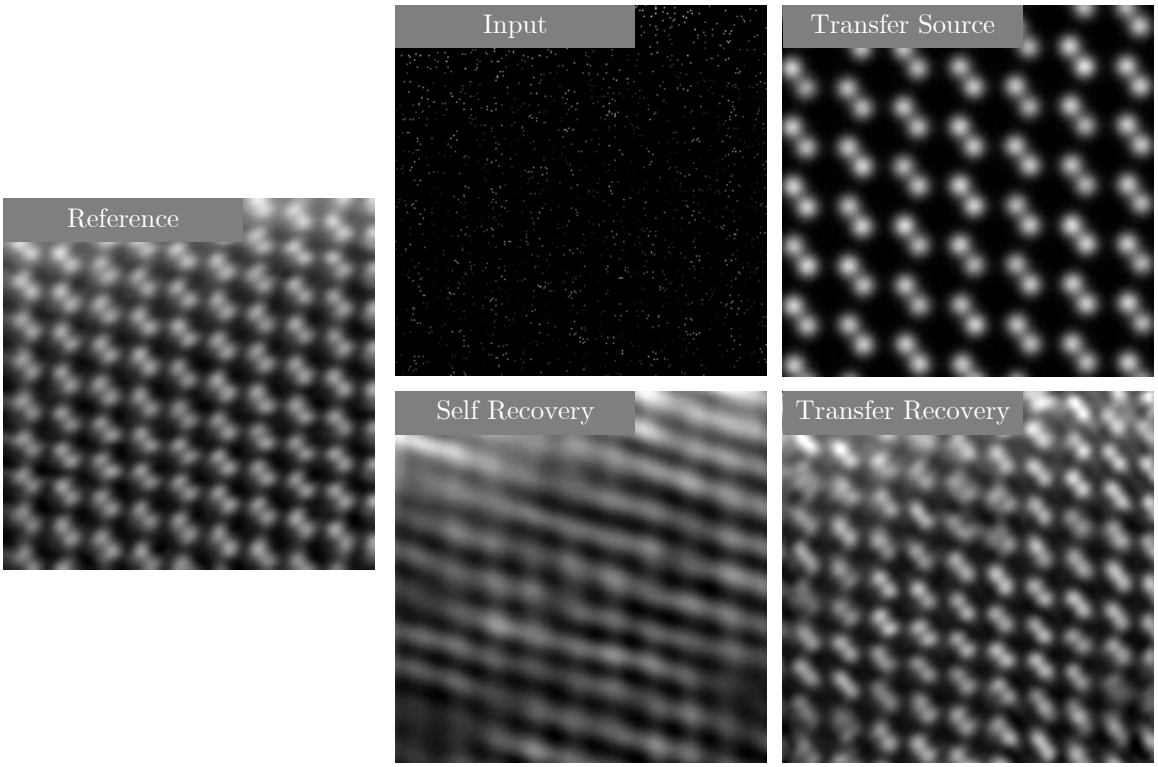

Figure 4.7: **Testing the transfer of a dictionary from a simulated image to an experimentally sub-sampled image.** The input image is a 3% UDS sampled version of the reference. The results show that transferring the dictionary from a simulated image provides a better reconstruction than the self-learned dictionary. Reference image courtesy of Dr Mounib Bahri. The convergence semi-angle used for the experiment here was 25mrad, an acceleration voltage of 200kV, and Scherzer defocus was also used. The scan-step is 0.125Å, which is finer than the Nyquist sampling rate of 0.2508Å.

error with respect to the observed data. It is also noted that the data was aligned to be on-axis using the Kikuchi bands observed in the Ronchigram, and the defocus/astigmatism corrected for to optimal conditions given the contaminating sample.

4.4 Conclusions

This chapter has presented results of compressive sensing applied to 2-D imaging modes in STEM, as well as demonstrating practical data acquisition using three separate systems. Furthermore, live CS-STEM has been presented as well as a novel method to improve image resolution using dictionary transfer from simulated STEM images.

As was discussed in section 2.4, beam damage, instability, and contamination may be reduced through a sub-sampled scan. This is due to fewer electrons striking the sample for each frame and better dose distribution. It is this distribution of dose that can reduce the diffusion

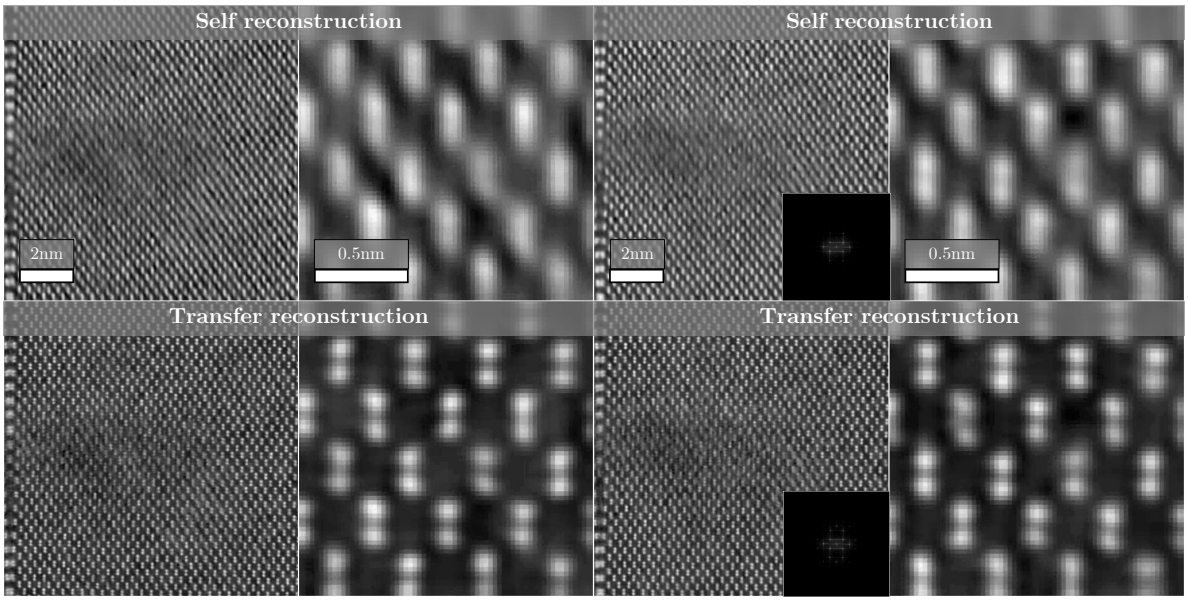

Figure 4.8: **Testing the transfer of a dictionary from a simulated image to an experimentally acquired image.** Two separate frames from experimental data which has been inpainted using a self-learned dictionary and a dictionary learned from a simulated image of the same sample. The results indicate higher resolution due to dictionary transfer from the simulated image, as evidenced by the increased intensity of higher order reflections in the overlaid power spectra. The convergence semi-angle used for the experiment here was 30.8mrad, an acceleration voltage of 300kV, and Scherzer defocus was also used. The scan-step is 0.2267Å, which is courser than the Nyquist sampling rate of 0.16Å.

of radicals between successive scan points.

In the case of imaging extremely sensitive samples, a delay can be added between successive frames, allowing the sample to relax and possibly recombine. In this case, there would be no speed increase, but having flexibility to manipulate the scan coils allows for complete control over the electron fluence.

The next steps along this research are to show that this method is robust when applied to practical live acquisition of defective samples, as well as beam sensitive materials. These complex samples will provide a test beyond that which is presented here.

5 | Applying Compressed Sensing Methods to STEM Simulations

5.1 Overview of STEM Simulations

To fully identify the atomic scale structure and composition of complex materials, interfaces and defects from experimental images, it is essential to use simulations to capture all the experimental parameters involved. The accurate simulation of scanning transmission electron microscopy (STEM) images [101, 172–174] is a computationally expensive task due to the nature of the scattering and detection process. The most common method used to obtain simulations is the multislice method [168–171, 175–195]. In the multislice approach, the 3-dimensional atomic potential of a sample is first approximated by a series of 2-dimensional (2D) infinitely thin potential slices, $V_{\mathrm{s}}^{\mathrm{2D}}$, where s is the index of a slice and $\vec{r}$ denotes a location in real space coordinates. For every probe position, the multislice approach involves the following steps. First, the incident wavefunction of the electron beam $\psi^{\mathrm{i}}_{(s+1)}(\vec{r})$ for a slice $s+1$, is computed from the exit wavefunction $\psi^{\mathrm{e}}_{(s)}(\vec{r})$ of the previous slice s and the atomic potential of that layer:

$$\psi^{\mathrm{i}}_{(s+1)}(\vec{r}) = \psi^{\mathrm{e}}_{(s)}(\vec{r}) \exp\left[j\sigma V_{\mathrm{s}}^{\mathrm{2D}}(\vec{r})\right] \ , \tag{1}$$

where σ denotes the beam-specimen interaction constant. Next, the exit wavefunction $\psi^{\mathrm{e}}_{(s+1)}(\vec{r})$ for slice $s+1$ is computed by propagating the incident wavefunction $\psi^{\mathrm{i}}_{(s+1)}(\vec{r})$ using Fresnel propagation model:

$$\psi^{\mathrm{e}}_{(s+1)}(\vec{r}) = \mathcal{F}^{-1}\left[\mathcal{F}\left[\psi^{\mathrm{i}}_{(s+1)}(\vec{r})\right] \exp\left(-j\pi\lambda|\vec{q}|^2 t\right)\right] \tag{2}$$

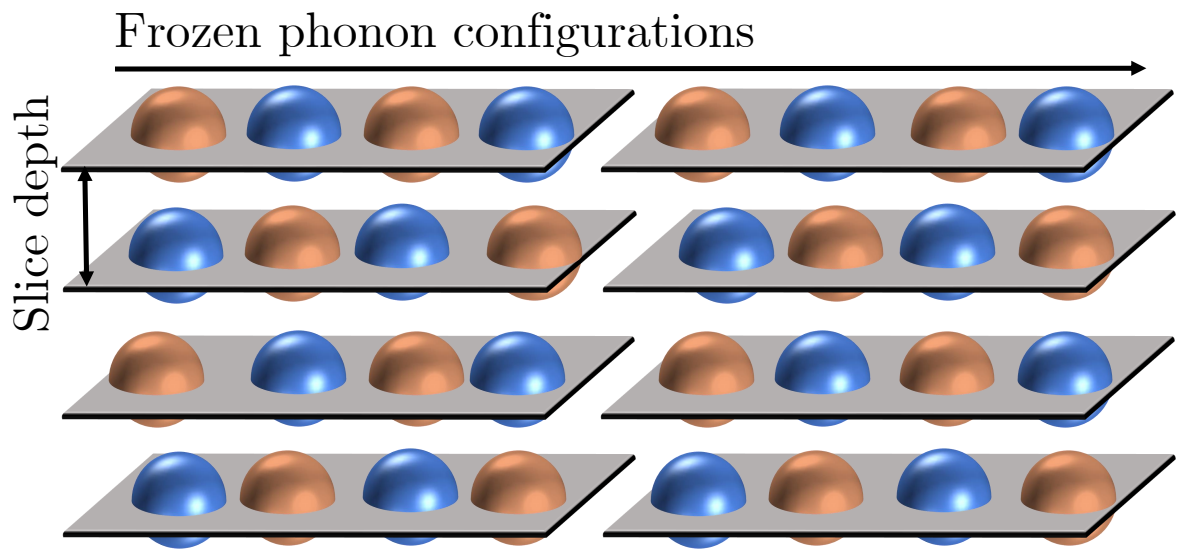

Figure 5.1: **Diagram explaining the frozen phonon model.** Each atom within the sample is slightly altered from its equilibrium position for each frozen phonon configuration and a multislice calculation is performed. The resulting simulation is an average over all the resulting multislice calculations.

where λ and t are the electron wavelength and slice depth, respectively [168]. Once the sample has been propagated through the sample, the exit probe is then determined in Fourier space. This must be repeated for all required STEM probe locations; hence the computation time of the STEM simulation is limited by computing multiple Fast Fourier Transforms (FFTs).

A more recent development through an algorithm called plane-wave reciprocal-space interpolated scattering matrix (PRISM) [168, 169] has led to much faster STEM simulations. This algorithm forms a basis set of plane waves (based on an interpolation factor where a higher interpolation factor introduces more error but faster simulation) which are independently propagated through the sample (in a multislice approach). After propagation, the plane waves form a scattering matrix, and can be superimposed with appropriate weighting to form a close approximation to the exit probe had it been calculated using the multislice approach. The set of plane waves are essentially shared between all probe locations, meaning that the multislice step must be completed only for the basis set of plane-waves, greatly reducing the computational load. This leads to significant speed up in computation times, with minimal loss of information compared to the traditional multislice method.

To account for all possible electron-phonon interactions, the frozen phonon approximation

(FPA) is also often used [170, 171, 191], where the intensity of multiple multislice calculations (layers) is averaged over various frozen phonon configurations (FPC), as depicted in Fig. 5.1. In STEM imaging, the electron probe is situated on a certain location for a certain amount of time, this is known as the dwell time. In STEM simulations, as described in equation 2, the potential is static or constant in time, whereas in practice it is evolving with time. It is safe to assume that given the *(i)* thickness of the sample ($\sim$ 100nm) and *(ii)* velocity of the electron ($\sim 0.78\times$ speed of light with a 300kV accelerating voltage) that it will only 'see' one instance of the potential, however if multiple electrons interact with the sample over the given dwell time this will directly effect observations on the detector. Since the atomic vibrations will scale with their temperature according to [196],

$$E \approx \frac{k_{\mathrm{B}}T}{2} \ , \tag{3}$$

which has an oscillation frequency $<<$ the reciprocal of the time the electron spends within the sample. This is why *snapshots* are an appropriate analogy.

Correct modelling of thermal diffuse scattering (TDS) is vital to calculate probability distributions arising from inelastically scattering electrons. In an ideal case, this would be modelled through quantum mechanics as in the work by Forbes *et al.*[18]. This model is described above in Section 2.2.3 when considering the change of electron wave-function when interacting with a sample. However, in the majority of cases the FPA is sufficient to represent thermal diffuse scattering.

Another consideration when performing STEM simulation, is to consider the real-space sampling. Similar to the Feynman path integral shown in Section 2.2.1, a continuous source (*i.e.*, the wave-function of an electron) is approximated into a discrete set of samples. The spacing between these samples (known as the real-space sampling) determines the resolution of the estimated sample potential and wave-function. This is the allowed positions for $\vec{r}$ as shown in Equation. 2. It is important that this space is sampled sufficiently in order to accurately represent the scattering, but also important to not oversample as this will increase calculation times.

Taking all the physics of these interactions into account at every beam location means the computation time of multislice STEM simulation scales with the number of required probe

locations, the number of FPCs, and the number of reciprocal space sampling points. A typical multislice STEM simulation (for a sample 10nm in thickness, 256×256 grid locations, using 20 FPCs and a real space sampling of 0.04 Å) operating on a system equipped with a graphics processing unit (GPU) for faster calculation can take of order hours, and even longer if a GPU is not available [168]. The PRISM method can perform the same simulation in significantly less time depending on the interpolation factor used (typically on the order of minutes or potentially seconds with a larger interpolation factor) at the expense of accuracy. For calculation of structures with a crystalline periodicity, a tiling method can be used to increase the number of effective probe locations. Following this, it is possible to calculate the minimum sampling frequency (*i.e.*, scan-step or scan-pitch) Δ_{p} (in m) based on Shannon-Nyquist sampling theorem.

For a one-dimensional signal which is continuous in space (or time, for here only the spatial continuity is required) and contains no spatial frequency higher than $k_{\max}$ (in units of m^{-1}), the signal can be completely determined from a discrete sampling set where the samples are spaced less than $1/(2k_{\max})$ meters apart. This is then extended into two dimensions by considering two orthogonal one-dimensional signals.

In the case of STEM simulations, the maximum spatial frequency contained in an image is determined by the probe convergence semi-angle α and electron wavelength λ, as per the aperture function dampening. The maximum spatial frequency is therefore $k_{\max} = 2\alpha/\lambda$. Following on from the Shannon-Nyquist sampling theorem, and assuming that the signal/image is continuous in space, then the minimum sampling frequency is given as,

$$\Delta_{\mathrm{p}} < \frac{\lambda}{4\alpha} \ , \tag{4}$$

Furthermore, the minimum necessary scan positions required (for structures satisfying periodic boundary conditions) is given by,

$$M_x M_y > \frac{4(a \times b)\alpha}{\lambda}, \tag{5}$$

where $(a \times b)$ is the size of the orthogonal scan area and $M_x \times M_y$ is the scan dimension (pixels), which agrees with the findings of Dwyer [195].

The resulting simulation can then be interpolated to an arbitrary size with correct interpolation parameters; Dwyer [195] suggests a sinc function interpolation or a Fourier interpolation, noting that the interpolation will only approximate if the structure does not contain periodic boundary conditions.

Following this, it may therefore be more beneficial to employ a sparse sampling approach, especially if the sample is non-periodic (*i.e.*, a grain boundary or defect containing sample). Furthermore, sparse-sampling can be employed if the sample is periodic and sampled according to Equation 5, as long as a suitable recovery algorithm is used.

5.2 Methods for compressed STEM simulations

Optimisation of STEM simulations is important for real-time analysis of specimen, whilst also retaining the information which is required to make accurate determinations of sample properties. As discussed, STEM simulations are computationally expensive due to the nature of the calculation methods. Although there exists alternative algorithms, *i.e.*, PRISM, they can often be limited by computer memory (RAM) or GPU memory (VRAM). Given the success of compressive sensing for experimental STEM acquisition, it was a natural question to ask whether the same ideas could be applied to STEM simulation. This section outlines those ideas, focusing on three key aspects of STEM simulation– probe sub-sampling, reciprocal space sampling, and the frozen phonon model.

This section is supported by two peer-reviewed journal articles, however the text is modified for the purpose of continuity throughout this document.

5.2.1 Probe sub-sampling

In STEM simulations, each probe location is independent from the other, *i.e.*, the calculation of the exit wave-function has no bearing upon the same calculation at a different probe location. This allows for STEM simulations to be computationally parallelised, hence calculation times can be significantly reduced by support of specialist hardware such as GPUs. By the same reasoning, it also means that any arbitrary sub-set of probe locations can be calculated without influencing the result at the acquired locations. The abTEM package allows for any arbitrary scanning regime to be implemented, hence sub-sampled scanning patterns are considered here.

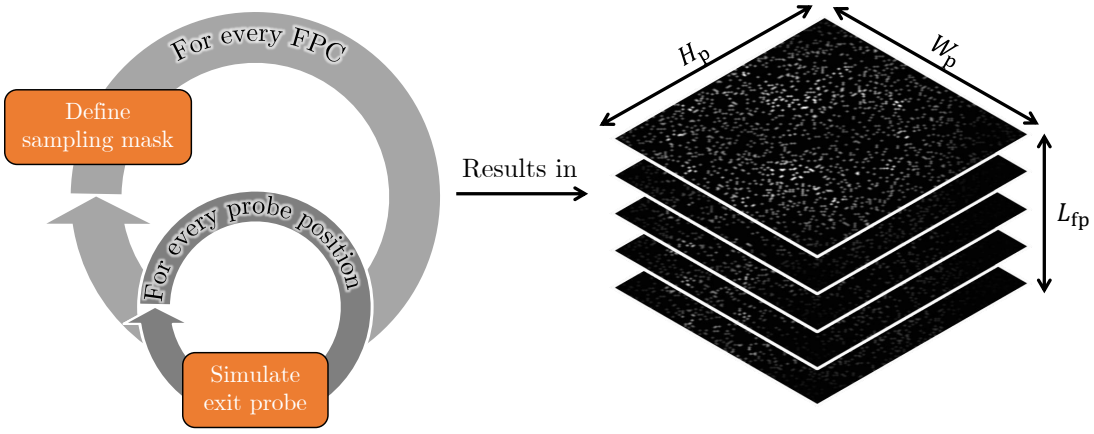

Figure 5.2: **Visualisation for acquisition of sub-sampled STEM simulations.** As the acquisition model in Eq. 6 describes, for each frozen phonon layer a sampling mask is defined, and the exit probe is simulated for each probe location given by that mask. This results in a sub-sampled three dimensional data-cube.

As previously discussed in section 5, a STEM simulation is a 3-D object which is collapsed to a 2-D object by taking an average of the data across the third dimension (*i.e.*, the frozen phonon model). Assume an electron probe system scanning a regular grid of H_p and W_p locations in the vertical and horizontal axis, respectively, collected in a probe locations set $\Omega_p := \{1, \cdots, H_p\} \times \{1, \cdots, W_p\}$. Let $\boldsymbol{r}_p := (r_p^h, r_p^w) \in \Omega_p$ denote the coordinates of a probe location, and the total number of probe locations given as $N_p = H_p W_p$. Additionally, let $\Omega_{fp} := \{1, \cdots, L_{fp}\}$ be the set of all frozen phonon layers where L_{fp} is the number of frozen phonon configurations/layers, and $\boldsymbol{l}_{fp} \in \Omega_{fp}$ denotes the layer index. Let $\mathcal{X} \in \mathbb{R}^{H_p \times W_p \times L_{fp}}$ be the discretised 3-D representation of fully sampled simulated STEM data; and $\mathcal{X}(\boldsymbol{r}_p, \boldsymbol{l}_{fp})$ be the simulated STEM data observed at probe location $\boldsymbol{r}_p$ and frozen phonon layer $\boldsymbol{l}_{fp}$. Each frozen phonon layer is therefore $\boldsymbol{X}^l_{\boldsymbol{l}_{fp}} := \mathcal{X}(\cdot, \boldsymbol{l}_{fp}) \in \mathbb{R}^{H_p \times W_p}$, with the final simulation being given as $\boldsymbol{X} := \frac{1}{L_{fp}} \sum_{i=1}^{L_{fp}} \boldsymbol{X}^l_i \in \mathbb{R}^{H_p \times W_p}$.

Each element of $\mathcal{X}$ is an independent step calculated using the multislice approximation (ignoring the PRISM method for now). This means that the computation time for a fully sampled STEM simulation is proportional to $|\Omega_p| \times |\Omega_{fp}|$ for a given reciprocal space sampling, slice depth, and sample thickness. This implies that the calculation time can be reduced if the size of the sampling sets Ω_p and Ω_{fp} are reduced.

Sensing model

A generalised sub-sampling strategy for calculating STEM simulations is now introduced. This is done by calculating $M_{\mathrm{p}}^{\mathrm{l}} \ll N_{\mathrm{p}}$ probe locations in the sub-sampling set $\Omega_{\mathrm{l}} \subset \Omega_{\mathrm{p}}$ which is equivalent to sub-sampling each of the frozen phonon layers independently[1]. This defines our acquisition model as,

$$\boldsymbol{Y}_{\boldsymbol{l}_{\mathrm{fp}}}^{\mathrm{l}} = \boldsymbol{P}_{\Omega_{\mathrm{l}}} \boldsymbol{X}_{\boldsymbol{l}_{\mathrm{fp}}}^{\mathrm{l}} + \boldsymbol{N}_{\boldsymbol{l}_{\mathrm{fp}}}^{\mathrm{l}} \in \mathbb{R}^{H_{\mathrm{p}} \times W_{\mathrm{p}}}, \quad \text{for } \boldsymbol{l}_{\mathrm{fp}} \in \Omega_{\mathrm{fp}}, \tag{6}$$

where $\boldsymbol{Y}_{\boldsymbol{l}_{\mathrm{fp}}}^{\mathrm{l}}$ is the sub-sampled measurements for frozen phonon layer $\boldsymbol{l}_{\mathrm{fp}}$ and $\boldsymbol{P}_{\Omega_{\mathrm{l}}}$ is a mask operator with $(\boldsymbol{P}_{\Omega_{\mathrm{l}}}(\boldsymbol{U}))_{(i,j)} = \boldsymbol{U}_{(i,j)}$ if $(i,j) \in \Omega_{\mathrm{l}}$ and $(\boldsymbol{P}_{\Omega}(\boldsymbol{U}))_{(i,j)} = 0$ otherwise, and $\boldsymbol{N}_{\boldsymbol{l}_{\mathrm{fp}}}$ is an additive noise. This is visualised and demonstrated in Fig. 5.2.

Targeted sampling of STEM simulations

Given that an atomic coordinates must be provided for the simulation, there is prior knowledge for the expected positions where there may be contrast within the final simulation. This means that a sampling operator (*i.e.*, mask) can be designed which specifically targets certain properties, such as intensity or intensity-gradient (or simply gradient). To design such a mask, a map which indicates atomic locations can be generated, modified (based on intensity or gradient) and then sampled from where the likelihood of sampling is proportional to the intensity of the modified map. Examples of different maps are given in Fig. 5.3.

The targeted sampling map is updated with a targeted sampling factor $F \in \{0, 1\}$ which defines how targeted it should be. That is, if $F = 1$ then the mask is purely targeted, and if $F = 0$ then the mask reverts to UDS. It is important to have control of this parameter such that the data is sufficiently sampled across the entire field of view.

To test the effectiveness of using Z-number intensity targeted sampling over random sampling (*i.e.*, uniform density sampling, UDS), various simulations are performed using different mask types and compared in Fig. 5.4.

[1] It is also possible to consider the case where $\Omega \subset \Omega_{\mathrm{p}}$ is a common mask shared across all frozen phonon layers, however only the general case is considered here.

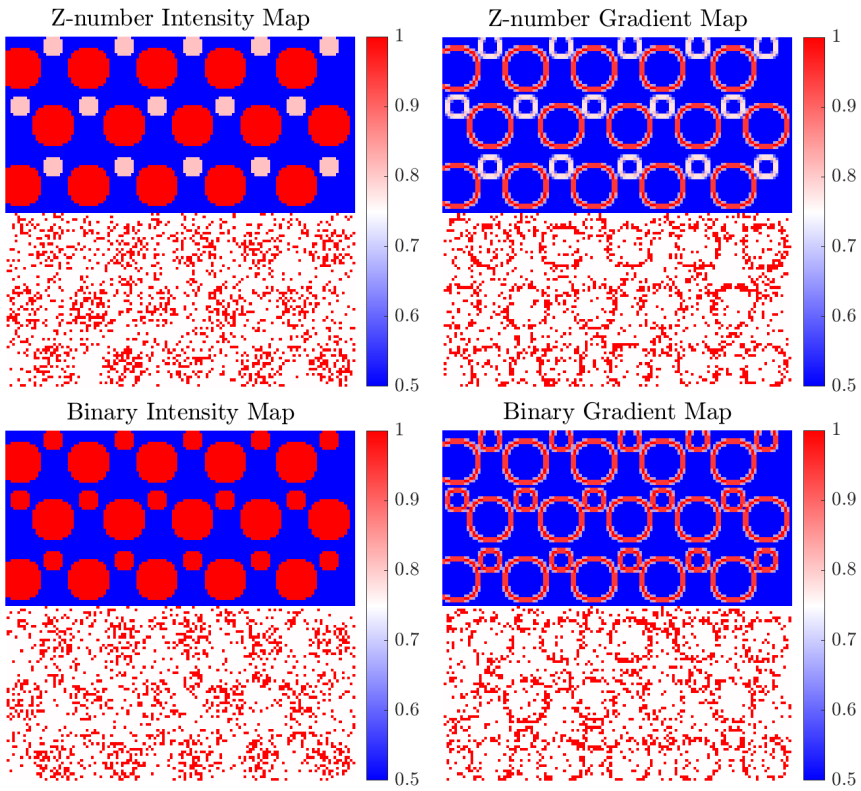

Figure 5.3: **Examples of different targeted sampling maps which the mask (overlaid) is drawn from.** By using prior knowledge of the sample, it is possible to design custom sampling masks which can optimise the recovery at low sampling rates. The examples shown here a subset of possible methods where the sampling is based on the intensity or the gradient of the map. The targeted sampling factor, F, is 0.5. The radii are determined from the ionic radii, although this is somewhat arbitrary and could be altered top be based on the bonding type, or perhaps the probe radius.

5.2.2 Optimising the frozen phonon model

The second step towards improving the efficiency of STEM simulation is to optimise how the frozen phonon model can be adapted through a targeted sampling method. The frozen phonon model is used to account for thermal diffuse scattering within the sample, by taking snapshots of the sample at some given time where the atom locations are slightly displaced from their equilibrium position depending on the Debye Waller factor (DWF) of the atom [170]. Each snapshot of atom positions is known as a FPC and as more configurations are considered, generally the more accurate the simulation is, given the final simulation is the average of simulations over all configurations (Fig. 5.1).

In practice, beyond some number of configurations (depending on sample type, resolution, thickness and spatial density), the improvement in simulation quality diminishes, however the

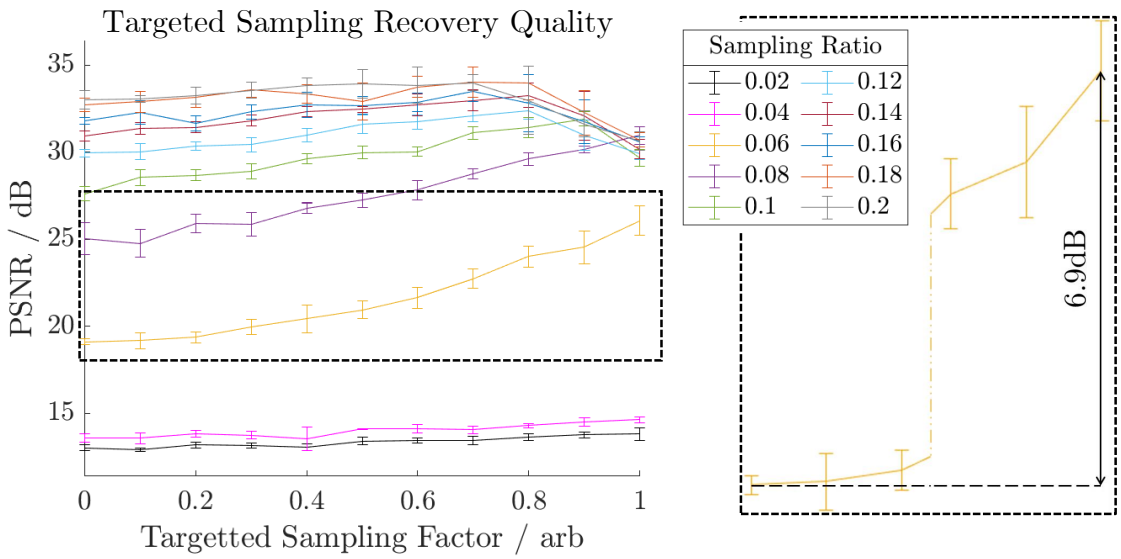

Figure 5.4: **Z-number intensity targeted sampling at different sampling ratios and targeted sampling factors versus reconstruction quality.** Targeted sampling can dramatically improve the quality of image reconstruction, especially at lower sampling ratios. This is highlighted for 6% on the right-hand side of the figure where a near 7dB improvement is seen in PSNR. The error bars are the standard deviation taken over 5 Monte-Carlo runs.

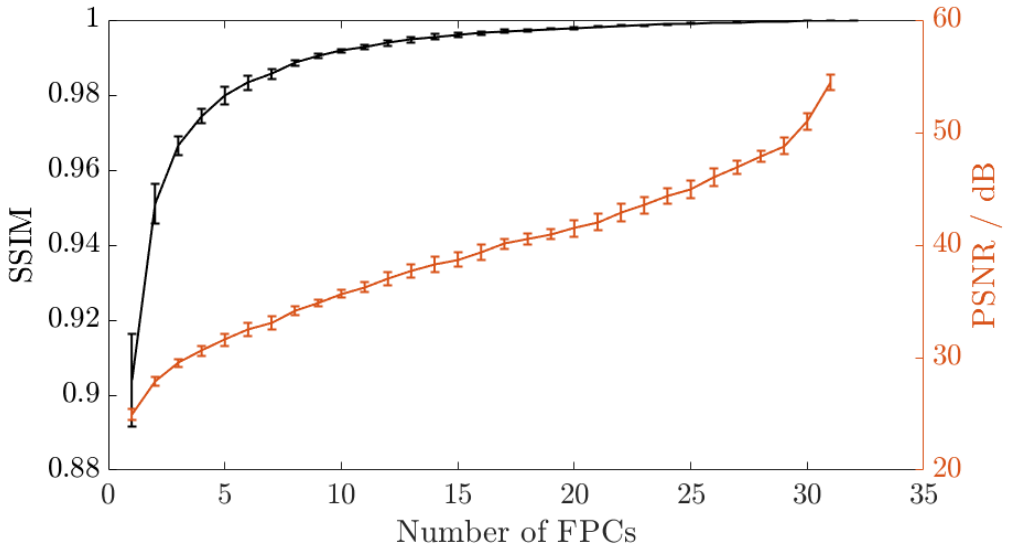

Figure 5.5: **Quality of simulation with respect to the number of frozen phonon configurations used.** Mutlislice simulations performed using different numbers of frozen phonon configurations. The calculation time scales linearly with respect to the number of configurations, but the quality improvement diminishes at around 10 configurations. The reference is the simulations performed with 32 configurations. Other parameters are detailed in table 5.1.

computation time scales linearly. Therefore, for the purpose of speeding up simulations, it is better to limit the number of configurations to the point where the improvement begins to diminish, as Fig 5.5 shows for a bulk MoS_2 sample.

Parameter	Value
Scan step (Å)	0.2
Real space sampling (Å)	0.1
Accelerating voltage (kV)	60
Slice depth (Å)	0.5
Sample thickness (Å)	$\sim$4

Table 5.1: **Parameters for simulations of MoS_2 with varying frozen phonon configurations.** The values for each parameter corresponding to Fig. 5.5.

As an example, consider sampling at a location where there is no atom. Here the number of configurations used make an insignificant contribution to the intensity of the pixel at that location, and therefore to continue sampling this position would be time inefficient. Instead it would be better to sample at atom sites more frequently where the frozen phonon approximation has more effect. This can be achieved by using a different targeted mask for each FPC rather than using the same mask each time. This will also increase the net sampling of the final simulation as the pixel values are averaged in the final step (Fig. 5.2), as well as reducing the total sampling ratio required for each independent configuration.

These two methods, when used in conjunction, can yield a final simulation that still includes the frozen phonon approximation, but decreases the simulation run-time significantly.

5.2.3 Real space sampling optimisation

Selecting an optimal real space sampling can be difficult, since it's not inherently obvious what this value should be without understanding the mathematical basis. The real space sampling $\Delta_\mathrm{r} \in \mathbb{R}^+$ refers to the resolution of the potential in real space, or the maximum scattering angle in real space. This value should be well selected to ensure that the correct contrast is estimated in simulation. The calculation time is approximately proportional to $1/(\Delta_\mathrm{r})^2$, and quality decreases as Δ_r increases.

However it is possible to estimate the optimal value based on parameters of the simulation set-up. The reciprocal space sampling is calculated according to,

$$\Delta_\mathrm{r} = \min\left[\frac{l_x}{n_x}, \frac{l_y}{n_y}\right] \ , \tag{7}$$

where $l_{x,y} \in \mathbb{R}^+$ is the size of the imported sample in the x and y dimensions, with units Å, and $n_{x,y} \in \mathbb{N}_1$ is the number of grid points which the sample potential is partitioned into. The task

is then to find the smallest values for n_x, n_y. To do this, consider the largest scattering angle which requires calculation. For high-angle Z-contrast simulation, this would be the outer angle of that detector $\theta_o \in \mathbb{R}^+$ (mrad). Given that the scattering in HAADF is considered incoherent, scattering beyond this outer angle should not impact the scattering within the detector range significantly. The maximum scattering vector for collection, $k_{max,c} \in \mathbb{R}^+$ is given as,

$$k_{max,c} = \frac{\theta_o \times 10^{-3}}{\lambda} \ , \tag{8}$$

where $\lambda \in \mathbb{R}^+$ is the wavelength with units Å. The values for n_x, n_y are then given as,

$$n_{x,y} = 2k_{max,c}l_{x,y} \ . \tag{9}$$

Further optimisations can be made such as finding the practical numbers closest to the values of $n_{x,y}$, as is done in the MULTEM code [171]. The factor of 2 arises due to symmetry. The optimal real space sampling is therefore,

$$\Delta_r = \frac{\lambda}{2\theta_o \times 10^{-3}} \ , \tag{10}$$

assuming that $n_{x,y}$ have been calculated as practical numbers.

As a demonstration, Fig. 5.6 is a measure of simulation quality with respect to real space sampling.

5.2.4 Conclusions of methods

The combination of sub-sampling, frozen phonon optimisation, and real space sampling optimisation will generate the fastest method for calculating STEM simualtions using both the PRISM and multislice methods. The important aspect of all this is that the quality of a simulated image is generally superior to experimental data. The lack of drift, various noises, increased stability, and exclusion of damage allows intended contrast to be calculated, whilst experimental images suffer these drawbacks. Therefore, does a STEM simulation require perfection if the experimental data is inherently imperfect?

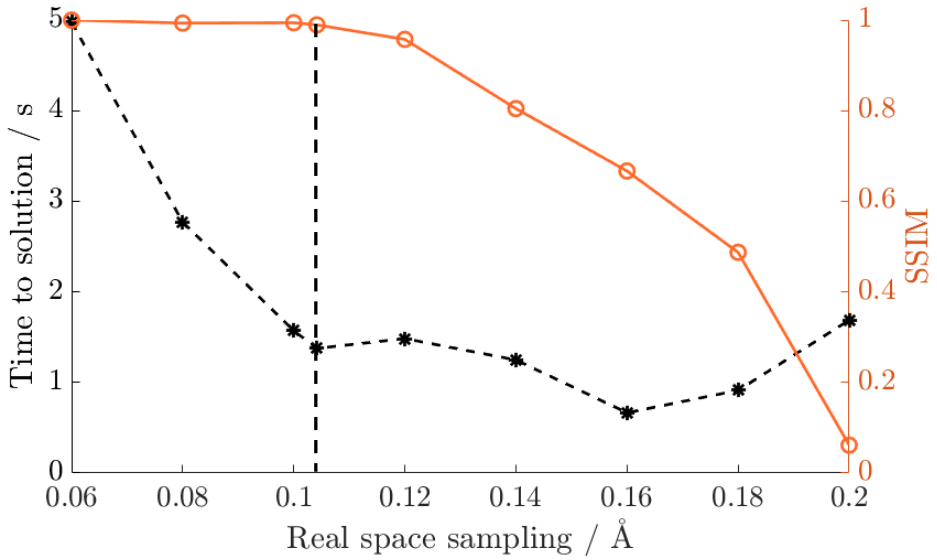

Figure 5.6: **Quality of simulation with respect to the real space sampling.** Mutlislice simulations performed using different values of real space sampling for one FPC. The vertical black dashed line indicates the optimal real space sampling based on Eq. 10.

5.3 Results

To test sub-sampling of simulations and to compare with existing simulation methods, compressed high angle annular dark field (HAADF) STEM simulations using the multislice and PRISM methods were calculated. In all cases, the simulations were performed using abTEM (version 1.0.0 beta 31) on a desktop computer equipped with an AMD Ryzen 5 2400G with Radeon Vega Graphics CPU @ 3.40 GHz, and one NVIDIA GeForce RTX 3060Ti GPU running CUDA 11.8. As already noted, it is important to note that all computation times are relative to the capability of abTEM and hence it is much better to consider the relative performance of the methods than the absolute computation times, as these are transferable to any other STEM simulation algorithm. All image recoveries were performed using a custom version of the BPFA algorithm written using CUDA so that it can be parallelised for maximum speed. The run-time of reconstructions is negligible for images of the sizes quoted (<2 seconds per reconstruction). Finally, each simulation is compared to a ground truth simulation using the Structural Similarity Index Measure (SSIM) [197] and Peak Signal to Noise Ratio (PSNR) [198]. As a rule of thumb, it is considered that a PSNR value greater than 20 dB is an acceptable reconstruction, anything over 25 dB is a very good reconstruction, and anything over 30 dB is visually indistinguishable from the ground truth.

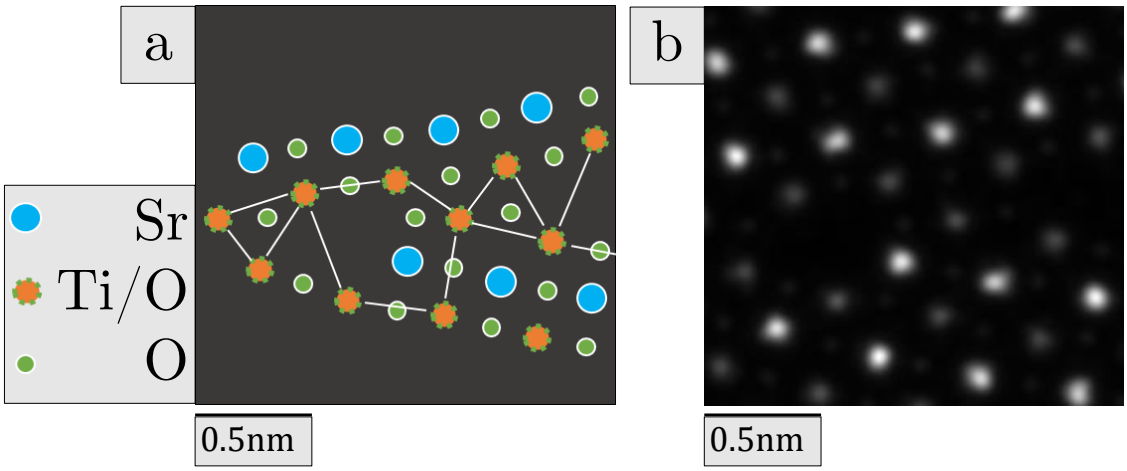

Figure 5.7: **(a) Model of the strontium titanate grain boundary and (b) multislice simulation of the structure.** The low energy grain boundary is selected due to the aperiodic structure. The model use is in line with that determined by Yang *et al.* [4].

5.3.1 Strontium titanate grain boundary

One important test of this method is its application to the calculation of contrast in grain boundaries. These complex structures are inherently difficult to interpret due to contrast variations within the void/boundary. For this purpose, a simulation of a low energy strontium titanate grain boundary ($SrTiO_3$ 22.6° $\Sigma 13(510)/[100]$) was performed (see Yang *et al.* [4] for more details on the structural model, calculation, and experimental set up).

According to Yang *et al.*, the grain boundary energy of the rigid body shift structure can be calculated through first principles, where the nonstoichiometric model is relaxed and shows a significant reduction in energy, to 0.81 ± 0.01 J/m^2. This model was chosen and simulated, as demonstrated in Fig. 5.7.

HAADF simulations were performed using both the multislice and PRISM methods through abTEM. The accelerating voltage was set at 200 kV, with a probe-forming aperture semi-angle of 24.5 mrad (Nyquist sampling of 0.256Å), and a collection semi-angle of $70 - 190$ mrad. The sample had a maximum depth of 4.5 nm (including a 0.5nm amorphous carbon layer on the surface), and both the fully sampled and sub-sampled simulations were performed with 10 frozen phonon configurations. The sub-sampled simulations were acquired at 5% sampling per frozen phonon layer, and each of the masks used was an Z-number based targeted sampling mask. All simulations had a real space sampling of 0.06Å. Therefore, the only compression was acquired through spatial sub-sampling of each frozen phonon layer. All simulations

were taken with 256×256 probe locations (scan step of 0.07Å) and each of the atoms in the input had a Debye-Waller factor determined from the method in [199] where model coefficients are modelled through phonon density-of-state curves (here a temperature of 300K is used).

The data in Fig. 5.8 shows that sub-sampling a multislice simulation is only slightly slower (24s) than using the PRISM method with an interpolation factor of 1. Furthermore, the sub-sampled multislice simulation yields functionally results to PRISM ($f = 1$) with respect to the fully sampled multislice simulation (greater than 0.9 SSIM and greater than 28dB PSNR). Using the PRISM method with sub-sampling, it is possible achieve simulation times on the order of seconds (7.5 s) with an interpolation factor of 4 and achieve SSIM values greater than 0.8, and PSNR values greater than 28dB. These values indicate a recovery that is not only functionally identical to the ground truth, but of a high quality too. Full reconstruction examples can be found in section A1.1, Fig. A1.1.

5.3.2 Monolayer molybdenum disulphide with monosulfur vacancies

2D materials are an active area of research within the electron microscopy community currently [200, 201], so naturally their simulation is also important. 2D materials are popular for their use as semiconductors, and understanding their properties is important for the development of nano-electronic devices.

The second test is therefore to determine whether the method can identify vacancies within the 2H phase of monolayer molybdenum disulfide (2H-MoS_2) [202]. Here, the monosulfur vacancy (Vs) case is specifically looked at due to it having the lowest formation energy [202] (Fig. 5.9).

Vacancies within the 2H-MoS2 structure can dramatically change the mechanical and electrical properties of this semiconductor material [203–205]. Hence, understanding the contrast through simulations is vital to determining the atomic structure to classify and statistically verify the abundance of defects and vacancies, not limited to just this specific example.

HAADF simulations were performed using both the multislice and PRISM methods through abTEM. The accelerating voltage was set at 60 kV, with a probe-forming aperture semi-angle of 39.1 mrad (Nyquist sampling of 0.3111Å), and a collection semi-angle of $86 - 200$ mrad. Both the fully sampled and sub-sampled simulations were performed with 10 frozen phonon configurations. The sub-sampled simulations were acquired at 5% sampling per frozen phonon

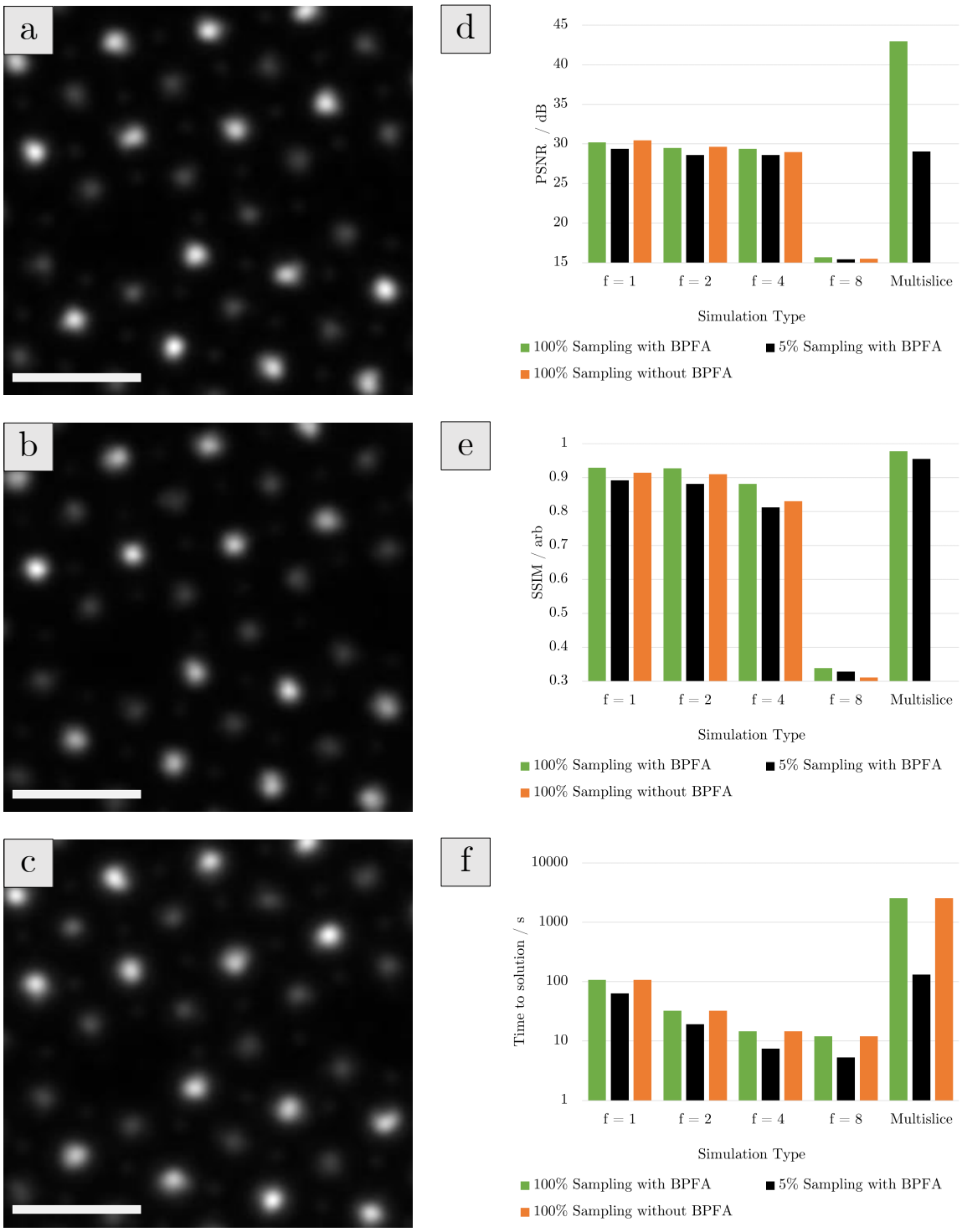

Figure 5.8: **Results for the $SrTiO_3$ grain boundary simulations.**(a) Reference simulation calculated using the multislice method, (b) compressed simulation calculated using the multislice method, and (c) simulation using the PRISM method with an interpolation factor of 2 of the $SrTiO_3$ grain boundary structure. (d-f) Plots of PSNR, SSIM and computation times of all the simulations respectively. The term f refers to the interpolation factor for PRISM simulations. The scale bar in (a-c) indicates 0.5 nm.

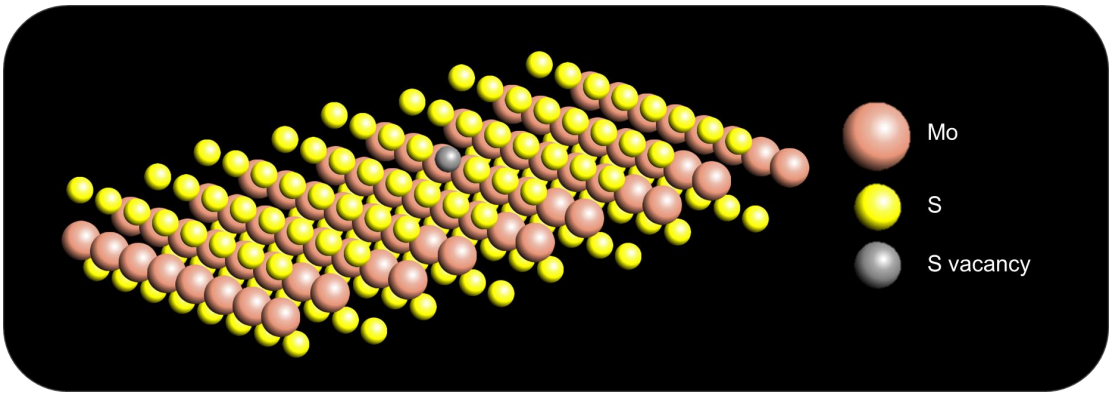

Figure 5.9: **Schematic showing the monolayer 2H-MoS_2 structure with a V_S present.** The model has been rotated such that the vacancy appears to sit on the top layer (for visibility), however it was in fact removed from the bottom layer for the simulations. The graphic was rendered using the OpenMX Viewer toolbox [5].

layer, and each of the masks used was an Z-number based targeted sampling mask. All simulations had a real space sampling of 0.05Å. Therefore, the only compression was acquired through spatial sub-sampling of each frozen phonon layer. All simulations were taken with 256×256 probe locations (scan-step of 0.063Å) and each of the atoms in the input had a Debye-Waller factor determined from the method in [199] where model coefficients are modelled through phonon density-of-state curves (here a temperature of 300K is used).

The results are similar to those in section 5.3.1 where the sub-sampled multislice method is faster than using PRISM with an interpolation factor of 1, and only slightly slower than an interpolation factor of 2. It also yields functionally identical results to both (approximately equal SSIM and PSNR values), as well as the fully sampled multislice simulation. Full reconstruction examples can be found in section A1.1, Fig. A1.2.

To validate that the monosulfur vacancy has been correctly simulated, an integrated line profile for simulations shown in Fig. 5.10(a-c) is taken and demonstrated in Fig. 9. The intensity profile across all three also shows functionally identical results, which agrees well with the experimentally observed results for contrast ratios of $0.5 : 1.0 : 2.3 - 2.5$ for Vs:S:Mo respectively [202]. The key difference however is that the PRISM method has a tailing effect in the vacuum regions, which is an artefact. This is due to the superposition of plane-waves, where some of the frequencies required to compensate for this are missing from the scattering matrix. Sub-sampling the multislice simulation does not introduce such artefacts.

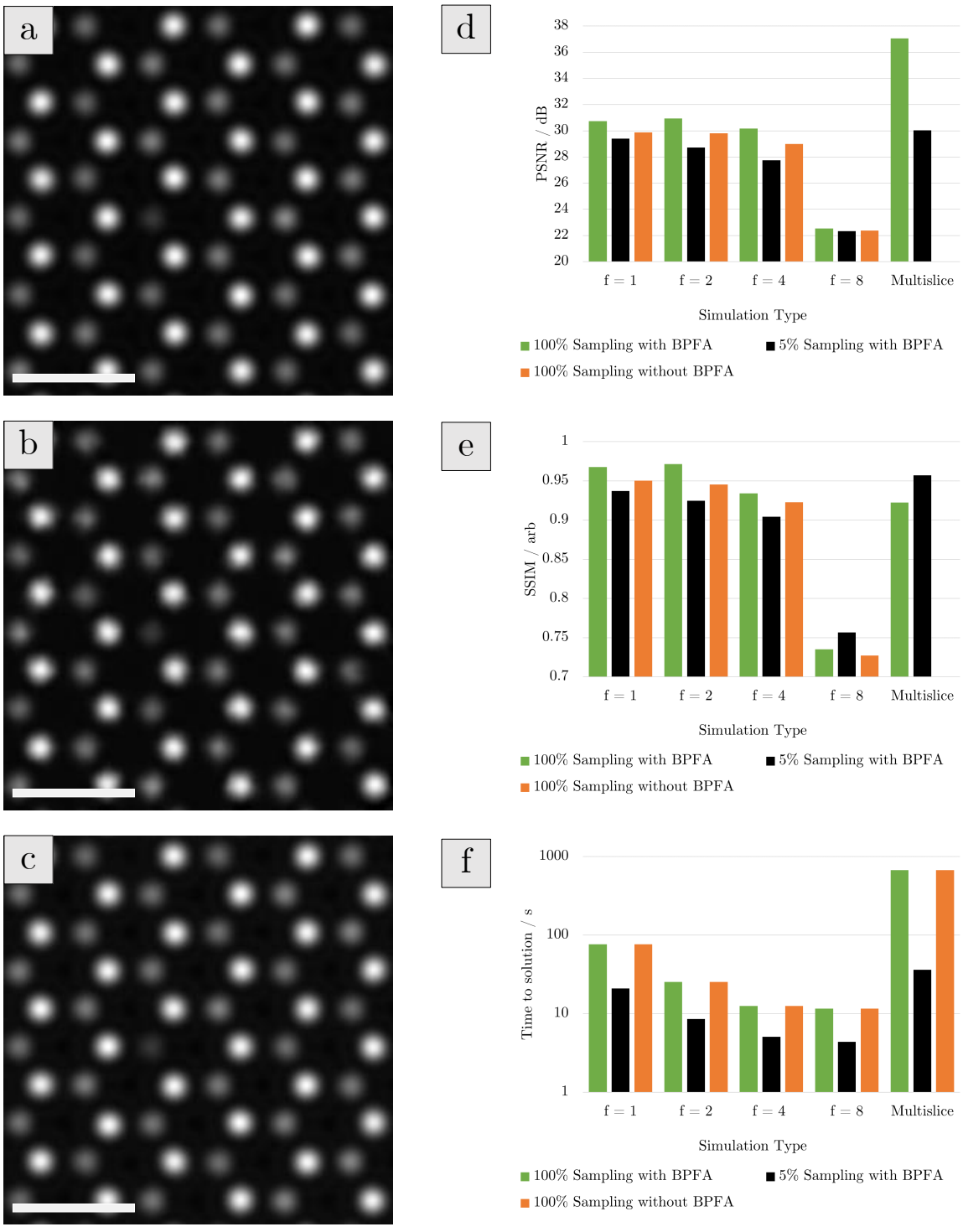

Figure 5.10: **Results for simulations of the 2H-MoS_2 structure with a V_S present.**(a) Reference simulation calculated using the multislice method, (b) compressed simulation calculated using the multislice method, and (c) simulation using the PRISM method with an interpolation factor of 2 of the 2H-MoS_2 structure with a V_S present. (d-f) Plots of PSNR, SSIM and computation times of all the simulations respectively. The term f refers to the interpolation factor for PRISM simulations. The scale bar in (a-c) indicates 0.5 nm.

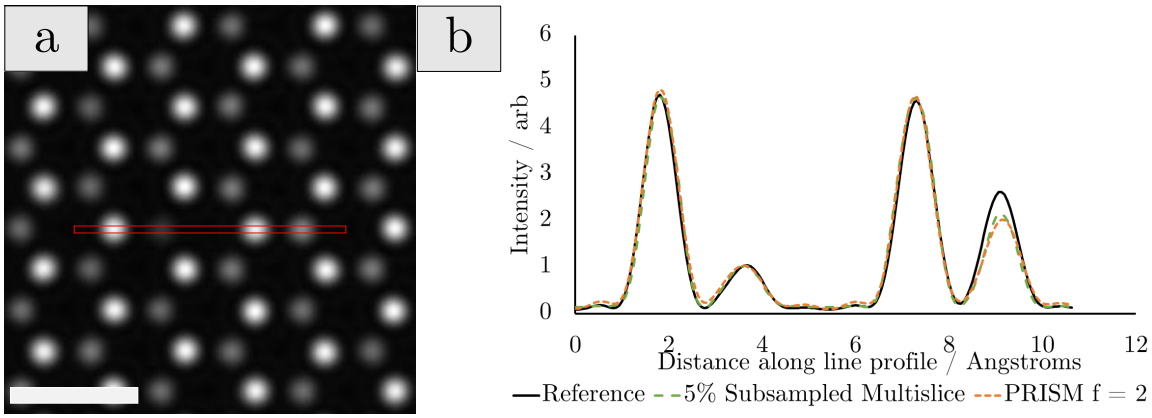

Figure 5.11: **Line profile plot for different simulation methods at the sulfur vacancy site.** (b) Integrated intensity line profiles of the images Fig. 5.10 (a-c) over the region marked by the red box in (a). The first peak, and third peak (from left to right) are molybdenum sites, the second peak is the sulfur vacancy site, and the fourth is a sulfur site. The scale bar in (a) indicates 0.5 nm.

5.3.3 Simultaneous Theoretical and Experimental Recovery

It is possible to also consider the use of simulations to seed the recovery of sub-sampled experimental data. One aspect of faster simulations is that matching simulation to experiment is more time efficient. This means more parameter testing can be performed in a shorter time frame. Given that the theory of electron scattering is very well understood [206], it would make sense to use simulations in a more practical aspect during acquisition. One of the ways this can be done is through (dictionary) transfer learning [207] where the dictionary from a simulation is used to seed the recovery of real sub-sampled data.

To test this method, we consider an yttrium silicide (Y_5Si_3) sample. The following paragraph is for completeness and can be skipped if the reader is familiar with the properties of the sample.

Yttrium silicide is part of the electride class of compound materials. An electride is a framework composed cation and anion sublattices. These sublattices have a net positive electric charge which are balanced by loosely bonded, interstitial anionic electrons [208]. Y_5Si_3 has been well proposed as a low Schottky barrier material for n-type silicon semiconductors thanks to its low Schottky barrier height of 0.27 eV [209]. It has also been recently proposed as an encapsulation material to capture radioactive volatile products within nuclear fission reactors [210]. All of this makes Y_5Si_3 a versatile material, hence understanding its properties are important. A recent paper from Q. Zheng *et al.* [208] looked at the local charge density of

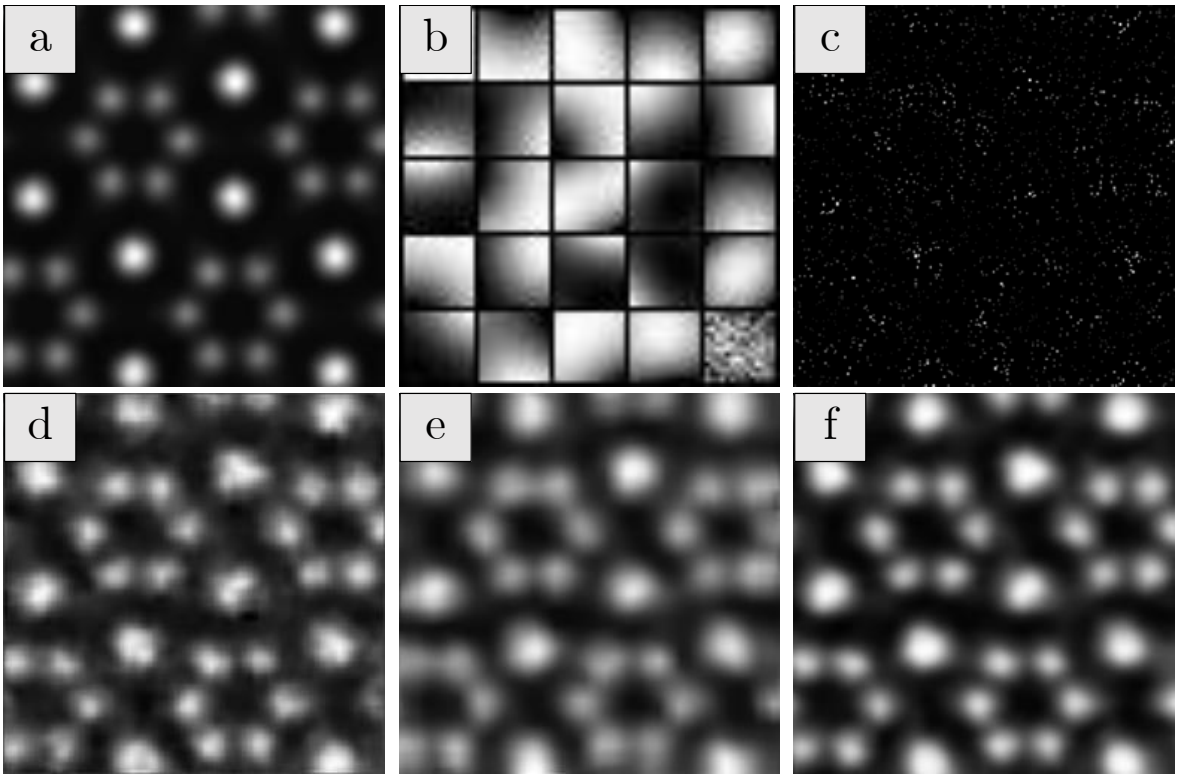

Figure 5.12: **Using a simulation to seed recovery for experimental Y_5Si_3 acquisition.** (a) Recovery from 5% sub-sampled multislice simulation of Y_5Si_3, and (b) the dictionary determined by BPFA. This dictionary is then used to reconstruct (c) a 3% sub-sampled acquisition of Y_5Si_3 giving (d) a reconstruction through OMP with a PSNR of 24.8 dB and an SSIM of 0.87. (e) Reconstruction using only BPFA to learn the dictionary and reconstruct at the same sampling rate with a PSNR of 22.8 dB and an SSIM of 0.86. Both comparisons are made to (f) the ground truth which was passed through BPFA at 100% sampling to denoise only.

Y_5Si_3 by performing differential phase contrast experiments. They also used simulations to verify the contrast in their HAADF STEM images, specifically the missing contrast of the silicon atoms. While the Zheng paper performed a matching between the experiment and theory, this example is going to consider the simulation as the seed for recovering the real STEM data, *i.e.*, they are solved and matched simultaneously.

A simulation (matching experimental parameters, see [208] for details) of Y_5Si_3 was performed and then recovered using BPFA. The dictionary which is learned is then used to recover a sub-sampled acquisition of Y_5Si_3 through an orthogonal matching pursuit (OMP) algorithm [152]. The result by passing the same acquisition through only BPFA is shown for comparison. The scan-step is 0.06Å, and the Nyquist sampling is 0.309Å.

Fig. 5.12 shows that by using transfer learning it is possible in this case to improve the reconstruction quality by 2 dB. This is likely due to the simulation being free of noise, and

therefore its dictionary is free of noise. Given that OMP approximates the correct weights to apply to each of the atoms, it can rescale the intensity if needed, matching directly to the real sub-sampled data. However, BPFA must learn a dictionary from the real sub-sampled data, which at low sampling rates can sometimes be challenging if the noise levels are also high, the dictionary transfer approach can show an advantage in the quality of reconstruction. This is demonstrated in the differences between Fig.5.12(d) and (e) where the yttrium columns are more refined using dictionary transfer than using BPFA alone. This is particularly important, as sampling at 3% [211, 212] using a scan generator correlates to a 33x speed up in image acquisition and 33x less total electron dose during the experiment. For beam sensitive materials this is excellent as the images acquired will be more representative of the pristine sample due to knock-on and radiolysis damage occurring fewer times overall [213]. For less beam-sensitive materials, the speed up in acquisition means fewer artefacts in the image caused by stage drift.

In practice, however, it is found that line-hop sampling (*i.e.*, random walk) [214] is a better alternative sampling strategy for experimental data and it balances sparsity of acquisition and the limiting effects of hysteresis. UDS sampling is the optimal set up for image recovery [215], however is limited in experiment due to hysteresis [214]. Comparison of UDS and line-hop can be found in previous work [211] and readers are referred to chapter 3 for examples of different mask types.

5.4 Conclusions

The methods have been implemented using both the PRISM and multislice algorithms, demonstrating a robustness of the sub-sub-sampling approach to increasing speed of simulations without significant loss of accuracy. It is also competitive in terms of its performance to the state-of-the-art method for faster simulation (PRISM). This highlights the effectiveness of sparse acquisition, in that much of the data requirements for STEM can be significantly reduced. Only a subset of the data is required to recover a functionally identical result through image inpainting.

It is observed with the experimental images that using a sub-sampling strategy need not be independent from other helpful approaches – other strategies for noise reduction, super resolution etc can be applied to the inpainted image. It is found that the same effect here with the simulations where the interpolation approach of PRISM can be used in conjunction with

sub-sampling at a lower interpolation factor without significant loss of information. For example, a simulation with an interpolation factor of 4 and sub-sampling at 5% gives significantly better results than a simulation with an interpolation factor of 8 at 100% sampling, running almost twice as fast in the process.

Although the demonstrations here have only shown results for HAADF simulations, the methods described are applicable to all STEM imaging modes such as bright field, annular bright field, and 4D-STEM. For example, 4D-STEM can approximate sample thickness using position average convergent beam electron diffraction by matching simulation to experiment [208].

Furthermore, it has also been demonstrated that STEM simulations can seed the dictionary for real image reconstruction, meaning faster 'live' reconstructions from experimental CS-STEM data. In the BPFA algorithm it is possible to begin the dictionary learning from a custom dictionary. By seeding the initial dictionary from the dictionary acquired from a simulation of the material in question, it could potentially speed up the convergence of the algorithm, or equally make the live reconstructions more accurate after fewer iterations. It is evidenced in this work that transfer learning with an OMP can give functionally identical results to the ground truth. This can be even faster to recover than just using BPFA to reconstruct, indicating further speed improvements.

In conclusion, sub-sampling a multislice simulation can be both faster and a better representation of the full multislice simulation than the PRISM method at 100% sampling (depending on the image quality metric used). This is analogous to using fewer but more intense probes in real CS-STEM, as opposed to a full raster scan at low dose acquisition [211]. This shows that in some cases it is better to have a more accurate calculation of individual probes than interpolating each probe estimate over the full scan area.

Transferring the dictionary learned from a simulation can also yield better results than blind inpainting the raw acquisition. This uses theory and experimental data together as opposed to just matching and comparing results. For microscopists applying this in a practical sense, it would speed up analysis of the imaging conditions, allowing for faster adjustment and hence less total electron dose on the sample. This could improve the final image quality for many materials by reducing the amount of sample damage, driving forward new science in beam sensitive materials.

As noted, the purpose of faster simulation is not to improve their accuracy *per se*, but to improve the efficiency of calculation for faster determination of properties and characteristics in conjunction with experimental data. In the case where the accuracy of experimental data is not perfect, it is proposed that the simulation itself does not need to be perfect if time is a constraint. Of course, one could calculate an improved simulation for in-depth analysis, but if it is possible to realise real-time simulations then they could be used to assist a microscopist during acquisition. It may also be possible for fast STEM simulations (in conjunction with deep learning) to interpret the correct adjustments needed for real-time acquisition automatically. This would allow for both faster and more efficient alignment, which is of particular importance to studies of beam sensitive materials.

6 | Applying Compressed Sensing Methods to 4-D STEM

6.1 Overview of 4-D STEM

4-dimensional STEM (4-D STEM) is a powerful acquisition method for obtaining high quality analysis for a range of specimens. In this imaging mode a series of diffraction patterns for each probe position in a 2D grid are recorded in the far field on a 2D pixelated detector, as shown in Fig. 6.1. This gives resolution in the real and reciprocal space, identifying the number of electron counts which were scattered to a range of angles. From this it is possible to extract phase information, electronic properties, and magnetic properties. These methods will be covered in more detail as the chapter progresses.

It is important to appreciate the developments over the last three decades which have lead to 4-D STEM becoming so popular, and why it is so. Prior to the widespread use of aberration correctors, Nellist *et al.*demonstrated one of the earliest cases of 4-D STEM where coherent micro-diffraction patterns were collected as a function of probe position and used for a super-resolved ptychographic reconstruction [216]. This allowed the resolution of the Si $\{004\}$ at 0.136nm; a much higher spatial resolution than was achieveable using high-angle annular dark field (HAADF) STEM on the instrument used. Another early demonstration by Zaluzec *et al.*, used position resolved diffraction to image distributions of magnetic induction in a Lorentz STEM imaging mode [217, 218].

4-D STEM has progressed significantly since these early demonstrations, with more recent examples of its application in ptychography having been used to recover the complex object wavefunction of weakly scattering objects, such as lithium ion cathode materials [219] and

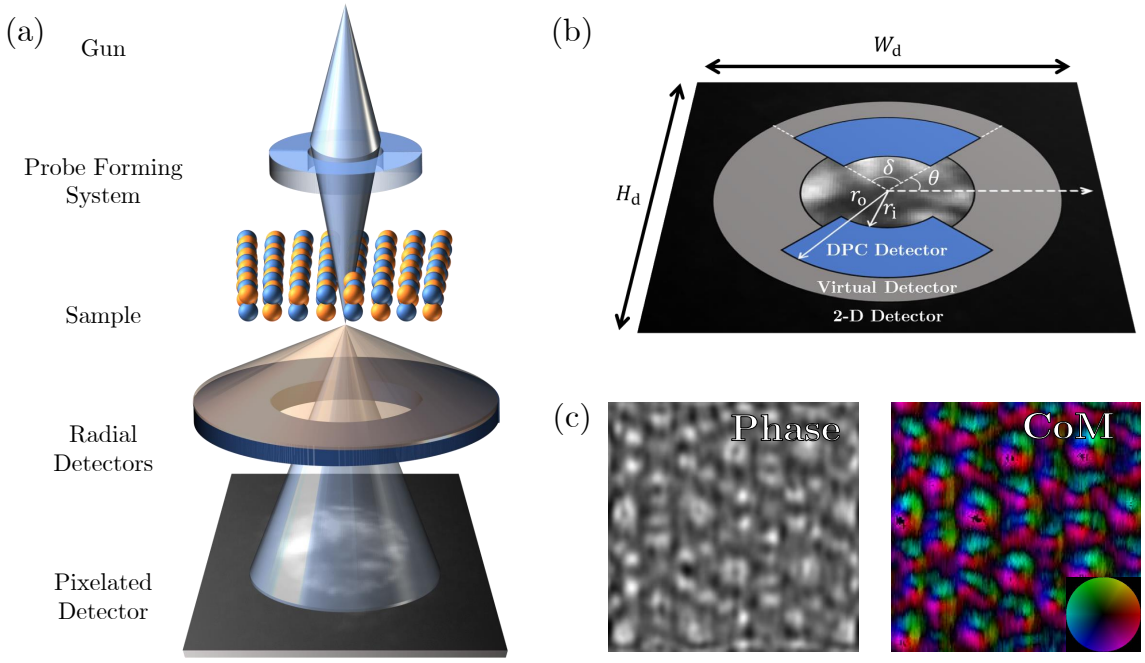

Figure 6.1: **Operating principles and analysis examples of 4-D STEM**.(a) Electrons are converged to form a probe which is rastered in 2-D across the sample plane. The transmitted electrons are collected using a 2-D detector in the far field for each probe position. (b) Examples of virtual detectors which can be applied to 4-D STEM data to emulate fixed integrating detectors typically found in a STEM. (c) Examples of analysis methods which utilise the diffraction data to extract phase information.

biological samples [220]. STEM ptychography has also been used to resolve praseodymium dumbbells at the limit set by thermal atomic motion [83]. 4-D STEM has become popular due to its versatility by way of multi-modal imaging using virtual detectors (VDs) [221], differential phase contrast (DPC) [222], centre of mass (CoM) analysis [223], and ptychography [224–228]. A major limitation in the application of 4-D STEM has been the need for long integration times to a achieve significant signal-to-noise ratio (SNR) in the presence of noise and dark current. However, the recent advent of commercially available direct electron detectors [229–232] means that CBED patterns can be acquired with little or no detector noise at up to 100,000 frames per second (fps), equivalent to an effective dwell time of 10µs per probe position [232, 233]. These detectors are specifically designed for 4-D STEM acquisition, however they are state-of-the-art, and the majority of detectors in use are typically operating at 1,000 fps up to 10,000 fps. While this is a significant improvement over earlier indirect scintillator coupled detectors operating at *ca.* 30 fps [234, 235], it is still at least a factor of ten too slow to be able to match the dwell time of traditional solid state monolithic STEM detectors. Hence,

4-D STEM experiments remain susceptible to drift and beam induced damage [213] which potentially limits its applicability to studies of beam sensitive organic and hybrid materials or to investigations of materials dynamics.

One option to overcome beam damage is to reduce the electron fluence at the sample [236, 237]. By reducing the fluence below a materials dependent threshold [238], or by using cryogenic temperatures [220], beam damage can be reduced. Furthermore, if combined with alternative methods to increase acquisition speeds such as low bit-depth electron counting [239, 240], the acquisition speed can be increased and sample drift can be reduced. However, given that the SNR is related to the number of detected electrons, and hence, with the fluence per probe position, a combination of fluence and fast acquisition quickly transitions the experiment to conditions that are below the minimum signal-to-noise requirements for 4-D methods such as ptychography [241].

An alternative method to overcome beam damage in STEM is by using techniques based on the theory of compressive sensing (CS) [145, 146], which is referred to here as probe sub-sampling. Probe sub-sampling in this context refers to controlling the set of positions of the STEM probe visits within a raster scan to reduce the number of acquisition points - thereby directly creating a faster scan and a lower fluence and flux at the sample. Probe sub-sampling has already been experimentally demonstrated for a variety of experimental STEM and SEM imaging modes [211, 212, 242–248], and has also been used to speed up the computational time for STEM simulations [249–251]. The key benefit for probe sub-sampling in STEM is that by acquiring data from fewer probe locations acquisition rates can be increased, which in turn reduces artefacts caused by sample drift as well as reducing the total cumulative electron fluence. Thus, samples which are susceptible to beam damage can be imaged at usable SNRs, without over exposure to the incident beam.

In this chapter, a new focused probe acquisition method is demonstrated which reduces beam damage and increases acquisition rate by probe sub-sampling. Only a subset of the CBED patterns are acquired and the BPFA is used to recover the full 4-D STEM data set from the sub-sampled measurements. Simulations of this method applied to a simulated 4-D STEM data set, as well as an experimental yttrium silicide data set are given, and demonstrate that 4-D STEM data acquisition can be reduced by at least 256$\times$ without significant quality loss in all imaging modes.

Previous work by Stevens *et al.* [245], showed the potential of recovering phase data from sub-sampled 4-D STEM data. Both probe sub-sampling and detector sub-sampling (where a subset of detector rows are readout at random) were used to show that with only 1% of the acquired data, inpainting followed by phase retrieval can recover functionally identical [1] results to a fully sampled experiment. In this work the inpainting of the 4-D data used a Kruskal-factor analysis technique [252]. Here, this approach is extended by using a new implementation of the BPFA algorithm which takes advantage of GPU acceleration, as well as providing a comparison between different 4-D STEM analysis techniques using inpainted data. The work of Zhang *et al.* [253] is also built upon, who showed that the number of detector pixels required for ptychographic reconstruction can be reduced significantly without loss of real space resolution by applying this analysis to an experimental data set (i.e. with noise).

6.2 Methods

This section presents the method for acquiring and inpainting sub-sampled 4-D STEM data, as well as the analysis methods which are common in 4-D STEM analysis. Both subsections contain rigorous mathematical descriptions which are consistent with previous notations.

6.2.1 Compressive 4-D STEM

The experimental set-up for the acquisition of a sub-sampled data set is shown in Fig. 6.1. Assume a pixelated detector with H_d and W_d pixels in the vertical and horizontal axis, respectively, collecting 2-D CBED patterns of size $H_\mathrm{d} \times W_\mathrm{d}$. Let $\Omega_\mathrm{d} \coloneqq \{1, \cdots, H_\mathrm{d}\} \times \{1, \cdots, W_\mathrm{d}\}$ be the set of all detector pixel locations and $\boldsymbol{k}_\mathrm{d} \coloneqq (k_\mathrm{d}^\mathrm{h}, k_\mathrm{d}^\mathrm{w}) \in \Omega_\mathrm{d}$ denote the coordinates of a detector pixel. Further assume an electron probe scanning a regular grid of H_p and W_p locations in the vertical and horizontal axis, respectively[2], collected in a probe locations set $\Omega_\mathrm{p} \coloneqq \{1, \cdots, H_\mathrm{p}\} \times \{1, \cdots, W_\mathrm{p}\}$. Let $\boldsymbol{r}_\mathrm{p} \coloneqq (r_\mathrm{p}^\mathrm{h}, r_\mathrm{p}^\mathrm{w}) \in \Omega_\mathrm{p}$ denote the coordinates of a probe location. Moreover, the total number of detector pixels and probe locations are denoted by, respectively, $N_\mathrm{p} = H_\mathrm{p} W_\mathrm{p}$ and $N_\mathrm{d} = H_\mathrm{d} W_\mathrm{d}$. Finally, given a scan step parameter Δ_p, in m, of the electron probe and detector pixel size Δ_d, in mrad, the location of the scanning probe and detector pixel can be converted from their index units to real units.

[1] Functionally identical results are defined as the preservation of features compared to the ground truth, such that the analysis is preserved in determining properties of the sample.

[2] Note that the coordinate axes of the pixelated detector and scanning probe are not necessarily the same.

Let $\mathcal{X} \in \mathbb{R}^{H_\mathrm{p} \times W_\mathrm{p} \times H_\mathrm{d} \times W_\mathrm{d}}$ be the discretised 4-D representation of fully sampled 4-D STEM data; and $\mathcal{X}(\boldsymbol{r}_\mathrm{p}, \boldsymbol{k}_\mathrm{d})$ be the 4-D STEM data observed at probe location $\boldsymbol{r}_\mathrm{p}$ and detector pixel $\boldsymbol{k}_\mathrm{d}$. A *CBED pattern* collected at probe location $\boldsymbol{r}_\mathrm{p}$ is denoted by $\boldsymbol{X}^{\mathrm{dp}}_{\boldsymbol{r}_\mathrm{p}} \coloneqq \mathcal{X}(\boldsymbol{r}_\mathrm{p}, \cdot) \in \mathbb{R}^{H_\mathrm{d} \times W_\mathrm{d}}$. In this work, the *virtual image* corresponding to a detector pixel $\boldsymbol{k}_\mathrm{d}$, represented as $\boldsymbol{X}^{\mathrm{vi}}_{\boldsymbol{k}_\mathrm{d}} \coloneqq \mathcal{X}(\cdot, \boldsymbol{k}_\mathrm{d}) \in \mathbb{R}^{H_\mathrm{p} \times W_\mathrm{p}}$, refers to a matrix collecting the data observed at detector pixel $\boldsymbol{k}_\mathrm{d}$ for all probe positions.

The compressed 4-D STEM to reduce beam damage and increase acquisition speed is now introduced. This is achieved by sub-sampling $M_\mathrm{p} \ll N_\mathrm{p}$ probe locations acquired in the sub-sampling set $\Omega \subset \Omega_\mathrm{p}$, which is equivalent to sub-sampling each of the virtual images (sharing a common mask determined by Ω). This defines the acquisition model as,

$$\boldsymbol{Y}^{\mathrm{vi}}_{\boldsymbol{k}_\mathrm{d}} = \boldsymbol{P}_\Omega(\boldsymbol{X}^{\mathrm{vi}}_{\boldsymbol{k}_\mathrm{d}}) + \boldsymbol{N}_{\boldsymbol{k}_\mathrm{d}} \in \mathbb{R}^{H_\mathrm{p} \times W_\mathrm{p}}, \quad \text{for } \boldsymbol{k}_\mathrm{d} \in \Omega_\mathrm{d}, \tag{1}$$

where $\boldsymbol{Y}^{\mathrm{vi}}_{\boldsymbol{k}_\mathrm{d}}$ is the sub-sampled measurements at detector pixel $\boldsymbol{k}_\mathrm{d}$ and $\boldsymbol{P}_\Omega$ is a mask operator with $(\boldsymbol{P}_\Omega(\boldsymbol{U}))_{(i,j)} = \boldsymbol{U}_{(i,j)}$ if $(i,j) \in \Omega$ and $(\boldsymbol{P}_\Omega(\boldsymbol{U}))_{(i,j)} = 0$ otherwise, and $\boldsymbol{N}_{\boldsymbol{k}_\mathrm{d}}$ is an additive noise.

The core of the recovery method assumes that the patches of every virtual image are sparse in a shared dictionary, *i.e.*, $\boldsymbol{x}^{\mathrm{vi}}_i = \boldsymbol{D}\boldsymbol{\alpha}_i$, where $\boldsymbol{x}^{\mathrm{vi}}_i \coloneqq \mathrm{vec}(\boldsymbol{X}^{\mathrm{vi}}_i) \in \mathbb{R}^{B^2}$ is a vectorised version of $\boldsymbol{X}^{\mathrm{vi}}_i$, $\boldsymbol{D} \in \mathbb{R}^{B^2 \times K}$ denotes the dictionary with K atoms and $\boldsymbol{\alpha}_i \in \mathbb{R}^K$ is a sparse vector of weights or coefficients for the i^{th} patch of the virtual image. Based on these definitions, the BPFA algorithm allows us to infer $\boldsymbol{D}$, $\boldsymbol{\alpha}_i$, and the noise statistics and in turn reconstruct the virtual images in a sequential fashion. The details on the BPFA can be found in Section. 3.2.1.

The advantages of this approach include the ability to infer both the noise variance and sparsity level of the signal in the dictionary, and allows for the learning of dictionary elements directly from sub-sampled data. This approach has been tested in previous reports [211, 212, 248, 249, 251] and has shown success when applied to electron microscopy data. Note that this approach learns a different dictionary for each virtual image and a BPFA instance is applied to every virtual image. This is not necessarily optimal, however the concept of learning a shared dictionary for all virtual images and applying a single instance of BPFA directly on the sub-sampled 4-D data is left to a future study.

In addition to probe sub-sampling, it is also possible to down sample the detector pixels

to eliminate redundancy. This can also be inferred as the optimisation of the reciprocal space sampling, Δ_{d}, which can be carried out by only reading out the set of rows which are within the sampling set. This is different to conventional detector pixel binning (which still requires reading of all rows within the total CBED pattern), since it does not consider nor acquire rows which do not belong to the sampling set.

Given the detector down sampling factor $f_{\mathrm{d}} \in \mathbb{N}$, uniformly read-out every $f_{\mathrm{d}}^{\mathrm{th}}$ row on the detector. This results in faster acquisition of CBED patterns of size $H_{\mathrm{d}}/f_{\mathrm{d}} \times W_{\mathrm{d}}$ pixels. To further reduce the size of the data-set, keep only the data from every $f_{\mathrm{d}}^{\mathrm{th}}$ column on the detector; resulting in CBED patterns with $M_{\mathrm{d}} = H_{\mathrm{d}} \cdot W_{\mathrm{d}}/f_{\mathrm{d}}^2$ entries. In this work, the detector down sampling ratio is defined as $M_{\mathrm{d}}/N_{\mathrm{d}} = 1/f_{\mathrm{d}}^2$. In practice, it could also be possible to vary the camera length to optimise Δ_{d} since the camera length is inversely proportional to the reciprocal space sampling. This would account for detectors where an individual row cannot be read out independently, but instead can only accept read-outs in blocks of rows.

Given the properties of applying detector down sampling, inpainting is not generally required hence the inpainting step remains unchanged. It is postulated that sparse detector sampling could further increase the rate of 4-D STEM data acquisition beyond that presented here.

6.2.2 Data analysis methods

Following acquisition of 4-D STEM data, various techniques such as VDs, DPC, CoM analysis, and phase retrieval techniques such as ptychography can be used for analysis. In all cases, the geometrical centre of the CBED patterns are aligned for consistent analysis.

Virtual detectors

A VD is analogous to fixed detectors which are typically used in STEM. A VD, as illustrated in Fig. 1(c), is characterised by inner and outer collection semi-angles r_{i}, $r_{\mathrm{o}} \in \mathbb{R}^+$, respectively (in mrad). Given those parameters, each 2-D CBED pattern is summed over a selected angular range. Setting $\Omega^{\mathrm{vd}} \coloneqq \Omega^{\mathrm{vd}}(r_{\mathrm{i}}, r_{\mathrm{o}}) \subset \Omega_{\mathrm{d}}$ as the set of detector pixel indices that falls within the radial range of the detector; and letting $\boldsymbol{Z}^{\mathrm{vd}} \in \mathbb{R}^{H_{\mathrm{p}} \times W_{\mathrm{p}}}$ be the VD image. Therefore, the value of the VD at probe location $\boldsymbol{r}_{\mathrm{p}}$, denoted by $z_{\boldsymbol{r}_{\mathrm{p}}}^{\mathrm{vd}}$, will be the sum of the 4-D STEM data at probe

location $\boldsymbol{r}_\mathrm{p}$ restricted to the pixels indexed in Ω_vd, *i.e.*,

$$z_{\boldsymbol{r}_\mathrm{p}}^{\mathrm{vd}} = \sum_{\boldsymbol{k}_\mathrm{d} \in \Omega^{\mathrm{vd}}} \mathcal{X}(\boldsymbol{r}_\mathrm{p}, \boldsymbol{k}_\mathrm{d}) \tag{2}$$

Examples of annular bright field (ABF) and low-angle annular dark field (LAADF) virtual detectors are shown in 6.2.

Differential phase contrast

DPC measures the projected electric field of a sample by quantifying the shift in the electron beam using a segmented (virtual) detector. As depicted in Fig. 1(c), a DPC detector is similar to a VD, but also includes an angular rotation $\theta \in [0, 2\pi)$ about the centre of the detector and an angular width $\delta \in [0, 2\pi)$. Let $\Omega^{\mathrm{dpc}^+} := \Omega^{\mathrm{dpc}^+}(r_\mathrm{i}, r_\mathrm{o}, \theta, \delta) \subset \Omega_\mathrm{d}$ be the set of detector pixel indices whose radii are within the radial range of the detector and whose angles are in the range of θ and $\theta + \delta$. Similarly, let $\Omega^{\mathrm{dpc}^-} := \Omega^{\mathrm{dpc}^-}(r_\mathrm{i}, r_\mathrm{o}, \theta, \delta)$ be the set of pixel indices whose radii are within the radial range of the detector and whose angles are in the range of $\theta + \pi$ and $\theta + \pi + \delta$. Consequently the DPC image is defined by $\boldsymbol{Z}^{\mathrm{dpc}} \in \mathbb{R}^{H_\mathrm{p} \times W_\mathrm{p}}$. Therefore, the value of the DPC image at probe location $\boldsymbol{r}_\mathrm{p}$, denoted by $z_{\boldsymbol{r}_\mathrm{p}}^{\mathrm{dpc}}$, will be the sum of the CBED pattern at that location and restricted to the pixels indexed in Ω^{dpc^+} minus the sum of that pattern restricted to the pixels indexed in Ω^{dpc^-}:

$$z_{\boldsymbol{r}_\mathrm{p}}^{\mathrm{dpc}} = \sum_{\boldsymbol{k}_\mathrm{d} \in \Omega^{\mathrm{dpc}^+}} \mathcal{X}(\boldsymbol{r}_\mathrm{p}, \boldsymbol{k}_\mathrm{d}) - \sum_{\boldsymbol{k}_\mathrm{d} \in \Omega^{\mathrm{dpc}^-}} \mathcal{X}(\boldsymbol{r}_\mathrm{p}, \boldsymbol{k}_\mathrm{d}) \ . \tag{3}$$

Example of a differential phase contrast virtual detector is shown in 6.2.

Centre of mass

The CoM field vector which quantifies the 2-D shift at probe location $\boldsymbol{r}_\mathrm{p}$ is denoted by $\boldsymbol{z}_{\boldsymbol{r}_\mathrm{p}}^{\mathrm{com}} \in \mathbb{R}^2$ to construct a full CoM vector field $\boldsymbol{Z}^{\mathrm{com}} \in \mathbb{R}^{H_\mathrm{p} \times W_\mathrm{p} \times 2}$. Let $\Omega^{\mathrm{bd}} := \Omega^{\mathrm{bd}}(\boldsymbol{k}_\mathrm{d}) \subset \Omega_\mathrm{d}$ be the set of detector pixel indices that falls within the desired shift measurement region (typically the bright field disk). Assume that each CBED pattern can be modelled as a non-uniform density lamina where the density is equivalent to the intensity of the signal in the CBED pattern. By

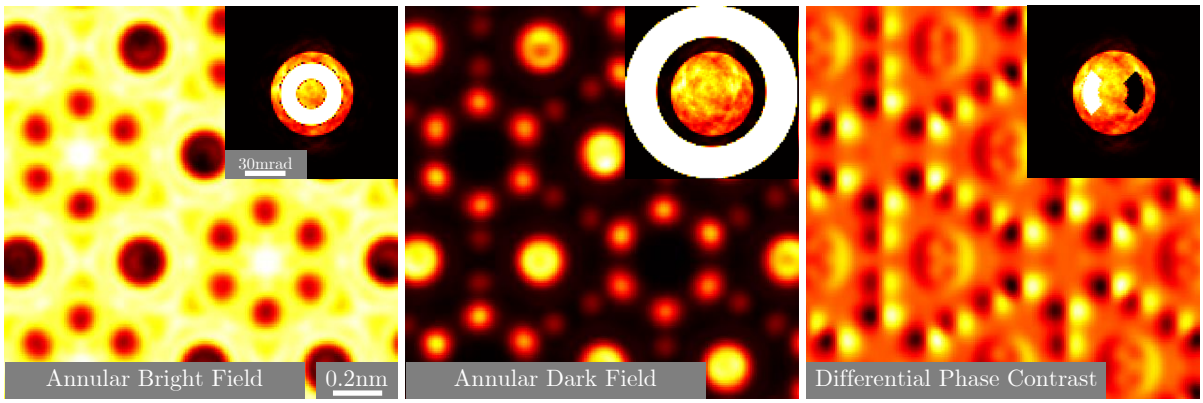

Figure 6.2: **Examples of virtual detectors for 4-D STEM analysis**.(left) Annular bright field virtual detector, (middle) low-angle annular dark field virtual detector, and (right) differential phase contrast virtual detector where for this specific detector, the white region has an amplitude of +1, and the black region of -1 when multiplied by the diffraction pattern.

using standard derivations to derive the CoM field coefficients, $z_{\boldsymbol{r}_\mathrm{p}}^{\mathrm{com}}$, as

$$z_{\boldsymbol{r}_\mathrm{p}}^{\mathrm{com}} = \frac{\sum_{\boldsymbol{k}_\mathrm{d}\in\Omega^{\mathrm{bd}}}(\boldsymbol{k}_\mathrm{d}-\boldsymbol{c}_\mathrm{d})\cdot\mathcal{X}(\boldsymbol{r}_\mathrm{p},\boldsymbol{k}_\mathrm{d})}{\sum_{\boldsymbol{k}_\mathrm{d}\in\Omega^{\mathrm{bd}}}\mathcal{X}(\boldsymbol{r}_\mathrm{p},\boldsymbol{k}_\mathrm{d})} \ , \tag{4}$$

where $\boldsymbol{c}_\mathrm{d} \in \mathbb{R}^2$ are the coordinates of the centre of the CBED pattern. Following this, the CoM displacement can be given as the magnitude or the angle of the vector $z_{\boldsymbol{r}_\mathrm{p}}^{\mathrm{com}}$.

In order to calculate the projected electric field, a field constant term is introduced as,

$$C_\mathrm{f} = \frac{hv}{e\lambda} \ , \tag{5}$$

where e, h, v are the elementary charge, Planck constant, and electron velocity respectively. The magnitude of the centre of mass at each probe location forms a pixel-wise centre of mass shift $z_{\boldsymbol{r}_\mathrm{p}}^{|\mathrm{com}|} = |z_{\boldsymbol{r}_\mathrm{p}}^{\mathrm{com}}|$. This is then converted to units of radians by multiplying the intensity in terms of pixels by $\Delta_\mathrm{d} \times 10^{-3}$, and the projected electric field $\boldsymbol{V} \in \mathbb{R}^{H_\mathrm{p}\times W_\mathrm{p}}$ is therefore given as,

$$\boldsymbol{V} = C_\mathrm{f}(\Delta_\mathrm{d}\times 10^{-3})\boldsymbol{z}^{|\mathrm{com}|} \ . \tag{6}$$

The projected charge density can also be estimated by taking the divergence of $z_{\boldsymbol{r}_\mathrm{p}}^{\mathrm{com}}$. The projected charge density $\boldsymbol{\rho} \in \mathbb{R}^{H_\mathrm{p}\times W_\mathrm{p}}$ is given as,

$$\boldsymbol{\rho} = \frac{C_\mathrm{f}\epsilon_0}{e}(\Delta_\mathrm{d}\times 10^{-3})\mathrm{div}(\boldsymbol{z}^{\mathrm{com}}) \ . \tag{7}$$

Wigner Distribution Deconvolution

Ptychography is a technique that recovers the complex object wavefunction illuminated by a (partially) coherent source, which in the case of STEM is a focused or intentionally defocused probe. There are a number of analytical and iterative algorithms [254–259] that recover the wave-function; here an adaptation of the Wigner distribution deconvolution (WDD) [260, 261] is used, which is one method for object phase recovery for focused probe illumination [219, 262–264].

Firstly a definition of the observed CBED patterns is introduced as,

$$\mathcal{X}(\boldsymbol{r}_p, \boldsymbol{k}_d) = |\mathcal{I}(\boldsymbol{r}_p, \boldsymbol{k}_d)|^2 \tag{8}$$

where,

$$\mathcal{I}(\boldsymbol{r}_p, \boldsymbol{k}_d) = \int P(\boldsymbol{r} - \boldsymbol{r}_p) o(\boldsymbol{r}) \exp\left(i2\pi \boldsymbol{r} \cdot \boldsymbol{k}_d\right) d\boldsymbol{r} \tag{9}$$

which implies that $\mathcal{X}$ is a convolution between the object transfer function $o(\boldsymbol{r})$ and probe function $P(\boldsymbol{r})$. To recover the object phase, the $\mathcal{H}$-matrix (or $\mathcal{H}$-array, for the sake of consistent notation) is calculated, which is the Fourier transform of a 4-D STEM data-set with respect to real space probe locations, followed by an inverse Fourier transform with respect to the detector pixel locations, *i.e.*,

$$\mathcal{H}(\boldsymbol{k}_\mathrm{p}, \boldsymbol{r}_\mathrm{d}) = \mathcal{F}^{-1}_{\boldsymbol{k}_\mathrm{d}}\Big[\mathcal{F}_{\boldsymbol{r}_\mathrm{p}}\big[\mathcal{X}(\boldsymbol{r}_\mathrm{p}, \boldsymbol{k}_\mathrm{d})\big]\Big] \ , \tag{10}$$

where $\boldsymbol{k}_\mathrm{p}$ are the reciprocal space coordinates of the probe locations and $\boldsymbol{r}_\mathrm{d}$ are real space coordinates with respect to the detector pixels.

For a general function $f(\boldsymbol{u})$, its Wigner distribution [260, 261] is defined as,

$$\mathcal{W}_f(\boldsymbol{u}, \boldsymbol{v}) = \mathcal{F}^{-1}_{\boldsymbol{v}'}\big[f(\boldsymbol{u} + \boldsymbol{v}') \cdot f^*(\boldsymbol{v}')\big] \ . \tag{11}$$

Using this definition of a Wigner distribution function in 11, it can be shown that the $\mathcal{H}$-array is the product of two Wigner distributions corresponding to the probe $\mathcal{W}_p$ and object $\mathcal{W}_o$, *i.e.*,

$$\mathcal{H}(\boldsymbol{k}_\mathrm{p}, \boldsymbol{r}_\mathrm{d}) = \mathcal{W}_P(-\boldsymbol{k}_\mathrm{p}, \boldsymbol{r}_\mathrm{d}) \cdot \mathcal{W}_O(\boldsymbol{k}_\mathrm{p}, \boldsymbol{r}_\mathrm{d}) \ , \tag{12}$$

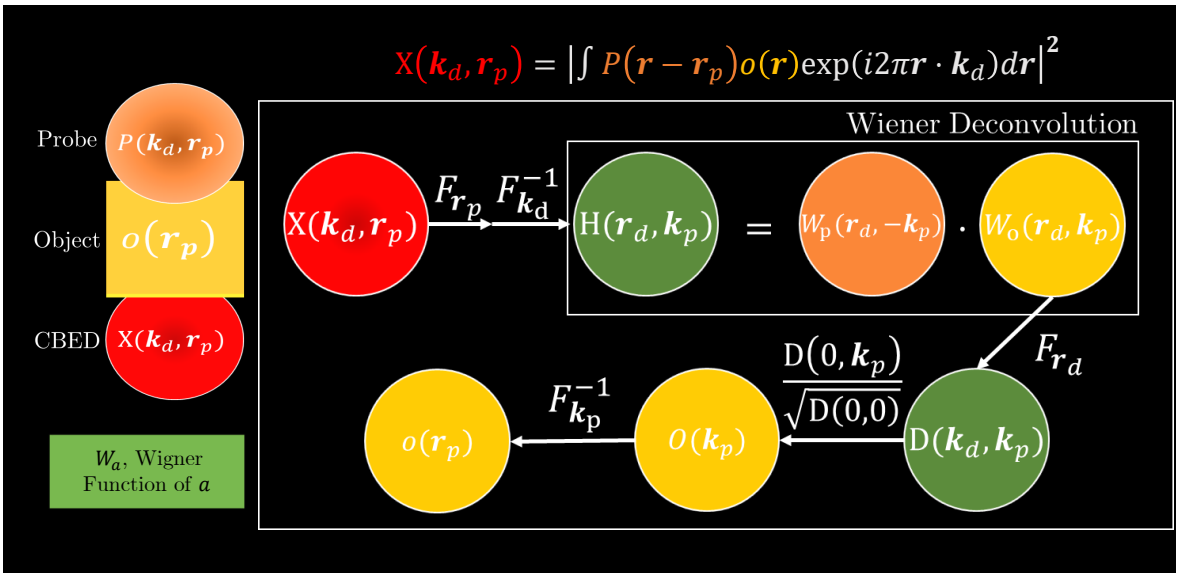

Figure 6.3: **Workflow of the WDD algorithm.** The 4-D STEM data undergoes multiple Fourier transforms, with the key step involving the Weiner deconvolution to separate the probe and object functions.

where $\mathcal{W}_P(\boldsymbol{k}_\mathrm{p}, \boldsymbol{r}_\mathrm{d})$ is estimated as the initial probe parameters.

The Wigner distribution of the object transfer function in the reciprocal space can then be computed by a Wiener deconvolution routine, with the inclusion of a small constant $\epsilon > 0$ to avoid division by zero, as

$$\mathcal{W}_O(\boldsymbol{k}_\mathrm{p}, \boldsymbol{r}_\mathrm{d}) = \frac{\mathcal{W}_P^*(-\boldsymbol{k}_\mathrm{p}, \boldsymbol{r}_\mathrm{d})\mathcal{H}(\boldsymbol{k}_\mathrm{p}, \boldsymbol{r}_\mathrm{d})}{|\mathcal{W}_P(-\boldsymbol{k}_\mathrm{p}, \boldsymbol{r}_\mathrm{d})|^2 + \epsilon} \quad . \tag{13}$$

Once $\mathcal{W}_O(\boldsymbol{k}_\mathrm{p}, \boldsymbol{r}_\mathrm{d})$ is computed in (13), it follows that

$$O^*(\boldsymbol{k}_\mathrm{d}) \cdot O(\boldsymbol{k}_\mathrm{p} + \boldsymbol{k}_\mathrm{d}) = \mathcal{L}(\boldsymbol{k}_\mathrm{p}, \boldsymbol{k}_\mathrm{d}) \coloneqq \mathcal{F}_{r_\mathrm{d}}\left[\mathcal{W}_O(\boldsymbol{k}_\mathrm{p}, \boldsymbol{r}_\mathrm{d})\right] \quad , \tag{14}$$

where $O(\boldsymbol{k}_\mathrm{p}) = \mathcal{F}_{r_\mathrm{p}}\left[o(\boldsymbol{r}_\mathrm{p})\right]$ is the Fourier transform of the object transfer function as a function of the spatial frequency of the probe location. It is clear from (14) that $|O(\boldsymbol{0})|^2 = \mathcal{L}(\boldsymbol{0}, \boldsymbol{0})$; and therefore,

$$O(\boldsymbol{k}_\mathrm{p}) = \frac{\mathcal{L}(\boldsymbol{k}_\mathrm{p}, \boldsymbol{0})}{\sqrt{\mathcal{L}(\boldsymbol{0}, \boldsymbol{0})e^{j\theta_0}}} \quad , \tag{15}$$

where θ_0 is the phase of the Fourier transform of the object transfer function at $\boldsymbol{k}_\mathrm{p} = \boldsymbol{0}$. Finally, an inverse Fourier transform on $O(\boldsymbol{k}_\mathrm{p})$ yields the object transfer function in the probe location coordinates:

$$o(\boldsymbol{r}_\mathrm{p}) = \mathcal{F}_{k_\mathrm{p}}^{-1}\left[O(\boldsymbol{k}_\mathrm{p})\right] \quad . \tag{16}$$

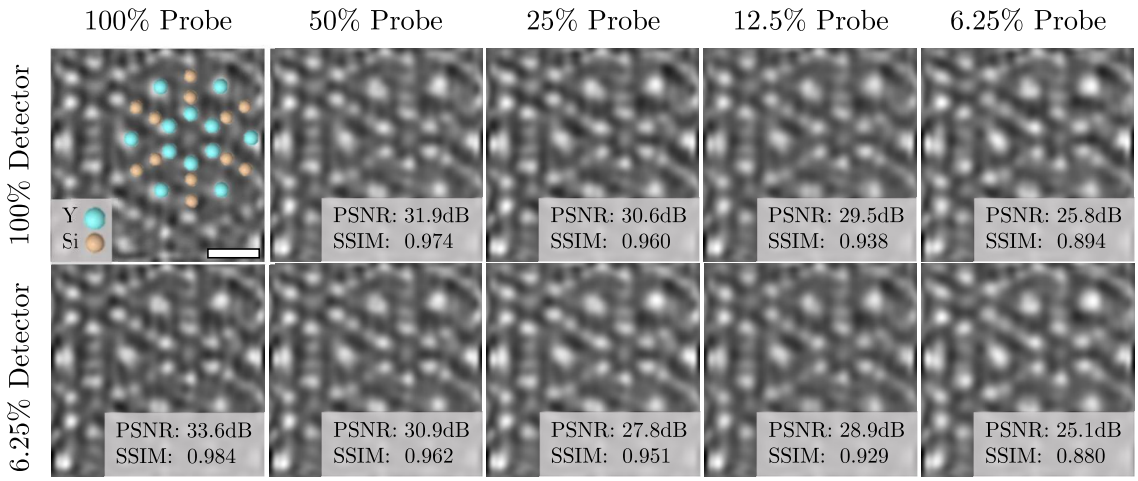

Figure 6.4: **Visual comparison of ptychographic phase retrieval quality for different probe sub-sampling and detector down sampling ratios.** The reference data is the full data-set passed through the BPFA algorithm (top row, leftmost column). The scale-bar indicates 5Å.

Note that the term $e^{j\theta_0}$ in (15) causes a global relative phase shift in the estimation of the Fourier transform of the object transfer function, equivalent to a spatial shift in real space. Without loss of generality, it is typical to set $\theta_0 = 0$. Furthermore, note that the estimated object transfer function recovered using the WDD in (16) is a function of $\boldsymbol{r}_{\mathrm{p}}$, *i.e.*, the real space coordinates of the probe location. Therefore, regardless of the number of detector pixels, the WDD estimation of the object transfer function has the same dimensionality as the scanning grid.

6.3 Results

This section presents results of applying the compressive 4-D STEM methods described above to two datasets. Firstly to a simulated CS experiment applied to experimental yttrium silicide data, and secondly experimentally sub-sampled 4-D data of a layered bismuth structure.

6.3.1 Experimental simulated compressed 4-D STEM of yttrium silicide

In order to model experimental acquisition, an experimental 4-D STEM data-set of Y_5Si_3 was used (with all scan positions) and applied random sub-sampling of the probe positions and down sampling of the CBED patterns. The experiment was carried out using a 100kV acceleration voltage, a 30mrad convergence semi-angle, and a scan-step of 0.108Å. Y_5Si_3 is an electride framework composed of cation and anion sublattices. These sublattices have a net positive electric charge which are balanced by loosely bonded, interstitial anionic electrons [208]. Y_5Si_3 has been proposed as a low Schottky barrier material for *n*-type silicon semiconductors due

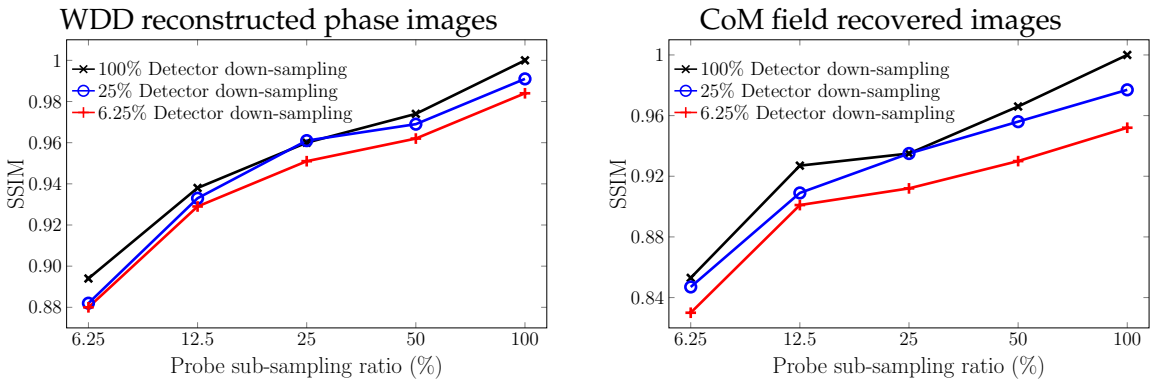

Figure 6.5: **SSIM of phase and CoM field recoveries with respect to probe and detector sampling ratios.** As the probe sub-sampling ratio increases, the quality of the phase and CoM field recovery increases. However, there is only a small difference in the image qualities as the detector down sampling ratio is decreased. This indicates significant redundancy within the 4-D data-set, which can be omitted through detector down sampling and probe sub-sampling. Example images of the phase images from this experiment are shown in Fig. 6.4.

to its low Schottky barrier height of 0.27eV [209]. It has also been recently proposed as an encapsulation material for radioactive volatile products within nuclear fission reactors [210]. Readers are referred to section 5.3.3 as well as the work of Zheng *et al.* [208] for more details on the sample.

In this study, probe sub-sampling ratios $M_\mathrm{p}/N_\mathrm{p} \in \{6.25, 12.5, 25, 50, 100\}\%$ were applied, as well as detector down sampling ratios $M_\mathrm{d}/N_\mathrm{d} \in \{6.25, 25, 100\}\%$. LAADF and annular BF (ABF) [265] virtual detector images, $(r_\mathrm{i},\ r_\mathrm{o}) = (30,\ 60)$ mrad and $(r_\mathrm{i},\ r_\mathrm{o}) = (10,\ 22)$ mrad were simulated together with DPC images with $(r_\mathrm{i},\ r_\mathrm{o}) = (10,\ 22)$ mrad and $(\theta,\ \delta) = (3\pi/4,\ \pi/2)$ rad. In addition, the recovered ptychographic phase images are calculated (Fig. 6.5 (left)). For this there are a number of analytical and iterative algorithms [254–259] that recover the complex ptychographic wave-function, and here a modification of the Wigner distribution deconvolution (WDD) algorithm [219, 260–264] is used as given within the *ptychoSTEM* package for MATLAB [228].

Fig. 6.5 (left figure) shows the quality of the ptychographic phase (using the structural similarity index measure (SSIM) [197] as our chosen metric) with respect to different probe sub-sampling and detector down sampling ratios. There is only a small degradation in the quality as the sampling at the detector is decreased; this implies the detector is over-sampled. Further observations show that probe sub-sampling can be used with BPFA to recover visually identical results in the phase recovery.

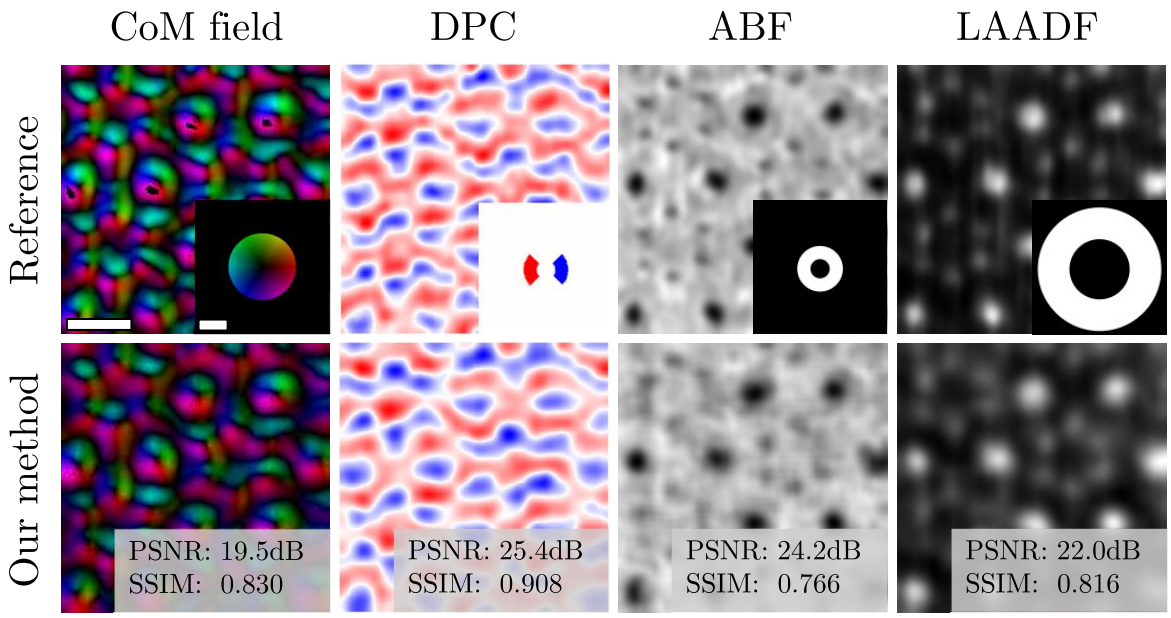

Figure 6.6: **Visual comparison of images recovered from sub-sampled 4-D STEM data.** CoM field, DPC, ABF, and LAADF images for 6.25% probe sampling and 6.25% detector down sampling after inpainting. The reference data is the full data passed through the BPFA algorithm (top row). The PSNR and SSIM values are overlaid, the spatial scale bar indicates 5Å, and the detector scale bar indicates 30 mrad.

Similarly, Fig. 6.5 (right figure) shows a comparison of the quality of CoM field analysis as a function of sub-sampling ratio, where visually identical results are achieved with respect to the reference data. Comparing the plots for the phase and CoM field recoveries shown in Fig. 6.5 suggests that ptychographic phase recovery is more robust in this case. This is possibly due to the fact that the WDD operates on a full 4-D data-set, while the CoM field is computed from individual CBED patterns.

Fig. 6.6 is a direct image comparison between the reference data and reduced sampling data ($M_\mathrm{p}/N_\mathrm{p} = M_\mathrm{d}/N_\mathrm{d} = 6.25\%$) when applied to CoM field analysis, DPC, ABF, and LAADF. It is clear that there is very little difference in the quality of the images from a visual perspective, and this is supported comparison of the corresponding peak signal-to-noise (PSNR) and SSIM values corresponding to each. Fig. 6.4 is a visual comparison of the data in Fig. 6.5 (left). As can be seen, the recovered phase data is almost indistinguishable, with all showing the expected location of yttrium and silicon atoms.

The results demonstrate the inherent redundancy within the 4-D STEM data-set. By utilising inpainting algorithms, it is possible to discard over 99.6% (see Fig. 6.4 bottom-right) of the original data-set whilst still recovering qualitatively identical results in the reconstructed

phase, CoM field, DPC and VD images, to those obtained from processing the full data-set. An example of these same methods applied to CdTe-Si interface are shown in section A1.2, Fig. A1.3.

6.3.2 Experimentally acquired compressed 4-D STEM

Having simulated compressive 4-D STEM and the possible quality of recovery, the next logical step is to test the method in practice. There are several challenges which must be overcome to achieve a sub-sampled 4-D STEM dataset acquisition, and it has taken many trials before arriving at a suitable method.

A JEOL 2100F (Cs corrected) equipped with a Direct Electron DE-16 camera and Direct Electron FreeScan scan generator is used as the 4-D STEM acquisition tool. In order to perform a standard 4-D STEM acquisition, the camera triggers the scan controller as to when the probe should be moved from its position to the next. This creates a send-receive-acquire triplet, between the camera-scan generator-microscope respectively. The hierarchy is important, since the timings are set by the sender. By changing the order, one can enforce which hardware is dominant and then by manipulating the parameters, can control where the probe is and which frames to capture.

The workflow is not too dissimilar to that of standard CS-STEM, however the inclusion of the camera involves an added layer of complexity. Firstly, the desired scanning pattern must be loaded into the FreeScan software, and the camera cooled and inserted. Given the custom scan pattern, the camera must be the signal receiver, not the sender; instead the scan generator must be the sender. The next important step is to ensure that the scan frequency matches the acquisition frequency of the camera so that the camera captures at the same rate as the scanning probe moves position.

Once complete, a compressive 4-D STEM data will be in a 3-D format, where each layer corresponds to a certain probe coordinate. To create a 4-D dataset, the diffraction patterns can be stored sparsely (*i.e.*, as a diffraction pattern with a corresponding position) or appended to a zero valued 4-D array, where the diffraction patterns from a sampled probe replace their zero valued corresponding diffraction pattern within the array, as depicted in Fig. 6.7.

This method was used to acquire 4-D STEM data of a layered bismuth structure collecting 12.5% of the probe locations in a UDS regime with a scan-step of 0.106Å over a 512×512

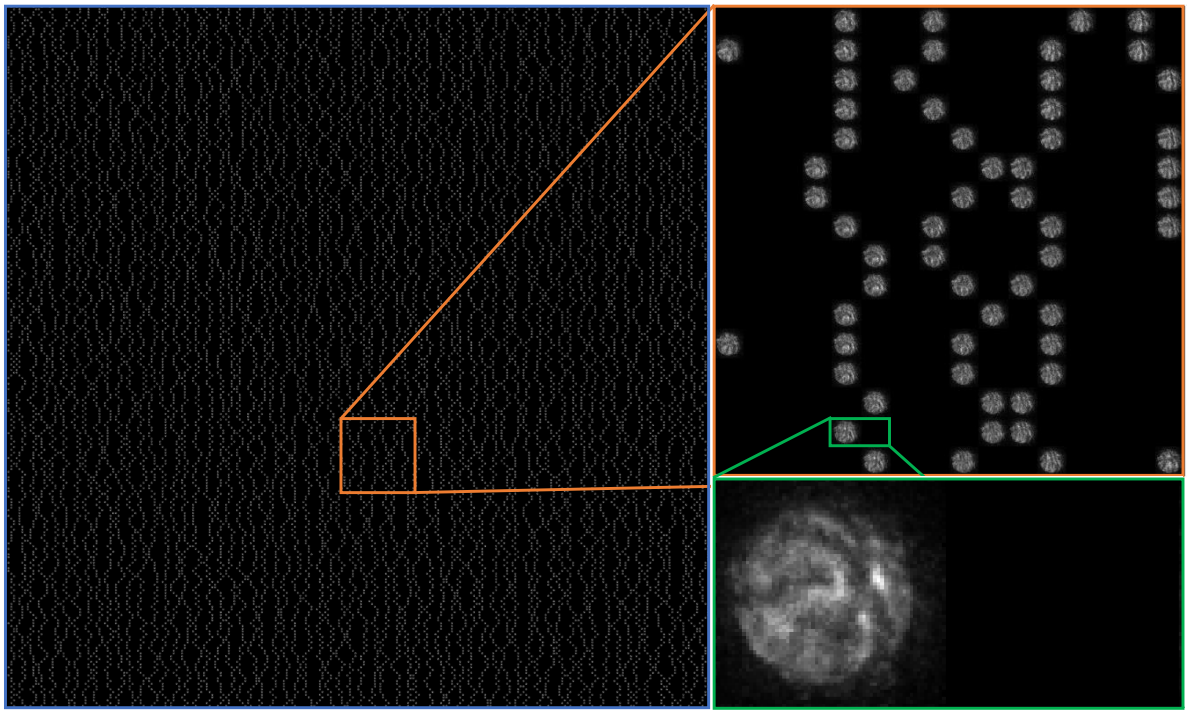

Figure 6.7: **Sub-sampled 4-D STEM of experimental data.** An example data array of 4-D STEM as acquired in experiment using a 25% line-hop sampling mask.

grid. The CBED patterns were collected using a Direct Electron DE-16 camera operating at $1,400$fps with a 256×256 readout region, camera length of 4cm, convergence semi-angle of 25mrad and an accelerating oltage of 200kV. The sample was prepared as a lamella using a FIB, then oriented onto the [110] axis to see the layers. The data was inpainted using the method described in section 6.2.1. The resulting object phase reconstruction and projected charge density distribution are shown in Fig. 6.8.

The results in Fig. 6.8 are promising and show that it is indeed possible to acquire and inpaint sub-sampled 4-D STEM data, as well as perform all the same analysis as would be expected. The streaking in the images is due to sample drift and vertical stage drift, and is not an artefact of the inpainting process. The total acquisition time was approximately 23s, which is equivalent to collecting all diffraction patterns using a camera running at $11,200$ fps. This could be significantly improved if a smaller read-out region was selected, however there were instabilities when the camera length was reduced to $\sim$ 2cm, meaning the CBED could not be made smaller on the detector using the projector lens system.

Given the success of applying this method to a near 20 year old instrument (the JEOL JEM 2100F Cs at Liverpool), with the correct hardware installed on a more modern system would

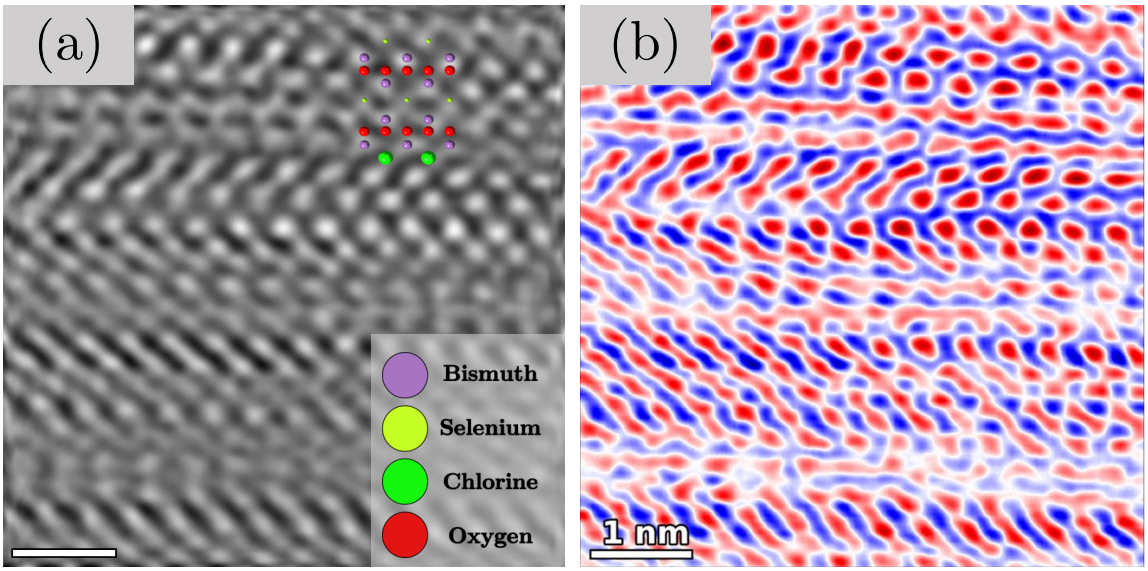

Figure 6.8: **Ptychographic reconstruction and projected charge density of the experimentally acquired compressive 4-D STEM data.** (a) Ptychographic reconstruction using the WDD and (b) the projected charge density calculated using the divergence of centre-of-mass. Scale bar indicates 1nm.

significantly improve the quality of the data. There is still a lot of work to do regarding optimisation of the technique, however this demonstration has shown that probe sub-sampling for 4-D STEM can be applied in practice.

6.4 Conclusions

The application of compressive sensing to 4-D STEM has significant benefits for upgrading existing microscopes through increased acquisition speeds, as well as the reduction of fluence and potential for beam damage. Given the inherent redundancy in 4-D STEM data, it is proposed that even lower sampling ratios could be employed using a multi-dimensional recovery algorithm, rather than performing sequential 2-D inpainting. The benefit of this is that by using a multi-dimensional recovery algorithm it is possible to leverage more data during the training process as well as the similarity between virtual images during the recovery step.

To further improve acquisition speeds, it may be possible to also include sparse detector sampling, analogous to probe sub-sampling, followed by inpainting the 4-D STEM data-set with minor modifications to the acquisition model. This could further increase acquisition speeds by assuming that each pixel has a fixed read-out time, and potentially allow for multiple 4-D STEM data-sets to be acquired rapidly.

Furthermore, it is postulated that time-resolved 4-D STEM is now not limited by the detector read-out speed, but can instead be acquired through reduced sampling strategies.

7 | Other works

7.1 Introduction

The main research foci of this thesis have been presented in the previous chapters, addressing the application of compressive sensing to STEM simulation and 4-D STEM. However, there are several other works which were performed alongside the main research topics. This chapter aims to summarise those works, their importance, and the learning outcomes.

7.2 Improving ePIE with a sparsity promoting regularization

This section presents a novel solution to improve the reconstruction quality of iterative ptychograms using a sparsity promoting l_0-norm regularization step, which ultimately also allows for fewer probes to be scanned without significant loss of information. This work was done in collaboration with Amirafshar Moshtaghpour (RFI) and Abner Velazco-Torrejon (RFI).

7.2.1 Introduction

As discussed in section 2.3.3, 4-D STEM data can be acquired using either a focused probe (as shown in section 6) or a defocused probe. In the case of defocused probe 4-D STEM, the goal is to recover the phase of the object through a phase retrieval algorithm, leveraging the redundancy between neighbouring probe locations to update an estimate of the incident probe and the object itself. This phase retrieval process is known as ptychography, which was initially developed by Hoppe [266] and extended for electron microscopy by Gerchberg [267], and then John Rodenburg and Richard H. T. Bates who pioneered the field to where it is today. It was initially considered a super-resolution technique [260, 261] since the theoretical resolution could outperform lenses at the time, and work by Nellist *et al.* [216] was the first demonstration

of practical results evidencing this, with a roughly 3× improvement in resolution. Rodenburg and Bates' work built upon the proposed method by Hoppe [266] for phase retrieval. Since then, ptychography has become a powerful imaging method for various fields such as x-ray microscopy [268] and visible light microscopy [269], not just in electron microscopy.

There are limitations to this method. The first is the inherent redundancy in the dataset which ultimately leads to overexposure of the sample, potentially resulting in beam damage and/or long acquisition times. The second is the size of the data collected which is typically on the order of several gigabytes, which is difficult to operate on during analysis, and a limitation to data throughput. This work aims to solve these issues through an more efficient acquisition method an improved version of ePIE by imposing sparsity on the final solution.

7.2.2 Principle of iterative phase retrieval

Work by Gerchberg [270] showed one of the earliest successful approaches to iterative phase retrieval, which Fienup [271, 272] also began developing towards the late seventies and early eighties. There are various iterative ptychography algorithms which can recover the object phase as well as the probe. These include the Ptychographical Iterative Engine (PIE) [254], the extended PIE (ePIE) [255], 3-D ePIE (3PIE) [257], Difference Map (DM) [273], Maximum Likelihood (ML) [274], Relaxed Average Alternating Reflections (RAAR) [275], Nonlinear Optimisation (NL) [276], Semi-implicit Douglas-Rachford (sDR) [277], as well as various gradient descent approaches [278].

In this thesis, the ePIE is used for iterative phase retrieval. The ePIE works by assuming the following forward model for the exit wave $\psi(\boldsymbol{r}, \boldsymbol{r_p})$,

$$\psi(\boldsymbol{r}, \boldsymbol{r_p}) = P(\boldsymbol{r} - \boldsymbol{r_p}) \cdot O(\boldsymbol{r}) \ , \tag{1}$$

where the signal measured on the detector $D(\boldsymbol{k}, \boldsymbol{r_p})$ (here, detector wave) is,

$$D(\boldsymbol{k}, \boldsymbol{r_p}) = |\mathcal{F}_{\boldsymbol{r}}[\psi]|^2 \ , \tag{2}$$

which is the same forward model given in section 6, Eq. 8. The inverse problem is then defined as,

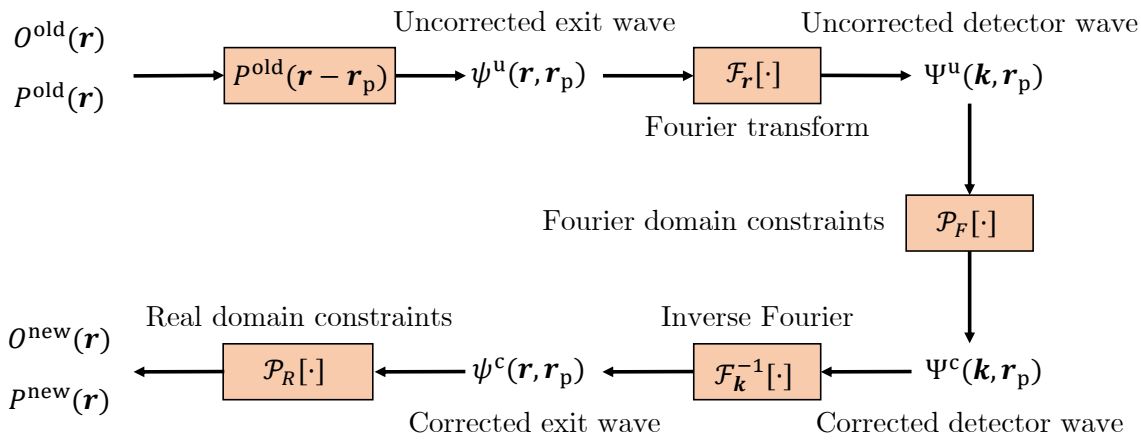

Figure 7.1: **Workflow of the ePIE algorithm.** The forward model is applied to initial estimates of the object and probe, which is then compared to the measurements in both Fourier and real domains. The final solution is the one which minimises the error between estimation and measurement.

$$\{\hat{O}(\boldsymbol{r}), \hat{P}(\boldsymbol{r})\} = \Pi(D(\boldsymbol{k}, \boldsymbol{r}_{\mathbf{p}})) \ , \tag{3}$$

where Π is a solver which takes in measured diffraction patterns and returns estimates of the object, $\hat{O}(\boldsymbol{r})$, and probe, $\hat{P}(\boldsymbol{r})$. For a non-convex problem such as this, there is no closed form solution which guarantees recovery, however it is possible to approximate a solution by the minimisation of error between the observations and the estimates.

The ePIE algorithm takes this approach to phase retrieval, and is summarised in Fig. 7.1. The algorithm begins with estimates of the object and probe in real space, with the later being approximated by considering the illumination source (although a random guess would be possible). A real space probe position is selected at random and an uncorrected exit wave is then calculated using the forward model in Eq. 1. Its Fourier transform is taken with respect to real space, and then it is compared with the measurement. This provides the first constraint in the Fourier domain, where the corrected detector wave is given as,

$$\Psi^{\text{c}}(\boldsymbol{k}, \boldsymbol{r}_{\mathbf{p}}) = \sqrt{D(\boldsymbol{k}, \boldsymbol{r}_{\mathbf{p}})} \exp\left(\angle \Psi^{\text{u}}(\boldsymbol{k}, \boldsymbol{r}_{\mathbf{p}})\right) \ , \tag{4}$$

i.e., replace the amplitude with that of the observed detector wave, but keep the phase of the uncorrected detector wave (see appendix A2.2). An inverse Fourier transform of the corrected detector wave is then taken with respect to reciprocal space, and this forms a corrected exit wave. This is then passed into the real domain constraints. The ePIE alternates between probe

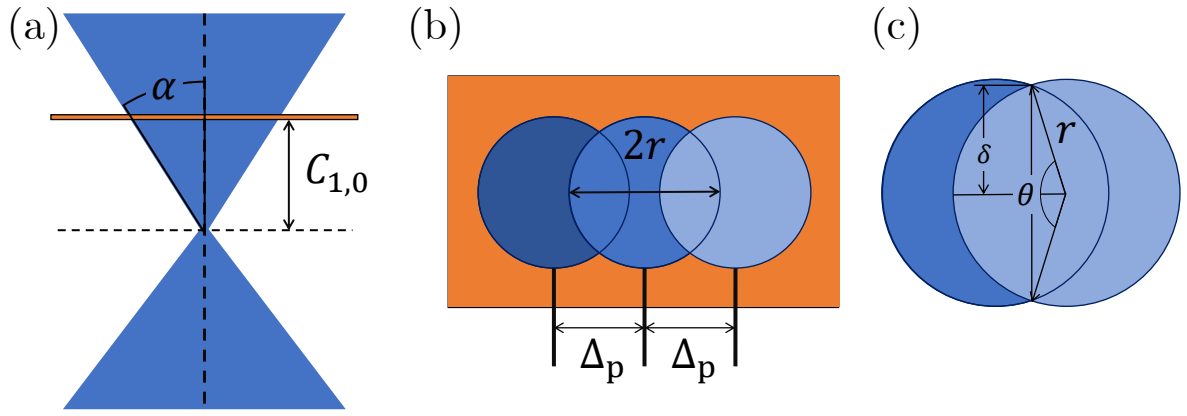

Figure 7.2: **Defocused probe set-up in STEM.** The overlap ratio can be calculated using the convergence semi-angle α, the defocus value $C_{1,0}$, and the scan-step Δ_p. (a) View perpendicular to the optical axis showing the defocus condition, (b) view from above parallel to the optical axis indicating the scan-step, and (c) geometry for calculating the probe overlap ratio.

and object updates, and here, an update step is defined as an update of the probe and object for one exit wave. An iteration is defined as the number of times the entire dataset is passed through the data. The real domain constraints are given as,

$$P^{\text{new}}(\boldsymbol{r}) = P^{\text{old}}(\boldsymbol{r}) + L_\beta \frac{O^{\text{old}*}(\boldsymbol{r}-\boldsymbol{r}_\text{p})}{\max_{\boldsymbol{r}}|O^{\text{old}}(\boldsymbol{r}-\boldsymbol{r}_\text{p})|^2}(\psi^\text{c}(\boldsymbol{r},\boldsymbol{r}_\mathbf{p}) - \psi^\text{u}(\boldsymbol{r},\boldsymbol{r}_\mathbf{p})) \tag{5}$$

$$O^{\text{new}}(\boldsymbol{r}) = O^{\text{old}}(\boldsymbol{r}) + L_\alpha \frac{P^{\text{new}*}(\boldsymbol{r}-\boldsymbol{r}_\text{p})}{\max_{\boldsymbol{r}}|P^{\text{new}}(\boldsymbol{r}-\boldsymbol{r}_\text{p})|^2}(\psi^\text{c}(\boldsymbol{r},\boldsymbol{r}_\mathbf{p}) - \psi^\text{u}(\boldsymbol{r},\boldsymbol{r}_\mathbf{p}))\ , \tag{6}$$

where L_α and L_β are learning rates. These constraints are derived in the appendix A2.2.

7.2.3 Importance of probe overlap

Defocused probe electron ptychography is a special case, whereby the electron probe formed in a STEM is deliberately defocused to increase the field-of-view for each incident electron probe. By doing this, the scan-step can be increased whilst retaining the same illuminated area, thus decreasing electron fluence. The scan-step and defocus should be set so that the overlap between neighbouring probe locations is sufficient [279].

The radius of the probe at the sample plane, $r \in \mathbb{R}$, is given as

$$r = C_{1,0}\tan(\alpha)\ . \tag{7}$$

The area of the isosceles angled triangle A_{T} seen in Fig. 7.2(c) is given as,

$$\begin{aligned} A_{\mathrm{T}} &= \delta\frac{\Delta_{\mathrm{p}}}{2} \\ &= \frac{\Delta_{\mathrm{p}}}{2}\sqrt{r^2 - \left(\frac{\Delta_{\mathrm{p}}}{2}\right)^2} \ , \end{aligned} \tag{8}$$

and the area of the segment subtended by θ, A_{S} is given as,

$$\begin{aligned} A_{\mathrm{S}} &= \frac{\theta}{2}r^2 \\ &= r^2\cos^{-1}\left(\frac{\Delta_{\mathrm{p}}}{2r}\right) \ . \end{aligned} \tag{9}$$

The area in the overlap region is then simply twice the difference between the area of the segment and the area of triangle *i.e.*, $A_{\mathrm{O}} = 2(A_{\mathrm{S}} - A_{\mathrm{T}})$. The overlap ratio R is then calculated as A_{O} divided by the area of the probe illuminating the sample,

$$R = \frac{2r^2\cos^{-1}\left(\frac{\Delta_{\mathrm{p}}}{2r}\right) - \Delta_{\mathrm{p}}\sqrt{r^2 - \left(\frac{\Delta_{\mathrm{p}}}{2}\right)^2}}{\pi r^2} \ . \tag{10}$$

The amount of overlap between probes has a direct effect upon the observed error in the recovered phase [279]. This is because there are more updates for each real position $\boldsymbol{r}$ where $\boldsymbol{r}$ is shared between more probes. The estimate of the phase at $\boldsymbol{r}$ in $O(\boldsymbol{r})$ is therefore more consistent with the observations. Typical overlap ratios range from $70-90\%$, implying $8-10$ probes contain the same real space position. In terms of beam damage and redundancy, this is not efficient, hence a new approach is considered which aims to reduce the amount of required overlap.

7.2.4 The l_0 regularized ePIE (LoRePIE)

In order to overcome the overlap limit, an l_0-norm regularization step is included during the real domain constraints. The l_0-norm regularization, or hard-thresholding, imposes a sparsity constraint on the transformation of an object into some sparsity basis, which in this case is the discrete cosine transform basis. For an object X, a basis transformation operator $\mathcal{A}$, and a threshold operator H_λ, l_0-norm regularization is defined as,

Parameter	Test 1	Test 2	Test 3	Test 4	Test 5
Sampling ratio (%)	100	25	11.11	6.25	4
Overlap ratio (%)	85	70	56	43	30
Electron fluence (e^{-}Å^{-2})	22.8	5.8	2.6	1.5	<1

Table 7.1: **Parameters for testing the effectiveness of LoRePIE.** .

$$\begin{aligned}\hat{X} &= \mathcal{A}^{-1}\Big[H_\lambda\big[\mathcal{A}[X]\big]\Big] \\ &= \mathcal{D}\big[X\big] \ , \end{aligned} \tag{11}$$

where λ defines the strength of the threshold, and $\mathcal{D}$ is the general regularization operator in this case. In the case of l_0, λ is the number of coefficients which are to be non-zero in following the application of H_λ, with the remaining values set to zero.

LoRePIE utilizes $\mathcal{D}$ during the real domain constraints acting on the object update step, *i.e.*,

$$O^{\text{new}}(\boldsymbol{r}) = \mathcal{D}\left(O^{\text{old}}(\boldsymbol{r}) + \alpha\frac{P^{\text{new}*}(\boldsymbol{r}-\boldsymbol{r}_{\text{p}})}{\max_{\boldsymbol{r}}|P^{\text{new}}(\boldsymbol{r}-\boldsymbol{r}_{\text{p}})|^2}(\psi^{\text{c}}(\boldsymbol{r},\boldsymbol{r}_{\mathbf{p}})-\psi^{\text{u}}(\boldsymbol{r},\boldsymbol{r}_{\mathbf{p}}))\right) \ , \tag{12}$$

which is one fast and simple process on the estimated object function, with only one additional parameter to tune.

7.2.5 Results

To test the effectiveness of LoRePIE versus ePIE, a 4-D STEM dataset of double-layered rotavirus particles was examined [220].

Acquisition parameters

The dataset was acquired with an acceleration voltage of $300kV$ in a defocused probe regime, with the defocus estimated as $-13\mu m$. The scan step was set to 31.25Å and 127×127 diffraction patterns were acquired in the far-field using a JEOL ARM 300CF equipped 256×256 Medipix3 direct electron detector. Given a convergence semi-angle of 1.03mrad, the sampling on the diffraction patterns was 0.023mrad and an 85% overlap ratio. The dwell time was 1ms using a 4pA beam current, hence the theoretical electron fluence was $22.8e^{-}$Å^{-2}.

Simulation set-up

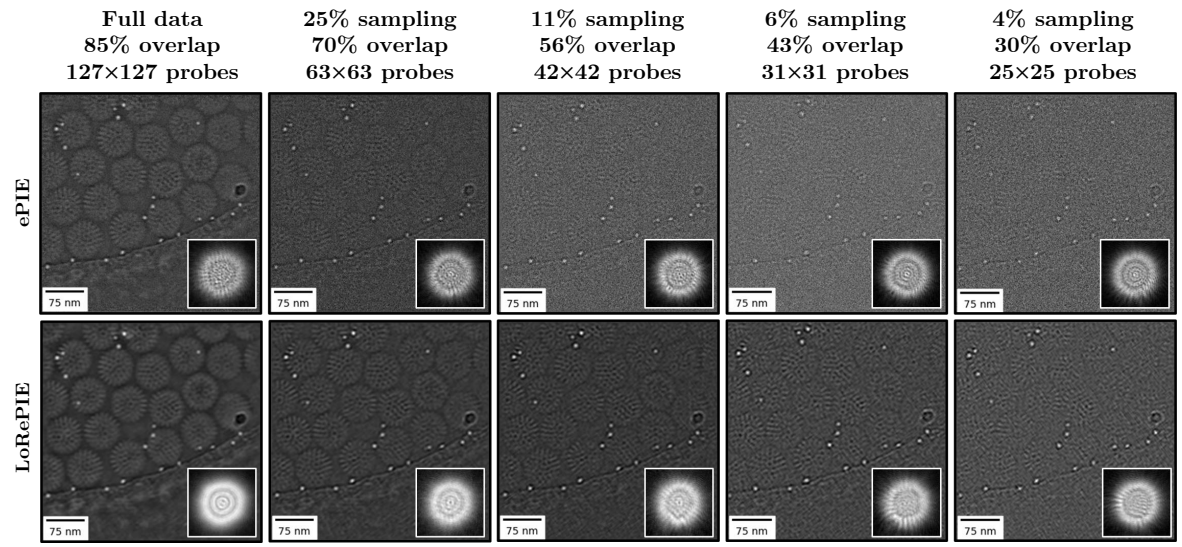

Figure 7.3: **Comparing LoRePIE to ePIE.** ePIE results (top row) for each sampling ratio, and LoRePIE results (bottom row) for the same parameters. LoRePIE returns visually improved object phase images compared to ePIE at all sampling ratios. Probe amplitudes are overlaid for reference. 4-D STEM data courtesy of Professor Peng Wang.

Simulated down-sampling of the probe locations were performed ranging from 100% sampling to 4%. The simulation parameters are given in table 7.1. The electron fluence values are calculated by considering the probe radius on the sample with the given parameters of the acquisition (dwell time, beam current). The calculations and simulations are given in the appendix A1.3.

The results in Fig. 7.3 show visually improved results for the object phase reconstructions. The images appear less noisy, and are significantly improved for lower overlap ratios. Furthermore, the probe estimates are significantly cleaner despite LoRePIE not changing the probe update step. This is likely due to a less-noisy object which in turn improves the overall probe function.

7.2.6 Conclusions

The results presented here show a novel solution to the probe overlap problem in defocused probe 4-D STEM. By using a simple regularization technique, the reconstructed object phase is significantly improved at low overlap ratio.

This solution could be extended to other regularization techniques such as the use of a different sparsity basis such as wavelet basis, or a dictionary learning or deep learning strategy. It is clear that simple computational imaging techniques can drastically improve results for

any existing iterative technique.

7.3 Characterisation of a CdTe-Si interface using 4-D STEM

This section is a brief summary of work done with Giuseppe Nicotra (CNR-IMM, Catania, Italy), Gianfranco Sfuncia (CNR-IMM, Catania, Italy), Daniel Nicholls (SenseAI Innovations Ltd, UK), and Sivananthan Laboratories (Bollingbrook, IL, USA). The goal of this collaboration was to both optimise the 4-D STEM acquisition process at CNR-IMM to increase speed, as well as characterise the CdTe-Si interface using 4-D STEM techniques. The microscope used was a JEOL JEM ARM 200F, equipped with a Gatan K2 Summit detector running at 400fps.

7.3.1 Objective

As discussed in section 6, 4-D STEM is a powerful data acquisition method to analyse complex materials such as low mass elements, biological samples, and defects, and interfaces. In 4-D STEM, a convergent beam electron diffraction (CBED) pattern is acquired at each probe location in a raster scan to collect a larger range of scattering information, resolved on the diffraction plane. Using this information, it is possible to generate different signals such as annular bright field, centre of mass (CoM) and retrieve the phase of the object through ptychography.

In this study, 4-D STEM was used to investigate the defect formation mechanisms of CdTe grown directly on Si produced by Sivananthan Laboratories through molecular beam epitaxial growth. The goal of this study was to use 4-D STEM to attempt to simultaneously resolve the CdTe and Si dumbbells through ptychography, a task which was not possible using traditional imaging methods on the same microscope. This study also used CoM analysis to understand the deflection of the electron beam caused by the interface. In combination, these analyses may highlight the dislocations formed to relax the mismatch strain [280], which plays a significant role in the performance of CdTe/Si substrate as an infrared detector.

7.3.2 Method

To acquire the dataset, a JEOL JEM-ARM200F at the Institute for Microelectronics and Microsystems in Catania was used. The microscope was aligned with a focused STEM probe at 200kV and a convergence semi-angle of 30 mrad. In this study, a 4-D STEM dataset with

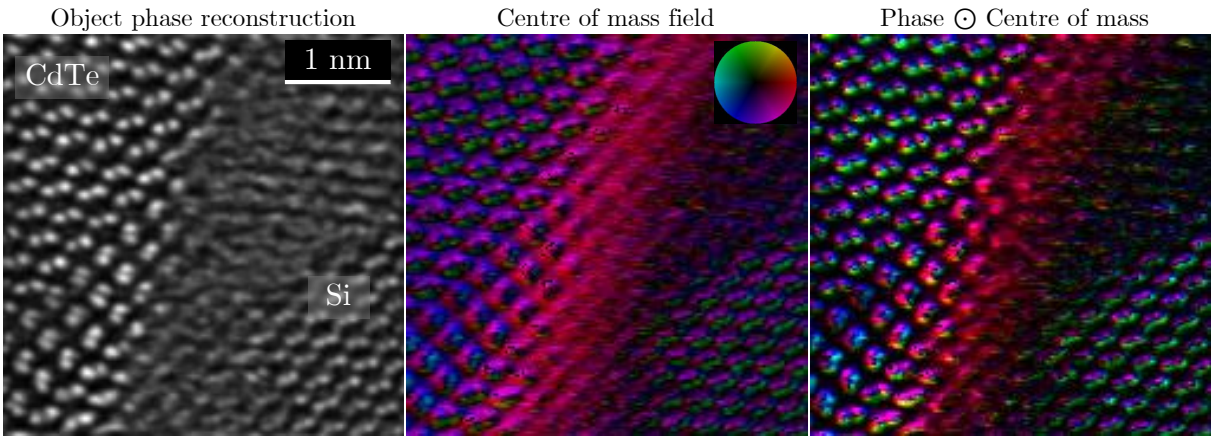

Figure 7.4: **4-D STEM results for CdTe-Si interface.** WDD reconstruced object phase (left), centre of mass field (centre), and product of the phase image with the centre of mass field (right).

diffraction patterns of size 128x128 was collected over a raster scan containing 170x170 probe positions using a scan-step of 0.25Å. To recover the phase image, the Wigner Distribution Deconvolution (WDD) algorithm was chosen which is well established technique for focused probe 4-D STEM phase retrieval. The centre of mass field was calculated using the method described in section 6 using the MAT 4-D STEM toolbox presented in section 7.4.

7.3.3 Results

As the results in Fig. 7.4 show, the CdTe and Si are well resolved using the WDD to recover the object phase. Furthermore, the CoM analysis proved equally useful for determining the beam deflection. As can be seen, the interface shows a direct preferential electric field orientation. This could be due to the difference in band gap between the CdTe and Si, potentially interface states, or interstitials or impurities within the interface. The reason shall be investigated at a later date as more data is required to validate the observation. Furthermore, the CdTe showed twinning at the interface, indicative of shear stress response at the interface as hypothesised above. Through ptychography, it was possible to simultaneously resolve both the Si and CdTe dumbbells, and resolve atomic phase information at the boundary. This demonstrates the power of focused probe ptychography to enhance the capabilities of STEM to resolve objects which are either irresolvable or difficult to resolve in traditional imaging techniques.

7.4 MAT 4-D STEM

Due to the amount of data that is typically acquired in a 4-D STEM acquisition, the different types of data analysis which can be performed are far greater than that of standard 2-D imaging. As discussed in section 6, these analysis types can reveal different properties depending on the virtual detector geometry, or the phase retrieval.

There are various 4-D STEM analysis toolboxes which aim to encompass all the different types within a package, such as the py4DSTEM [281, 282] and LiberTEM [283]. py4DSTEM has become especially popular due to its large community of users, and LiberTEM for teaming up with some camera manufacturers to read in their native file types. In the case of py4DSTEM, there are extensive tutorials which guide users through, but the analysis still requires a knowledge of python to manipulate data correctly. In order to overcome this, a simple MATLAB based GUI was developed for 4-D STEM analysis, called MAT 4-D STEM. MAT 4-D STEM aims to support MATLAB users, as well as those with little knowledge of programming languages to interactively investigate their 4-D STEM data. The MAT 4-D STEM also has scripting functionality, as well as support for ptychography.

7.4.1 MAT 4-D STEM GUI

The GUI is designed to be simple and modular, promoting reproducible results and analysis. The GUI is composed of four items, *Experimental*, *Detector*, *Visualise*, and *Inpaint*. In this section, each shall be explained in more detail.

Experimental

4-D STEM analysis requires knowledge of the experimental set-up such as the accelerating voltage, convergence semi-angle, and scan-step. For this reason, the Experimental tab is the first to be updated with respect to the MAT 4-D STEM GUI.

Experimental parameters can be saved into a .mat file which can be stored and loaded with a time stamp, encouraging reproducible results and faster analysis. Users can also prepare data with built in de-noising tools such as filters or regularisation. There is also a custom analytical de-noising tool for data which is corrupted by Poisson noise, although its efficacy has not been tested to a full extent. There is also the option to re-orient data, as well as crop, bin, down-sample, and pad diffraction patterns. Padding can also be applied to real-space to

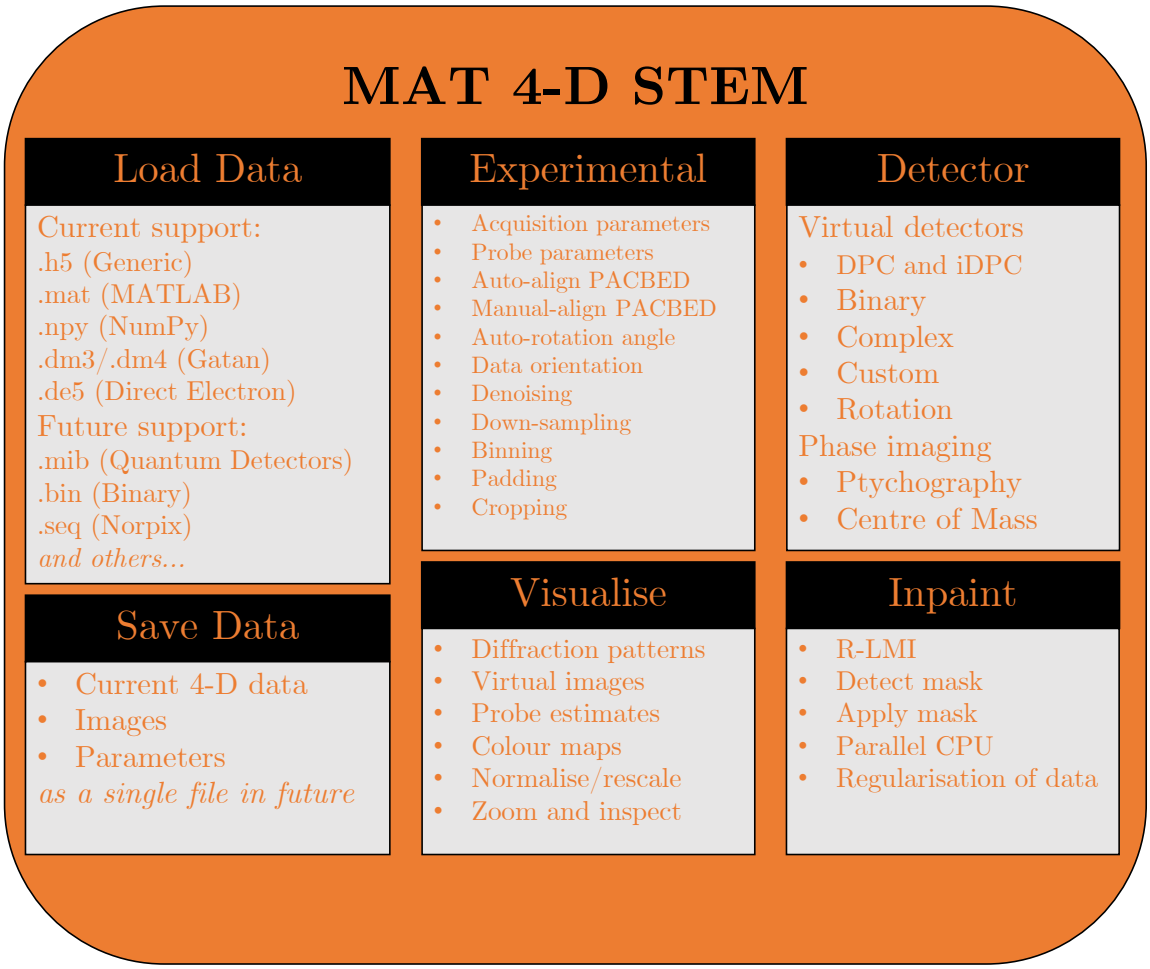

Figure 7.5: **Overview of MAT 4-D STEM support and analysis tools.** MAT 4-D STEM is currently a work-in-progress, but already support preliminary analysis modes as well as multiple data types.

reduce wrapping artefacts.

The position averaged CBED (PACBED) can be aligned by hand using a built in *Draw circle* function, or the PACBED can be automatically aligned using a binary threshold. The projector lens rotation can also be automatically corrected using the divergence and curl of DPC data. Probe parameters/aberrations can also be updated without reinitialising the code, especially useful for parameter tuning within the WDD.

Detector

Virtual detectors, such as those described in section 6, allow a range of flexibility to 4-D STEM analysis that cannot be achieved with fixed radial detectors in STEM mode. MAT 4-D STEM allows users to define inner and outer detector angles, rotation of DPC/iDPC detectors, as well as control over the sector angle in DPC detectors. The CoM collection angle can also be

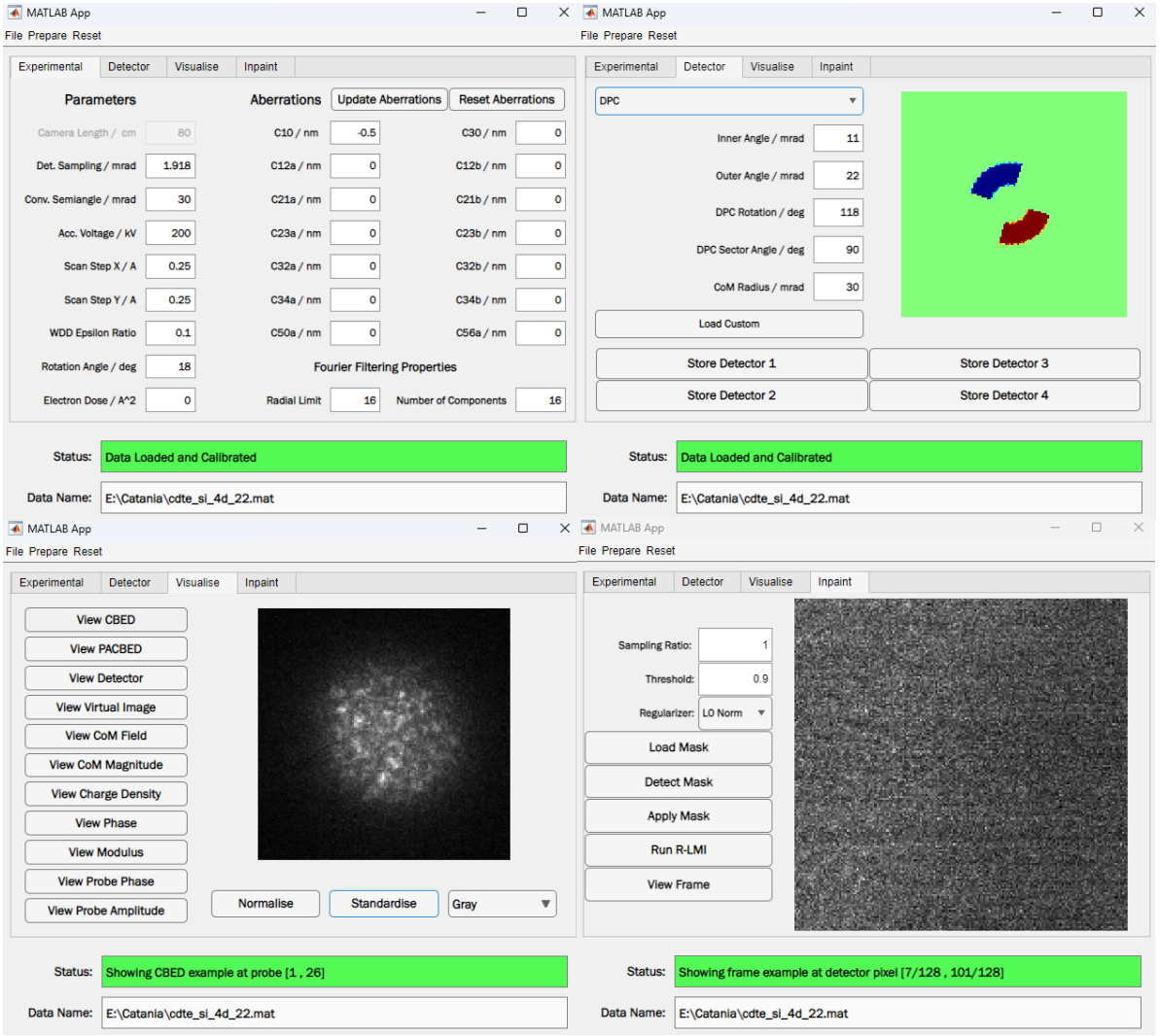

Figure 7.6: **The four key components of MAT 4-D STEM.** MAT 4-D STEM is designed with simplicity of use. The four key components are Experimental (top left), Detector (top right), Visualise (bottom left), and Inpaint (bottom right). These elements create a modular design, ideal for simple, reproducible analysis.

adjusted to exclude or include certain signals, and the detectors can be easily stored if there are preferential detector types. Detectors can be easily visualised as they are updated, and they are automatically centred given the centring of the PACBED.

MAT 4-D STEM also includes an adaptation of the WDD algorithm allowing for ptychographic phase retrieval. In future, iterative techniques shall also be implemented for further use cases. MAT 4-D STEM can also perform CoM analysis to generate projected electric field and projected electric charge density. In all cases, MAT 4-D STEM will eventually have GPU support for faster analysis using the mex compiler.

Visualise

Data is visualised within the *Visualise* tab. Individual CBED patterns can be observed, as well as the PACBED, virtual (detector) images, CoM analysis, and ptychographic reconstructions. The images can be manipulated through normalisation for saving, standardisation (rescales all data between 0 and 1), as well as various colour maps for enhancing visualisation.

Inpaint

The final tab is based on inpainting 4-D STEM data using R-LMI from section 3.2.2. A sub-sampled 4-D STEM data can be loaded in and the mask can be detected, then the data inpainted using automatic parameters determined by the input. If the data is already fully sampled, the regularisation tool can be used to denoise the data, and users can select whether to use a l_0-norm or l_1-norm regularisation. This section is fully parallelised on the CPU, and is relatively fast for data processing/inpainting.

7.4.2 Future of MAT 4-D STEM

MAT 4-D STEM has been developed from custom 4-D STEM analysis scripts, and as such is still *under construction*. In future, MAT 4-D STEM will receive a revamp, enhancing the user experience as well as providing more tools for data analysis such as iterative phase retrieval, and GPU support. The goal is to release version 1.0.0 by June next year, in time for summer conferences. The toolbox will also be open source, with contributions encouraged from the community.

8 | Discussion, Conclusions, and Future Work

The goal of this thesis was to develop the use of computational imaging techniques within electron microscopy to reduce acquisition time and potentially beam damage of samples, specifically focusing on the application of compressive sensing techniques to STEM simulation and 4-D STEM. As a result, a novel strategy for fast STEM simulation was developed, as well as the first practical demonstration of sub-sampled 4-D STEM. In addition, the use of elegant solutions to promote sparsity have improved results for experimental data, indicating the power of computational imaging techniques for electron microscopy.

8.1 Chapter summaries

In Chapter 3, compressive sensing and image inpainting were outlined, as well as describing what it means to sample at the Nyquist-rate for a STEM dataset acquisition. This chapter then took the ideas of efficient sampling into a compressed sensing framework, highlighting why it is possible to recover data from few sub-sampled measurements, as well as how it is performed in practice.

This chapter then extended into a new method for inpainting, R-LMI, which can be employed to inpaint simple, periodic datasets. In contrast, the BPFA was shown to inpaint aperiodic or defected datasets, indicating why BPFA is used throughout this work. Another key component of this chapter refers to the patch-size parameter and how a lower bound can be derived based on the type of sampling mask or sampling ratio which was used. This now allows for faster estimation of parameters.

However, there are a number of questions which are yet to be answered, such as

- Could a deep-learning (*i.e.*, generative adversarial network (GAN)) tool be used to inpaint the data? The short answer is yes, but there are inherent drawbacks to this versus the dictionary learning based approach. Neural networks can be very good at inpainting data through generative techniques, most commonly used to recover large regions of missing data in corrupted images [284]. The issue with applying this to electron microscopy is that it would be detrimental if the network generated an artefact, such as inpainting a defect or vacancy with the wrong solution. Furthermore, a network requires extensive training data (typically thousands of images), and this process takes a long time. Granted, other solutions involve creating smaller, less-generalised networks which can be updated according to new data, but the uncertainty of whether a data is *correct* would still linger. By contrast, the BPFA only inpaints what it can see and has no knowledge of the global structure. Chapter 3 demonstrated where this could fail, but also demonstrated how it can be corrected for by parameter tuning. Future work will look into the use of neural networks which could potentially perform remedial tasks, such as sparse coding steps.

- The tuning of BPFA hyperparameters requires extensive research. Although efforts were made to address this in a closed form, the number of parameters and their influences are beyond the scope of this thesis. However, it is a problem that does need solving if the use of the BPFA were to be rolled out to the community. One possible solution is to use a gradient descent which leverages the residual in the recovery as a cost function for parameter tuning. Although this solution may be realisable, it may not be realistic with respect to fast inpainting.

- A question remains regarding the optimal sampling strategy for STEM imaging. How many probe locations should be visited and in what order? There is no definitive answer so far. The best assumption to make is that for any given dataset which matches the criteria that it is compressible in some sparsity basis, it can be sub-sampled if the smallest feature will be sampled atleast once. That is, if the feature is the size of a pixel in the image, then sub-sampled will not work. Of course, in most applications the user is aware of the minimum feature size (atomic radii can be estimated, nanoparticles generally have a fixed size distribution), and as such can balance their sampling accordingly. For periodic

structures, this is not the case and the sampling can be greatly reduced such that the feature would then become the unit cell/s. As for the order of sampling, the limitations of hysteresis, as discussed in section 2.4.3, limit the possible scanning regimes. Work by others in the University of Liverpool group aim to address this issue by modelling the scan-coil dynamics [137].

Chapter 4 looked at the application of compressive sensing and computational imaging to 2-dimensional STEM data, specifically HAADF imaging. The main focus here was on how to control the STEM probe, how to collect data, and how to inpaint the data. This chapter demonstrated sub-sampling on three different systems at the University of Liverpool, Rosalind Franklin Institute, and CNR-IMM. This was important since it showed that compressive sensing can be applied to various systems without requiring a complete overhaul of the existing system.

This chapter then concluded with dictionary transfer from a simulated image to experimental data. The results showed that by using a simulated image to form the dictionary, a better reconstruction can be formed. This may be controversial and does probably ask more questions than it answers. Why does the resolution improve and what would happen if the same process was applied to an off-axis or astigmatic image? The answer is always that the algorithm's job is to find the solution which minimises the residual given the inputs. The solution given is the solution to that problem, and it may be that it is wrong in some cases. On the other hand, by running a self-learned dictionary at the same time, the two images can be compared. This then allows the user to verify whether the expected answer matches to their experiment. If it does then the microscope/sample are well aligned, if not then there are possibly other experimental factors to tune such as tilt or astigmatism.

The later may be useful for samples which become so damaged during acquisition that only short exposure of the beam is possible. Take a metal organic framework, say one that damages even under very low doses. If the signal acquired were low, then the signal could be improved by transferring the dictionary of the simulated structure. It is only equivalent to how simulations are used to characterise results, except now the matching is done through a dictionary matching.

A number of things could be expanded on,

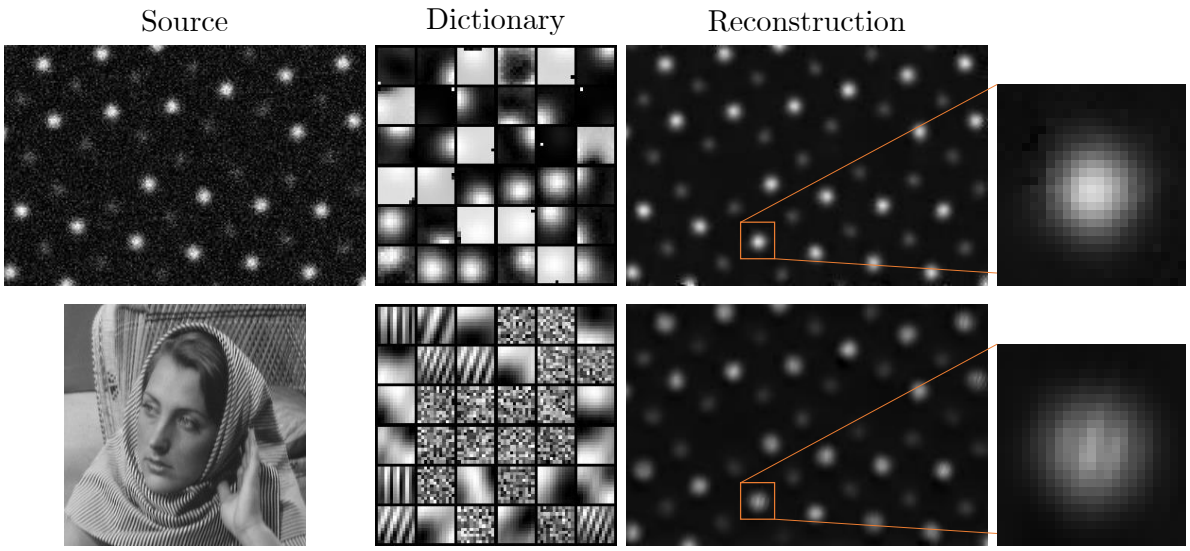

Figure 8.1: **Importance of using the correct dictionary.** The dictionary learned from the source to be inpainted (top row) provides a better recovery than the same image inpainted using a different source (bottom row).

- How can a dictionary be quantified as to whether it is suitable or not? Consider a dictionary which is not suited for a certain sample, say a dictionary containing stripes which is meant to inpaint an atomic resolution image, as in Fig. 8.1. The dictionary transferred from a poor source can produce artefacts (in this case stripes), although the overall structure remains intact. It would be useful to then have a way to define what makes that dictionary worse for the given input. The number of dictionary atoms used per overlapping patch could be a good measure, as this would imply the dictionary is not as useful for providing a sparse image representation. This would, however, mean reconstructing the data to find out.

- Does a new simulation have to be performed if defocus, magnification, or orientation change? This means more experiments need to be performed to validate the efficacy of dictionary transfer from a simulated image. Extending this, it is vital that the same results in Fig. 4.8 can be replicated but with changes to the defocus, tilt and astigmatism to see if the result presented is a true representation, despite the assumptions and observations during acquisition.

Chapter 5 gave evidence that compressive sensing methods can be applied in the simulation of STEM images, and showed that simulations can be sped up dramatically through a targeted sampling approach. By treating the simulation as a 3-D data cube, the redundancy through the frozen phonon configurations can be exploited and the correct contrast estimated

for complex structures such as grain boundaries and samples containing defects or vacancies. Furthermore, it was shown that the method can be applied to the multislice and PRISM algorithms, increasing speed in both cases.

At the end of this chapter, a possible use case other than transfer was postulated- automatic STEM alignment. Assume that the dictionary transferred from a STEM simulation shows higher resolution versus the self-learned reconstruction, as in that shown in Fig. 4.8. Now, assume that same set-up, but the microscope is optimally aligned at ideal focus conditions. It is postulated that the self-trained dictionary would produce the result closest to the transfer case when this condition is satisfied. To actually verify this, more testing is required as in the earlier discussion within this chapter.

Another option would be to develop a digital twin that is a computational symmetry of the microscope. Fast simulation could be part of this framework along with sub-sampling, inpainting, denoising, and possibly deep learning.

Chapter 6 focused on the application of compressive sensing to 4-D STEM. This chapter posed that sub-sampling of the probe locations, as well as down-sampling of the diffraction patterns can dramatically increase the rate of data acquisition. As part of this section, a tests were performed on experimental data. It was shown that this data could be reduced to <0.4% of its original size, whilst retaining functionally identical results.

Following this, the first experimentally acquired sub-sampled data was presented. It was shown that sub-sampling can be employed in 4-D STEM acquisition, however the results could be improved. Several areas could be improved upon, such as,

- Can the method be used in a live mode? This is something that has yet been tested, although there are plans to do so in the near future as part of a collaboration with the RFI. In theory, the answer is yes. The practical issues are yet to be explored but the application of live CS STEM has shown that it should be possible if all the correct signals and data can be transferred to the SenseAI software.
- Given the method, it would be a great test to image a material which has yet been visualised using 4-D STEM. One such material would be a MOF sample, and there are plans to try this out as part of an international collaboration.
- The ideas of detector down-sampling should be justified with a mathematical deriva-

tion. Currently, the observations are empirical, however it should be justified through theoretical calculations.

Chapter 7 presented other works, including the LoRePIE algorithm to improve quality of phase retrieval for low overlap ratios, and the MAT 4-D STEM toolbox. The LoRePIE algorithm has many areas for future research, such as the addition of other regularisation techniques such as soft thresholding, or a dictionary based regularisation. The method should also be tested for other iterative algorithms, which should be relatively simple to implement given the modularity of the technique. It would also be interesting to see if sub-sampling can be combined with LoRePIE to further reduce sampling of diffraction patterns, which could further reduce dose, perhaps below $1e^{-}Å^{2}$.

8.2 Future work and final remarks

In the application of computational imaging and compressive sensing to STEM, it has been shown that various methods can be improved through sub-sampling. By exploiting information across multiple dimensions, signal can be improved for scattering of weaker signals. Consider a STEM which can simultaneously acquire signals from various sources such as EDS, HAADF, 4-D STEM, and EELS. An example of such as system is shown in Fig. 8.2. Consider the following signals;

1. HAADF
2. 4-D STEM (ABF, CoM, ptychography)
3. EDX/EDS
4. EELS
5. Beam tilt (tomography)
6. *in-situ*
7. Simulations

and assume that there exists a method for acquiring them all simultaneously. It is postulated that through a computational imaging strategy, each of the signals would be combined to improve signal across the entire n-dimensional data. This would make for more efficient char-

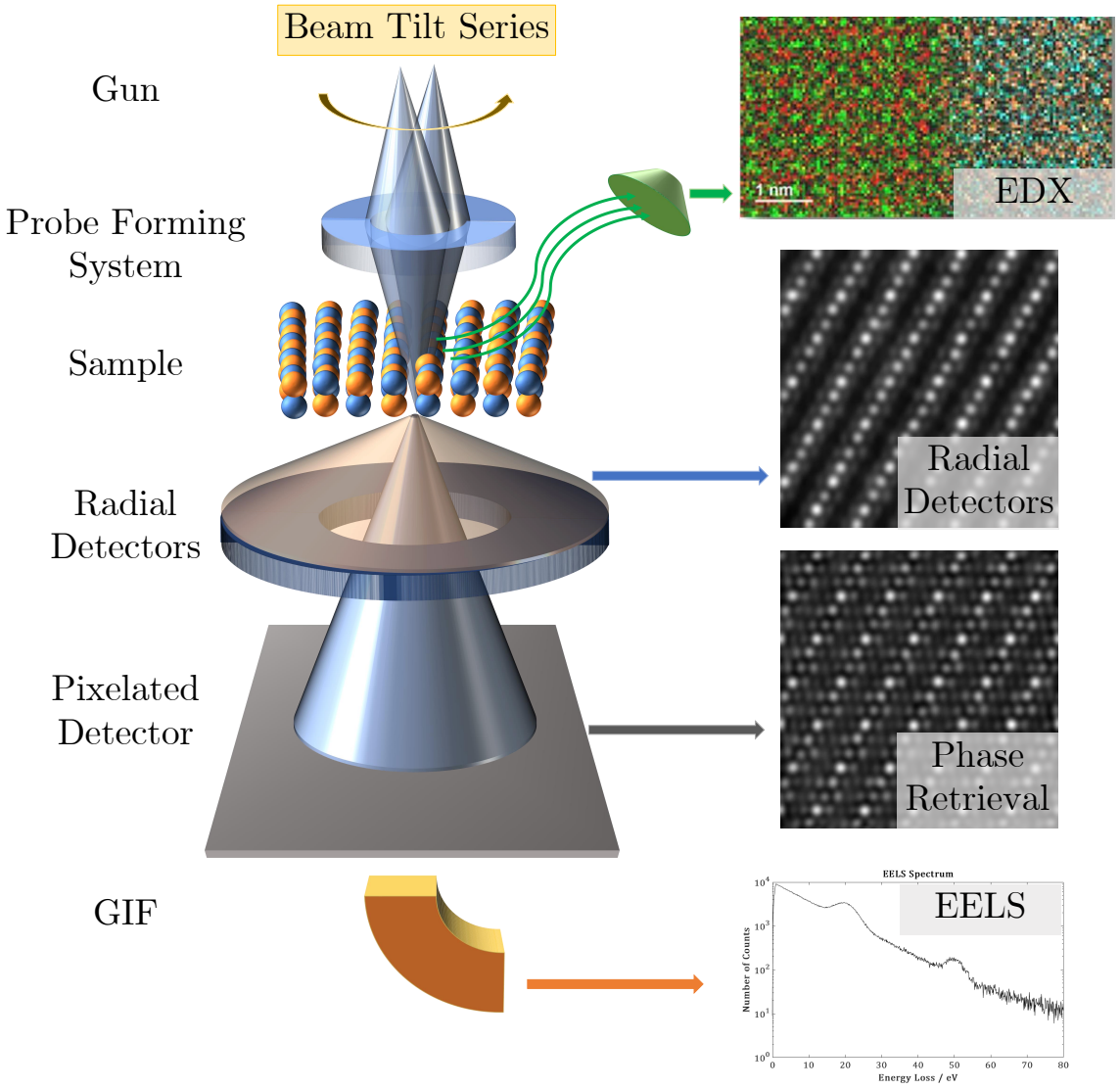

Figure 8.2: **The multi-dimensional STEM set-up.** By acquiring signals from multiple sources such as EDX, EELS, HAADF, and 4-D STEM, a multidimensional STEM data can be formed where signals from some sources can improve the signal-to-noise of others.

acterisation, where the entire possible information (using a STEM) could be extracted without requiring multiple scans over the same region. This would also reduce exposure, potentially leading to reduced beam damage and more reliable sample representation.

Perhaps this idea is no longer limited by scattering cross sections, nor data throughput, since compressive sensing and computational image can enhance the signal observed with fewer measurements.

In future work, the conclusions from each of the chapters aim to be addressed. The goal is to ultimately develop the self-driving or automated STEM, effectively by incorporating com-

putational imaging strategies and techniques into the acquisition process. There are still a lot of unanswered questions, ideas, and avenues to explore, and it is an exciting time to be at the cutting edge of method development in electron microscopy.

Bibliography

[1] W. Bollmann, "Interference effects in the electron microscopy of thin crystal foils," *Physical Review*, vol. 103, no. 5, p. 1588, 1956.

[2] E. James, N. Browning, A. Nicholls, M. Kawasaki, Y. Xin, and S. Stemmer, "Demonstration of atomic resolution Z-contrast imaging by a JEOL JEM-2010F scanning transmission electron microscope," *Microscopy*, vol. 47, no. 6, pp. 561–574, 1998.

[3] A. Crewe and J. Wall, "A scanning microscope with 5 Å resolution," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 375–393, 1970.

[4] H. Yang, H. Lee, M. Sarahan, Y. Sato, M. Chi, P. Moeck, Y. Ikuhara, and N. D. Browning, "Quantifying stoichiometry-induced variations in structure and energy of a $SrTiO_3$ symmetric $\Sigma 13$ $\{510\}/<100>$ grain boundary," *Philosophical Magazine*, vol. 93, no. 10-12, pp. 1219–1229, 2013.

[5] Y.-T. Lee and T. Ozaki, "OpenMX Viewer: A web-based crystalline and molecular graphical user interface program," *Journal of Molecular Graphics and Modelling*, vol. 89, pp. 192–198, 2019.

[6] O. Krivanek, N. Dellby, and A. Lupini, "Towards sub-Å electron beams," *Ultramicroscopy*, vol. 78, no. 1-4, pp. 1–11, 1999.

[7] E. Knapek and J. Dubochet, "Beam damage to organic material is considerably reduced in cryo-electron microscopy," *Journal of molecular biology*, vol. 141, no. 2, pp. 147–161, 1980.

[8] J. Wells, *Developing a fast and robust inpainting algorithm implementation for multidimensional electron microscopy data.* PhD thesis, University of Liverpool, pending submission.

[9] J. J. Thomson, *Cathode rays*. No. 4, Academic Reprints, 1897.

[10] J. J. Thomson, "XL. Cathode rays," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 44, no. 269, pp. 293–316, 1897.

[11] A. Einstein, "Über einem die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt," *Annalen der physik*, vol. 4, 1905.

[12] L. De Broglie, *Recherches sur la théorie des quanta*. PhD thesis, Migration-université en cours d'affectation, 1924.

[13] P. A. M. Dirac, "The quantum theory of the electron," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 117, no. 778, pp. 610–624, 1928.

[14] R. P. Feynman, *QED: The strange theory of light and matter*, vol. 90. Princeton University Press, 2006.

[15] R. P. Feynman, *The Principle of Least Action in Quantum Mechanics*. PhD thesis, Princeton University, 1942.

[16] R. P. Feynman, A. R. Hibbs, and D. F. Styer, *Quantum mechanics and path integrals*. Courier Corporation, 2010.

[17] E. Schrödinger, "An undulatory theory of the mechanics of atoms and molecules," *Physical review*, vol. 28, no. 6, p. 1049, 1926.

[18] B. Forbes, A. Martin, S. D. Findlay, A. J. D'Alfonso, and L. J. Allen, "Quantum mechanical model for phonon excitation in electron diffraction and imaging using a Born-Oppenheimer approximation," *Physical Review B*, vol. 82, no. 10, p. 104103, 2010.

[19] A. Barnes, "Rutherford Scattering," 2002. URL: https://www.personal.soton.ac.uk/ab1u06/teaching/phys3002/course/02_rutherford.pdf, Accessed: 05/08/23.

[20] X. Zou, S. Hovmöller, and P. Oleynikov, *Electron crystallography: electron microscopy and electron diffraction*, vol. 16. Oxford University Press, 2011.

[21] M. M. Disko, C. C. Ahn, and B. Fultz, *Transmission electron energy loss spectrometry in materials science*, vol. 2. Minerals, Metals, & Materials Society, 1992.

[22] N. Browning, M. Chisholm, and S. Pennycook, "Atomic-resolution chemical analysis using a scanning transmission electron microscope," *Nature*, vol. 366, no. 6451, pp. 143–146, 1993.

[23] P. Batson, "Simultaneous STEM imaging and electron energy-loss spectroscopy with atomic-column sensitivity," *Nature*, vol. 366, no. 6457, pp. 727–728, 1993.

[24] N. Browning, D. Wallis, P. Nellist, and S. Pennycook, "EELS in the STEM: Determination of materials properties on the atomic scale," *Micron*, vol. 28, no. 5, pp. 333–348, 1997.

[25] C. C. Ahn, *Transmission electron energy loss spectrometry in materials science and the EELS atlas*. John Wiley & Sons, 2006.

[26] E. Okunishi, H. Sawada, Y. Kondo, and M. Kersker, "Atomic resolution elemental map of EELS with a Cs corrected STEM," *Microscopy and Microanalysis*, vol. 12, no. S02, pp. 1150–1151, 2006.

[27] K. Kimoto, T. Asaka, T. Nagai, M. Saito, Y. Matsui, and K. Ishizuka, "Element-selective imaging of atomic columns in a crystal using STEM and EELS," *Nature*, vol. 450, no. 7170, pp. 702–704, 2007.

[28] M. Varela, J. Gazquez, and S. J. Pennycook, "Stem-eels imaging of complex oxides and interfaces," *Mrs bulletin*, vol. 37, no. 1, pp. 29–35, 2012.

[29] O. L. Krivanek, N. Dellby, M. F. Murfitt, M. F. Chisholm, T. J. Pennycook, K. Suenaga, and V. Nicolosi, "Gentle STEM: ADF imaging and EELS at low primary energies," *Ultramicroscopy*, vol. 110, no. 8, pp. 935–945, 2010.

[30] R. F. Egerton, *Electron energy-loss spectroscopy in the electron microscope*. Springer Science & Business Media, 2011.

[31] R. Senga and K. Suenaga, "Single-atom electron energy loss spectroscopy of light elements," *Nature communications*, vol. 6, no. 1, p. 7943, 2015.

[32] M. Kociak and O. Stéphan, "Mapping plasmons at the nanometer scale in an electron microscope," *Chemical Society Reviews*, vol. 43, no. 11, pp. 3865–3883, 2014.

[33] D. Shindo, T. Oikawa, D. Shindo, and T. Oikawa, "Energy dispersive x-ray spectroscopy," *Analytical electron microscopy for materials science*, pp. 81–102, 2002.

[34] A. d'Alfonso, B. Freitag, D. Klenov, and L. Allen, "Atomic-resolution chemical mapping using energy-dispersive x-ray spectroscopy," *Physical Review B*, vol. 81, no. 10, p. 100101, 2010.

[35] L. J. Allen, A. J. D'Alfonso, B. Freitag, and D. O. Klenov, "Chemical mapping at atomic resolution using energy-dispersive x-ray spectroscopy," *MRS bulletin*, vol. 37, no. 1, pp. 47–52, 2012.

[36] R. Birch and M. Marshall, "Computation of bremsstrahlung x-ray spectra and comparison with spectra measured with a Ge (Li) detector," *Physics in Medicine & Biology*, vol. 24, no. 3, p. 505, 1979.

[37] J. Cazaux, "From the physics of secondary electron emission to image contrasts in scanning electron microscopy," *Journal of electron microscopy*, vol. 61, no. 5, pp. 261–284, 2012.

[38] K. D. Vernon-Parry, "Scanning electron microscopy: an introduction," *III-Vs review*, vol. 13, no. 4, pp. 40–44, 2000.

[39] H. Seiler, "Secondary electron emission in the scanning electron microscope," *Journal of Applied Physics*, vol. 54, no. 11, pp. R1–R18, 1983.

[40] O. Guise, C. Strom, and N. Preschilla, "Stem-in-sem method for morphology analysis of polymer systems," *Polymer*, vol. 52, no. 5, pp. 1278–1285, 2011.

[41] Y. Yamazawa, S. Okada, Z. Yasenjiang, T. Sunaoshi, and K. Kaji, "The first results of the low voltage cold-FE SEM/STEM system equipped with EELS," *Microscopy and Microanalysis*, vol. 22, no. S3, pp. 50–51, 2016.

[42] M. Knoll and E. Ruska, "Das elektronenmikroskop," *Zeitschrift für physik*, vol. 78, pp. 318–339, 1932.

[43] E. Ruska, "The development of the electron microscope and of electron microscopy," *Reviews of modern physics*, vol. 59, no. 3, p. 627, 1987.

[44] L. De Broglie, "Waves and quanta," *Nature*, vol. 112, no. 2815, pp. 540–540, 1923.

[45] M. M. Freundlich, "Origin of the Electron Microscope: The history of a great invention, and of a misconception concerning the inventors, is reviewed.," *Science*, vol. 142, no. 3589, pp. 185–188, 1963.

[46] N. Bohr, "I. On the constitution of atoms and molecules," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 26, no. 151, pp. 1–25, 1913.

[47] N. Bohr, "XXXVII. On the constitution of atoms and molecules," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 26, no. 153, pp. 476–502, 1913.

[48] N. Bohr, "LXXIII. On the constitution of atoms and molecules," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 26, no. 155, pp. 857–875, 1913.

[49] T. Mulvey, "Origins and historical development of the electron microscope," *British journal of applied physics*, vol. 13, no. 5, p. 197, 1962.

[50] T. Palucka, "Overview of electron microscopy," *URL: https://authors. library. caltech. edu/5456/1/hrst. mit. edu/hrs/materials/public/ElectronMicroscope/EM_HistOverview. htm (visited on 04/17/2018)*, 2002.

[51] "Zeiss: Electron microscopy - how it all began." `https://www.zeiss.com/corporate/en/about-zeiss/past/history/technological-milestones/electron-microscopy.html`. Accessed: 2023-08-25.

[52] H. Inada, H. Kakibayashi, S. Isakozawa, T. Hashimoto, T. Yaguchi, and K. Nakamura, "Hitachi's development of cold-field emission scanning transmission electron microscopes," *Advances in Imaging and Electron Physics*, vol. 159, pp. 123–186, 2009.

[53] "Milestones: The Company: JEOL Ltd.." `https://www.jeol.com/corporate/outline/history.php`. Accessed: 2023-08-25.

[54] R. Heidenreich, "Electron microscope and diffraction study of metal crystal textures by means of thin sections," *Journal of Applied Physics*, vol. 20, no. 10, pp. 993–1010, 1949.

[55] R. Cahn, "Plastic deformation of alpha-uranium; twinning and slip," *Acta metallurgica*, vol. 1, no. 1, pp. 49–70, 1953.

[56] M. J. Whelan, P. B. Hirsch, R. Horne, and W. Bollmann, "Dislocations and stacking faults in stainless steel," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 240, no. 1223, pp. 524–538, 1957.

[57] K. Yada and T. Hibi, "Fine Lattice Fringes around 1Å Resolved by the Axial Illumination," *Journal of Electron Microscopy*, vol. 18, no. 4, pp. 266–271, 1969.

[58] D. B. Williams and C. B. Carter, *The transmission electron microscope*, p. 156. Springer, 1996.

[59] O. Scherzer, "Über einige fehler von elektronenlinsen," *Zeitschrift für Physik*, vol. 101, no. 9-10, pp. 593–603, 1936.

[60] O. Scherzer, "Spharische und chromatische korrektur von elektronen-linsen," *Optik*, vol. 2, pp. 114–132, 1947.

[61] R. F. Klie, "Reaching a new resolution standard with electron microscopy," *Physics*, vol. 2, p. 85, 2009.

[62] J. Zach and M. Haider, "Aberration correction in a low voltage sem by a multipole corrector," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 363, no. 1-2, pp. 316–325, 1995.

[63] M. Haider, H. Rose, S. Uhlemann, E. Schwan, B. Kabius, and K. Urban, "A spherical-aberration-corrected 200 kv transmission electron microscope," *Ultramicroscopy*, vol. 75, no. 1, pp. 53–60, 1998.

[64] F. Quigley, P. McBean, P. O'Donovan, J. J. Peters, and L. Jones, "Cost and Capability Compromises in STEM Instrumentation for Low-Voltage Imaging," *Microscopy and Microanalysis*, vol. 28, no. 4, pp. 1437–1443, 2022.

[65] D. C. Bell, C. J. Russo, and D. V. Kolmykov, "40keV atomic resolution TEM," *Ultramicroscopy*, vol. 114, pp. 31–37, 2012.

[66] S. Takeda and H. Yoshida, "Atomic-resolution environmental TEM for quantitative in-situ microscopy in materials science," *Microscopy*, vol. 62, no. 1, pp. 193–203, 2013.

[67] T. Uchiyama, H. Yoshida, Y. Kuwauchi, S. Ichikawa, S. Shimada, M. Haruta, and S. Takeda, "Systematic morphology changes of gold nanoparticles supported on CeO_2 during CO oxidation," *Angewandte Chemie International Edition*, vol. 50, no. 43, pp. 10157–10160, 2011.

[68] Y. Kuwauchi, H. Yoshida, T. Akita, M. Haruta, and S. Takeda, "Intrinsic catalytic structure of gold nanoparticles supported on TiO_2," *Angewandte Chemie (International ed. in English)*, vol. 51, no. 31, pp. 7729–7733, 2012.

[69] P. L. Hansen, J. B. Wagner, S. Helveg, J. R. Rostrup-Nielsen, B. S. Clausen, and H. Topsøe, "Atom-resolved imaging of dynamic shape changes in supported copper nanocrystals," *Science*, vol. 295, no. 5562, pp. 2053–2055, 2002.

[70] X. H. Liu and J. Y. Huang, "In situ TEM electrochemistry of anode materials in lithium ion batteries," *Energy & Environmental Science*, vol. 4, no. 10, pp. 3844–3860, 2011.

[71] F. Wu and N. Yao, "Advances in sealed liquid cells for in-situ TEM electrochemial investigation of lithium-ion battery," *Nano Energy*, vol. 11, pp. 196–210, 2015.

[72] Y. Yuan, K. Amine, J. Lu, and R. Shahbazian-Yassar, "Understanding materials challenges for rechargeable ion batteries with in situ transmission electron microscopy," *Nature communications*, vol. 8, no. 1, p. 15806, 2017.

[73] W. Li, N. D. Browning, and B. Layla Mehdi, "Electron microscopies for batteries," in *Batteries: Materials principles and characterization methods*, pp. 6–1, IOP Publishing Bristol, UK, 2021.

[74] Q. Yu, M. Legros, and A. Minor, "In situ tem nanomechanics," *Mrs Bulletin*, vol. 40, no. 1, pp. 62–70, 2015.

[75] D. B. Williams and C. B. Carter, *The transmission electron microscope*, p. 491. Springer, 1996.

[76] D. B. Williams and C. B. Carter, *The transmission electron microscope*, p. 485. Springer, 1996.

[77] D. B. Williams and C. B. Carter, *The transmission electron microscope*, p. 495. Springer, 1996.

[78] O. Scherzer, "The theoretical resolution limit of the electron microscope," *Journal of Applied Physics*, vol. 20, no. 1, pp. 20–29, 1949.

[79] D. B. Williams and C. B. Carter, *The transmission electron microscope*, pp. 486–487. Springer, 1996.

[80] D. B. Williams and C. B. Carter, *The transmission electron microscope*, p. 492. Springer, 1996.

[81] C. D. Meinhart and S. T. Wereley, "The theory of diffraction-limited resolution in microparticle image velocimetry," *Measurement science and technology*, vol. 14, no. 7, p. 1047, 2003.

[82] D. B. Williams and C. B. Carter, *The transmission electron microscope*, p. 5. Springer, 1996.

[83] Z. Chen, Y. Jiang, Y.-T. Shao, M. E. Holtz, M. Odstrčil, M. Guizar-Sicairos, I. Hanke, S. Ganschow, D. G. Schlom, and D. A. Muller, "Electron ptychography achieves atomic-resolution limits set by lattice vibrations," *Science*, vol. 372, no. 6544, pp. 826–831, 2021.

[84] S. J. Pennycook and P. D. Nellist, *Scanning transmission electron microscopy: imaging and analysis*. Springer Science & Business Media, 2011.

[85] M. Von Ardenne, "Das Elektronen-Rastermikroskop: Theoretische Grundlagen," *Zeitschrift für Physik*, vol. 109, no. 9-10, pp. 553–572, 1938.

[86] M. Von Ardenne, "Das Elektronen-Rastermikroskop: Praktische Ausführung," *Zeitschrift für technische Physik*, vol. 19, pp. 407–416, 1938.

[87] M. von Ardenne, "On the history of scanning electron microscopy, of the electron microprobe, and of early contributions to transmission electron microscopy," in *The beginnings of electron microscopy*, vol. 16, pp. 1–21, Academic Press Orlando, 1985.

[88] M. von Ardenne, "über eine Atomumwandlungsanlage für Spannungen bis zu 1 Million Volt," *Zeitschrift für Physik*, vol. 121, no. 3-4, pp. 236–267, 1943.

[89] A. V. Crewe, "Scanning Electron Microscopes: Is High Resolution Possible? Use of a field-emission electron source may make it possible to overcome existing limitations on resolution.," *Science*, vol. 154, no. 3750, pp. 729–738, 1966.

[90] A. Crewe, D. Eggenberger, J. Wall, and L. Welter, "Electron gun using a field emission source," *Review of Scientific Instruments*, vol. 39, no. 4, pp. 576–583, 1968.

[91] A. V. Crewe, M. Isaacson, and D. Johnson, "A simple scanning electron microscope," *Review of Scientific Instruments*, vol. 40, no. 2, pp. 241–246, 1969.

[92] A. V. Crewe, J. Wall, and J. Langmore, "Visibility of single atoms," *science*, vol. 168, no. 3937, pp. 1338–1340, 1970.

[93] A. V. Crewe, "A high-resolution scanning electron microscope," *Scientific American*, vol. 224, no. 4, pp. 26–35, 1971.

[94] A. Crewe, M. Isaacson, and D. Johnson, "Electron energy loss spectra of the nucleic acid bases," *Nature*, vol. 231, no. 5300, pp. 262–263, 1971.

[95] M. Isaacson, D. Johnson, and A. Crewe, "Electron beam excitation and damage of biological molecules; its implications for specimen damage in electron microscopy," *Radiation research*, pp. 205–224, 1973.

[96] J. Wall, J. Langmore, M. Isaacson, and A. Crewe, "Scanning transmission electron microscopy at high resolution," *Proceedings of the National Academy of Sciences*, vol. 71, no. 1, pp. 1–5, 1974.

[97] I. R. Wardell and P. E. Bovey, "Chapter 6: A History of Vacuum Generators' 100-kV Scanning Transmission Electron Microscope," in *Advances in Imaging and Electron Physics*, vol. 159 of *Advances in Imaging and Electron Physics*, pp. 221–285, Elsevier, 2009.

[98] O. Krivanek, N. Dellby, A. Spence, R. Camps, and L. Brown, "Aberration correction in the STEM," in *Proceedings of the Institute of Physics Electron Microscopy and Analysis Group Conference*, vol. 153, pp. 35–39, 1997.

[99] M. Haider, S. Uhlemann, E. Schwan, H. Rose, B. Kabius, and K. Urban, "Electron microscopy image enhanced," *Nature*, vol. 392, no. 6678, pp. 768–769, 1998.

[100] S. Pennycook and J. Narayan, "Direct imaging of dopant distributions in silicon by scanning transmission electron microscopy," *Applied physics letters*, vol. 45, no. 4, pp. 385–387, 1984.

[101] S. J. Pennycook, "Seeing the atoms more clearly: STEM imaging from the Crewe era to today," *Ultramicroscopy*, vol. 123, pp. 28–37, 2012.

[102] S. Pennycook and D. Jesson, "High-resolution incoherent imaging of crystals," *Physical review letters*, vol. 64, no. 8, p. 938, 1990.

[103] S. Pennycook and D. Jesson, "High-resolution Z-contrast imaging of crystals," *Ultramicroscopy*, vol. 37, no. 1-4, pp. 14–38, 1991.

[104] D. Jesson and S. Pennycook, "Incoherent imaging of thin specimens using coherently scattered electrons," *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences*, vol. 441, no. 1912, pp. 261–281, 1993.

[105] D. Jesson and S. J. Pennycook, "Incoherent imaging of crystals using thermally scattered electrons," *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences*, vol. 449, no. 1936, pp. 273–293, 1995.

[106] R. F. Loane, P. Xu, and J. Silcox, "Incoherent imaging of zone axis crystals with ADF STEM," *Ultramicroscopy*, vol. 40, no. 2, pp. 121–138, 1992.

[107] D. Jesson, S. Pennycook, J.-M. Baribeau, and D. Houghton, "Direct imaging of surface cusp evolution during strained-layer epitaxy and implications for strain relaxation," *Physical review letters*, vol. 71, no. 11, p. 1744, 1993.

[108] D. Jesson, S. Pennycook, J. Tischler, J. Budai, J.-M. Baribeau, and D. Houghton, "Interplay between evolving surface morphology, atomic-scale growth modes, and ordering during Si_x Ge_{1-x} epitaxy," *Physical review letters*, vol. 70, no. 15, p. 2293, 1993.

[109] M. Chisholm and S. Pennycook, "Structural origin of reduced critical currents at YBa2Cu3O7–δ grain boundaries," *Nature*, vol. 351, no. 6321, pp. 47–49, 1991.

[110] S. Pennycook, N. Browning, M. McGibbon, A. McGibbon, D. Jesson, and M. Chisholm, "Direct determination of interface structure and bonding with the scanning transmission electron microscope," *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 354, no. 1719, pp. 2619–2634, 1996.

[111] H. Vonharrach, A. Nicholls, D. Jesson, and S. Pennycook, "First results of a 300 kV high-resolution field-emission STEM," *Electron Microscopy and Analysis 1993*, no. 138, pp. 499–502, 1993.

[112] Y. Xin, S. Pennycook, N. Browning, P. Nellist, S. Sivananthan, F. Omnes, B. Beaumont, J. Faurie, and P. Gibart, "Direct observation of the core structures of threading dislocations in GaN," *Applied physics letters*, vol. 72, no. 21, pp. 2680–2682, 1998.

[113] V. Ronchi, "Forty years of history of a grating interferometer," *Applied optics*, vol. 3, no. 4, pp. 437–451, 1964.

[114] E. James and N. Browning, "Practical aspects of atomic resolution imaging and analysis in STEM," *Ultramicroscopy*, vol. 78, no. 1-4, pp. 125–139, 1999.

[115] S. J. Pennycook and P. D. Nellist, *Scanning transmission electron microscopy: imaging and analysis*, p. 117. Springer Science & Business Media, 2011.

[116] S. J. Pennycook and P. D. Nellist, *Scanning transmission electron microscopy: imaging and analysis*, pp. 102–103. Springer Science & Business Media, 2011.

[117] S. J. Pennycook and P. D. Nellist, *Scanning transmission electron microscopy: imaging and analysis*, pp. 92–93. Springer Science & Business Media, 2011.

[118] R. Egerton, P. Li, and M. Malac, "Radiation damage in the TEM and SEM," *Micron*, vol. 35, no. 6, pp. 399–409, 2004.

[119] R. Egerton, "Mechanisms of radiation damage in beam-sensitive specimens, for TEM accelerating voltages between 10 and 300 kV," *Microscopy research and technique*, vol. 75, no. 11, pp. 1550–1556, 2012.

[120] R. Egerton, "Radiation damage to organic and inorganic specimens in the TEM," *Micron*, vol. 119, pp. 72–87, 2019.

[121] H.-P. Komsa, J. Kotakoski, S. Kurasch, O. Lehtinen, U. Kaiser, and A. V. Krasheninnikov, "Two-dimensional transition metal dichalcogenides under electron irradiation: defect production and doping," *Physical review letters*, vol. 109, no. 3, p. 035503, 2012.

[122] S. de Graaf and B. J. Kooi, "Radiation damage and defect dynamics in 2D WS2: a low-voltage scanning transmission electron microscopy study," *2D Materials*, vol. 9, no. 1, p. 015009, 2021.

[123] R. Egerton, "Control of radiation damage in the TEM," *Ultramicroscopy*, vol. 127, pp. 100–108, 2013. Frontiers of Electron Microscopy in Materials Science.

[124] D. C. Joy and C. S. Joy, "Dynamic charging in the low voltage SEM," *Microscopy and Microanalysis*, vol. 1, no. 3, pp. 109–112, 1995.

[125] R. Fleck, C. Bisson, C. Hecksel, and J. B. Gilchrist, "Preparing lamellae from vitreous biological samples using a dual-beam scanning electron microscope for cryo-electron tomography," *Journal of Visualized Experiments*, 2021.

[126] T. Shaffner and R. Van Veld, "Charging effects in the scanning electron microscope," *Journal of Physics E: Scientific Instruments*, vol. 4, no. 9, p. 633, 1971.

[127] K. H. Kim, Z. Akase, T. Suzuki, and D. Shindo, "Charging effects on SEM/SIM contrast of metal/insulator system in various metallic coating conditions," *Materials transactions*, vol. 51, no. 6, pp. 1080–1083, 2010.

[128] H. Heide, "Die Objektverschmutzung im Elektronenmikroskop und das Problem der Strahlenschädigung durch Kohlenstoffabbau," *Z. angew. Phys.*, vol. 15, pp. 116–128, 1963.

[129] M. Hugenschmidt, K. Adrion, A. Marx, E. Müller, and D. Gerthsen, "Electron-Beam-Induced Carbon Contamination in STEM-in-SEM: Quantification and Mitigation," *Microscopy and Microanalysis*, vol. 29, no. 1, pp. 219–234, 2023.

[130] J. Hillier, "On the investigation of specimen contamination in the electron microscope," *Journal of Applied Physics*, vol. 19, no. 3, pp. 226–230, 1948.

[131] R. Egerton and C. Rossouw, "Direct measurement of contamination and etching rates in an electron beam," *Journal of Physics D: Applied Physics*, vol. 9, no. 4, p. 659, 1976.

[132] S. Hettler, M. Dries, P. Hermann, M. Obermair, D. Gerthsen, and M. Malac, "Carbon contamination in scanning transmission electron microscopy and its impact on phase-plate applications," *Micron*, vol. 96, pp. 38–47, 2017.

[133] P. Rödiger, H. D. Wanzenboeck, G. Hochleitner, and E. Bertagnolli, "Evaluation of chamber contamination in a scanning electron microscope," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena*, vol. 27, no. 6, pp. 2711–2717, 2009.

[134] D. R. Mitchell, "Contamination mitigation strategies for scanning transmission electron microscopy," *Micron*, vol. 73, pp. 36–46, 2015.

[135] Y. Murooka, *Parallel electron energy loss spectroscopy of electron hole drilling in calcite.* PhD thesis, University of Cambridge, 1994.

[136] D. B. Williams and C. B. Carter, *The transmission electron microscope*, p. 158. Springer, 1996.

[137] D. Nicholls, J. Wells, A. W. Robinson, A. Moshtaghpour, A. I. Kirkland, and N. D. Browning, "Scan Coil Dynamics Simulation for Subsampled Scanning Transmission Electron Microscopy," *arXiv preprint arXiv:2307.08441*, 2023.

[138] S. Ning, T. Fujita, A. Nie, Z. Wang, X. Xu, J. Chen, M. Chen, S. Yao, and T.-Y. Zhang, "Scanning distortion correction in STEM images," *Ultramicroscopy*, vol. 184, pp. 274–283, 2018.

[139] T. Mullarkey, J. J. Peters, C. Downing, and L. Jones, "Using your beam efficiently: Reducing electron dose in the STEM via flyback compensation," *Microscopy and Microanalysis*, vol. 28, no. 4, pp. 1428–1436, 2022.

[140] L. Jones and P. D. Nellist, "Identifying and correcting scan noise and drift in the scanning transmission electron microscope," *Microscopy and Microanalysis*, vol. 19, no. 4, pp. 1050–1060, 2013.

[141] L. Kovarik, A. Stevens, A. Liyu, and N. D. Browning, "Implementing an accurate and rapid sparse sampling approach for low-dose atomic resolution STEM imaging," *Applied Physics Letters*, vol. 109, no. 16, 2016.

[142] E. T. Whittaker, "XVIII.—On the functions which are represented by the expansions of the interpolation-theory," *Proceedings of the Royal Society of Edinburgh*, vol. 35, pp. 181–194, 1915.

[143] H. Nyquist, "Certain topics in telegraph transmission theory," *Transactions of the American Institute of Electrical Engineers*, vol. 47, no. 2, pp. 617–644, 1928.

[144] C. Shannon, "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, pp. 10–21, jan 1949.

[145] D. L. Donoho and E. J. Candès, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[146] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE transactions on information theory*, vol. 52, no. 2, pp. 489–509, 2006.

[147] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 59, no. 8, pp. 1207–1223, 2006.

[148] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.

[149] Y. Wiaux, L. Jacques, G. Puy, A. M. Scaife, and P. Vandergheynst, "Compressed sensing imaging techniques for radio interferometry," *Monthly Notices of the Royal Astronomical Society*, vol. 395, no. 3, pp. 1733–1742, 2009.

[150] L.-l. Xiao, K. Liu, D.-p. Han, and J.-y. Liu, "A compressed sensing approach for enhancing infrared imaging resolution," *Optics & Laser Technology*, vol. 44, no. 8, pp. 2354–2360, 2012.

[151] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proceedings of 27th Asilomar conference on signals, systems and computers*, pp. 40–44, IEEE, 1993.

[152] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on information theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[153] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[154] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5505–5514, 2018.

[155] G. Pilikos and N. Philip, "Beta process factor analysis for efficient seismic compressive sensing with uncertainty quantification," in *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*, pp. 1–5, IEEE, 2018.

[156] J. Paisley and L. Carin, "Nonparametric factor analysis with beta process priors," in *Proceedings of the 26th annual international conference on machine learning*, pp. 777–784, 2009.

[157] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of machine learning research*, vol. 1, no. Jun, pp. 211–244, 2001.

[158] M. Zhou, H. Chen, J. W. Paisley, L. Ren, G. Sapiro, and L. Carin, "Non-parametric Bayesian dictionary learning for sparse image representations.," in *NIPS*, vol. 9, pp. 2295–2303, 2009.

[159] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[160] S. Sertoglu and J. Paisley, "Scalable Bayesian nonparametric dictionary learning," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, pp. 2771–2775, 2015.

[161] J. W. Paisley, D. M. Blei, and M. I. Jordan, "Bayesian Nonnegative Matrix Factorization with Stochastic Variational Inference," 2014.

[162] X. Ding, J. Paisley, Y. Huang, X. Chen, F. Huang, and X.-p. Zhang, "Compressed sensing MRI with Bayesian dictionary learning," in *2013 IEEE International Conference on Image Processing*, pp. 2319–2323, IEEE, 2013.

[163] Y. Huang, J. Paisley, Q. Lin, X. Ding, X. Fu, and X.-P. Zhang, "Bayesian nonparametric dictionary learning for compressed sensing MRI," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5007–5019, 2014.

[164] D. Nicholls, J. Wells, A. Stevens, Y. Zheng, J. Castagna, and N. D. Browning, "Sub-sampled imaging for STEM: Maximising image speed, resolution and precision through reconstruction parameter refinement," *Ultramicroscopy*, vol. 233, p. 113451, 2022.

[165] Y. Wang, Z. Liang, H. Zheng, and R. Cao, "Recent progress on defect-rich transition metal oxides and their energy-related applications," *Chemistry–An Asian Journal*, vol. 15, no. 22, pp. 3717–3736, 2020.

[166] S. Wallis, "Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods," *Journal of Quantitative Linguistics*, vol. 20, no. 3, pp. 178–208, 2013.

[167] M. J. Beal, *Variational algorithms for approximate Bayesian inference*. University of London, University College London (United Kingdom), 2003.

[168] A. Pryor, C. Ophus, and J. Miao, "A streaming multi-GPU implementation of image simulation algorithms for scanning transmission electron microscopy," *Advanced structural and chemical imaging*, vol. 3, no. 1, p. 8, 2017.

[169] J. Madsen and T. Susi, "The abTEM code: transmission electron microscopy from first principles," *Open Research Europe*, vol. 1, p. 24, 2021.

[170] J. Barthel, "Dr. Probe: A software for high-resolution STEM image simulation," *Ultramicroscopy*, vol. 193, pp. 1–11, 2018.

[171] I. Lobato, S. van Aert, and J. Verbeeck, "MULTEM: A new multislice program to perform accurate and fast electron diffraction and imaging simulations using Graphics Processing Units with CUDA," *Ultramicroscopy*, vol. 156, pp. 9–17, 2015.

[172] A. V. Crewe, "An introduction to the STEM," *Journal of ultrastructure research*, vol. 88, no. 2, pp. 94–104, 1984.

[173] A. V. Crewe, "Scanning transmission electron microscopy," *Journal of microscopy*, vol. 100, no. 3, pp. 247–259, 1974.

[174] V. Beck and A. V. Crewe, "High resolution imaging properties of the STEM," *Ultramicroscopy*, vol. 1, no. 2, pp. 137–144, 1975.

[175] J. M. Cowley and A. F. Moodie, "The scattering of electrons by atoms and crystals. I. A new theoretical approach," *Acta Crystallographica*, vol. 10, no. 10, pp. 609–619, 1957.

[176] P. Goodman and A. F. Moodie, "Numerical evaluations of N-beam wave functions in electron scattering by the multi-slice method," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 30, no. 2, pp. 280–290, 1974.

[177] K. Ishizuka, "A practical approach for STEM image simulation based on the FFT multislice method," *Ultramicroscopy*, vol. 90, no. 2–3, pp. 141–150, 2002.

[178] K. Ishizuka and N. Uyeda, "A new theoretical and practical approach to the multislice method," *Acta Crystallographica Section A*, vol. 33, no. 5, pp. 815–824, 1977.

[179] E. J. Kirkland, *Advanced Computing in Electron Microscopy*. Springer, 2010.

[180] E. J. Kirkland, R. F. Loane, and J. Silcox, "Simulation of annular dark field STEM images using a modified multislice method," *Ultramicroscopy*, vol. 23, no. 1, pp. 77–96, 1987.

[181] P. A. Stadelmann, "EMS–a software package for electron diffraction analysis and HREM image simulation in materials science," *Ultramicroscopy*, vol. 21, no. 2, pp. 131–146, 1987.

[182] P. Stadelmann, "Image analysis and simulation software in transmission electron microscopy," *Microscopy and Microanalysis*, vol. 9, no. S03, pp. 934–935, 2003.

[183] R. Kilaas, "MacTempas a program for simulating high resolution TEM images and diffraction patterns." `http://www.totalresolution.com/`, 2019.

[184] C. T. Koch, *Determination of core structure periodicity and point defect density along dislocations*. PhD thesis, Arizona, 2002.

[185] M. de Graef, *Introduction to Conventional Transmission Electron Microscopy*. Cambridge: Cambridge University Press, 2003.

[186] J. M. Zuo and J. C. Mabon, "Web-based electron microscopy application software: Web-EMAPS," *Microscopy and Microanalysis*, vol. 10, no. S02, pp. 214–215, 2004.

[187] E. Carlino, V. Grillo, and P. Palazzari, "Accurate and fast multislice simulations of HAADF image contrast by parallel computing," in *Microscopy of Semiconducting Materials 2007*, pp. 177–180, Springer, 2008.

[188] A. Rosenauer and M. Schowalter, "STEMSIM—a New Software Tool for Simulation of STEM HAADF Z-Contrast Imaging," in *Microscopy of Semiconducting Materials 2007*, pp. 305–308, Springer Netherlands, 2008.

[189] S. K. Walton, K. Zeissler, W. R. Branford, and S. Felton, "MALTS: A tool to simulate Lorentz transmission electron microscopy from micromagnetic simulations," *IEEE Transactions on Magnetics*, vol. 49, no. 8, pp. 4931–4934, 2013.

[190] M. Bar-Sadan, J. Barthel, H. Shtrikman, and L. Houben, "Direct imaging of single Au atoms within GaAs nanowires," *Nano Letters*, vol. 12, no. 5, pp. 2443–2447, 2012.

[191] I. Lobato, S. van Aert, and J. Verbeeck, "Progress and new advances in simulating electron microscopy datasets using MULTEM," *Ultramicroscopy*, vol. 168, pp. 17–27, 2016.

[192] W. van den Broek, X. Jiang, and C. T. Koch, "FDES, a GPU-based multislice algorithm with increased efficiency of the computation of the projected potential," *Ultramicroscopy*, vol. 158, pp. 24–33, 2015.

[193] E. C. Cosgriff, A. J. D'Alfonso, L. J. Allen, S. D. Findlay, A. I. Kirkland, and P. D. Nellist, "Three-dimensional imaging in double aberration-corrected scanning confocal electron microscopy, Part I," *Ultramicroscopy*, vol. 108, no. 12, pp. 1678–1687, 2008.

[194] J. O. Oelerich, L. Duschek, J. Belz, A. Beyer, S. D. Baranovskii, and K. Volz, "STEMsalabim: A high-performance computing cluster friendly code for scanning transmission electron microscopy image simulations of thin specimens," *Ultramicroscopy*, vol. 177, pp. 84–92, 2017.

[195] C. Dwyer, "Simulation of scanning transmission electron microscope images on desktop computers," *Ultramicroscopy*, vol. 110, no. 6, pp. 656–665, 2010.

[196] M. Born and K. Sarginson, "The effect of thermal vibrations on the scattering of X-rays," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 179, no. 976, pp. 69–93, 1941.

[197] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[198] A. Horé and D. Ziou, "Is there a relationship between peak-signal-to-noise ratio and structural similarity index measure?," *IET Image Processing*, vol. 7, no. 1, pp. 12–24, 2013.

[199] V. Sears and S. Shelley, "Debye–Waller factor for elemental crystals," *Acta Crystallographica Section A: Foundations of Crystallography*, vol. 47, no. 4, pp. 441–446, 1991.

[200] R. Mas-Balleste, C. Gomez-Navarro, J. Gomez-Herrero, and F. Zamora, "2D materials: to graphene and beyond," *Nanoscale*, vol. 3, no. 1, pp. 20–30, 2011.

[201] K. Novoselov, A. Mishchenko, A. Carvalho, and A. Castro Neto, "2D materials and van der Waals heterostructures," *Science*, vol. 353, no. 6298, p. aac9439, 2016.

[202] W. Zhou, X. Zou, S. Najmaei, Z. Liu, Y. Shi, J. Kong, J. Lou, P. M. Ajayan, B. I. Yakobson, and J.-C. Idrobo, "Intrinsic structural defects in monolayer molybdenum disulfide," *Nano letters*, vol. 13, no. 6, pp. 2615–2622, 2013.

[203] K. F. Mak, C. Lee, J. Hone, J. Shan, and T. F. Heinz, "Atomically thin MoS_2: a new direct-gap semiconductor," *Physical review letters*, vol. 105, no. 13, p. 136805, 2010.

[204] B. Radisavljevic, A. Radenovic, J. Brivio, V. Giacometti, and A. Kis, "Single-layer MoS_2 transistors," *Nature nanotechnology*, vol. 6, no. 3, pp. 147–150, 2011.

[205] Z. Yin, H. Li, H. Li, L. Jiang, Y. Shi, Y. Sun, G. Lu, Q. Zhang, X. Chen, and H. Zhang, "Single-layer MoS_2 phototransistors," *ACS nano*, vol. 6, no. 1, pp. 74–80, 2012.

[206] S. J. Pennycook, "Z-contrast STEM for materials science," *Ultramicroscopy*, vol. 30, no. 1-2, pp. 58–69, 1989.

[207] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[208] Q. Zheng, T. Feng, J. A. Hachtel, R. Ishikawa, Y. Cheng, L. Daemen, J. Xing, J. C. Idrobo, J. Yan, N. Shibata, Y. Ikuhara, B. C. Sales, S. T. Pantelides, and M. Chi, "Direct visualization of anionic electrons in an electride reveals inhomogeneities," *Science Advances*, vol. 7, no. 15, p. eabe6819, 2021.

[209] T. Isogai, H. Tanaka, T. Goto, A. Teramoto, S. Sugawa, and T. Ohmi, "Formation and property of yttrium and yttrium silicide films as low schottcky barrier material for n-type silicon," *Japanese journal of applied physics*, vol. 47, no. 4S, p. 3138, 2008.

[210] N. Kuganathan, A. Chroneos, and R. W. Grimes, "One-dimensional yttrium silicide electride (y5si3: e-) for encapsulation of volatile fission products," *Journal of Applied Physics*, vol. 129, no. 24, p. 245105, 2021.

[211] D. Nicholls, A. Robinson, J. Wells, A. Moshtaghpour, M. Bahri, A. Kirkland, and N. Browning, "Compressive scanning transmission electron microscopy," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1586–1590, 2022.

[212] D. Nicholls, J. Wells, A. W. Robinson, A. Moshtaghpour, M. Kobylynska, R. A. Fleck, A. I. Kirkland, and N. D. Browning, "A Targeted Sampling Strategy for Compressive Cryo Focused Ion Beam Scanning Electron Microscopy," *arXiv preprint arXiv:2211.03494*, 2022.

[213] R. Egerton, "Radiation Damage and Nanofabrication in TEM and STEM," *Microscopy Today*, vol. 29, no. 3, p. 56–59, 2021.

[214] N. D. Browning, J. Castagna, A. I. Kirkland, A. Moshtaghpour, D. Nicholls, A. W. Robinson, J. Wells, and Y. Zheng, "The advantages of sub-sampling and inpainting for scanning transmission electron microscopy," *Applied Physics Letters*, vol. 122, no. 5, 2023.

[215] L. Carin, D. Liu, and B. Guo, "Coherence, compressive sensing, and random sensor arrays," *IEEE Antennas and Propagation Magazine*, vol. 53, no. 4, pp. 28–39, 2011.

[216] P. Nellist, B. McCallum, and J. M. Rodenburg, "Resolution beyond the information limit in transmission electron microscopy," *Nature*, vol. 374, no. 6523, pp. 630–632, 1995.

[217] N. J. Zaluzec, "Lorentz STEM: A digital approach to an old technique," *Microscopy and Microanalysis*, vol. 7, no. S2, pp. 222–223, 2001.

[218] N. J. Zaluzec, "Quantitative measurements of magnetic vortices using position resolved diffraction in Lorentz STEM," *Microscopy and Microanalysis*, vol. 8, no. S02, pp. 376–377, 2002.

[219] J. G. Lozano, G. T. Martinez, L. Jin, P. D. Nellist, and P. G. Bruce, "Low-dose aberration-free imaging of Li-rich cathode materials at various states of charge using electron ptychography," *Nano letters*, vol. 18, no. 11, pp. 6850–6855, 2018.

[220] L. Zhou, J. Song, J. Kim, X. Pei, C. Huang, M. Boyce, L. Mendonça, D. Clare, A. Siebert, C. Allen, E. Liberti, D. Stuart, X. Pan, P. Nellist, P. Zhang, A. Kirkland, and P. Wang, "Low-dose phase retrieval of biological specimens using cryo-electron ptychography," *Nature communications*, vol. 11, no. 1, pp. 1–9, 2020.

[221] C. Ophus, "Four-dimensional scanning transmission electron microscopy (4D-STEM): From scanning nanodiffraction to ptychography and beyond," *Microscopy and Microanalysis*, vol. 25, no. 3, pp. 563–582, 2019.

[222] N. Shibata, S. D. Findlay, Y. Kohno, H. Sawada, Y. Kondo, and Y. Ikuhara, "Differential phase-contrast microscopy at atomic resolution," *Nature Physics*, vol. 8, no. 8, pp. 611–615, 2012.

[223] K. Müller-Caspary, F. F. Krause, T. Grieb, S. Löffler, M. Schowalter, A. Béché, V. Galioit, D. Marquardt, J. Zweck, P. Schattschneider, J. Verbeeck, and A. Rosenauer, "Measurement of atomic electric fields and charge densities from average momentum transfers using scanning transmission electron microscopy," *Ultramicroscopy*, vol. 178, pp. 62–80, 2017.

[224] W. Hoppe, "Beugung im inhomogenen Primärstrahlwellenfeld. III. Amplituden-und Phasenbestimmung bei unperiodischen Objekten," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 25, no. 4, pp. 508–514, 1969.

[225] W. Hoppe and G. Strube, "Beugung in inhomogenen Primärstrahlenwellenfeld. II. Lichtoptische analogieversuche zur Phasenmessung von gitterinterferenzen," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 25, no. 4, pp. 502–507, 1969.

[226] R. Hegerl and W. Hoppe, "Dynamische theorie der kristallstrukturanalyse durch elektronenbeugung im inhomogenen primärstrahlwellenfeld," *Berichte der Bunsengesellschaft für physikalische Chemie*, vol. 74, no. 11, pp. 1148–1154, 1970.

[227] W. Hoppe, "Trace structure analysis, ptychography, phase tomography," *Ultramicroscopy*, vol. 10, no. 3, pp. 187–198, 1982.

[228] H. Yang, R. Rutte, L. Jones, M. Simson, R. Sagawa, H. Ryll, M. Huth, T. Pennycook, M. Green, H. Soltau, Y. Kondo, B. Davis, and P. Nellist, "Simultaneous atomic-resolution electron ptychography and Z-contrast imaging of light and heavy elements in complex nanostructures," *Nature Communications*, vol. 7, no. 1, pp. 1–8, 2016.

[229] A. Faruqi, R. Henderson, M. Pryddetch, P. Allport, and A. Evans, "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*, vol. 546, no. 1-2, pp. 170–175, 2005.

[230] H. Ryll, M. Simson, R. Hartmann, P. Holl, M. Huth, S. Ihle, Y. Kondo, P. Kotula, A. Liebel, K. Müller-Caspary, A. Rosenauer, R. Sagawa, J. Schmidt, H. Soltau, and L. Strüder, "A pnCCD-based, fast direct single electron imaging camera for TEM and STEM," *Journal of Instrumentation*, vol. 11, no. 04, p. P04006, 2016.

[231] A. Faruqi and G. McMullan, "Direct imaging detectors for electron microscopy," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 878, pp. 180–190, 2018.

[232] J. Ciston, I. J. Johnson, B. R. Draney, P. Ercius, E. Fong, A. Goldschmidt, J. M. Joseph, J. R. Lee, A. Mueller, C. Ophus, A. Selvarajan, D. E. Skinner, T. Stezelberger, C. S. Tindall, A. M. Minor, and P. Denes, "The 4D camera: Very high speed electron counting for 4D-STEM," *Microscopy and Microanalysis*, vol. 25, no. S2, pp. 1930–1931, 2019.

[233] H. T. Philipp, M. W. Tate, K. S. Shanks, L. Mele, M. Peemen, P. Dona, R. Hartong, G. van Veen, Y.-T. Shao, Z. Chen, J. Thom-Levy, D. A. Muller, and S. M. Gruner, "Very-High Dynamic Range, 10,000 Frames/Second Pixel Array Detector for Electron Microscopy," *Microscopy and Microanalysis*, vol. 28, no. 2, pp. 425–440, 2022.

[234] I. MacLaren, T. A. Macgregor, C. S. Allen, and A. I. Kirkland, "Detectors— the ongoing revolution in scanning transmission electron microscopy and why this important to material characterization," *APL Materials*, vol. 8, no. 11, p. 110901, 2020.

[235] A. Faruqi and R. Henderson, "Electronic detectors for electron microscopy," *Current opinion in structural biology*, vol. 17, no. 5, pp. 549–555, 2007.

[236] K. C. Bustillo, S. E. Zeltmann, M. Chen, J. Donohue, J. Ciston, C. Ophus, and A. M. Minor, "4D-STEM of beam-sensitive materials," *Accounts of Chemical Research*, vol. 54, no. 11, pp. 2543–2551, 2021.

[237] G. Li, H. Zhang, and Y. Han, "4D-STEM Ptychography for Electron-Beam-Sensitive Materials," *ACS Central Science*, 2022.

[238] R. Egerton, "Dose measurement in the TEM and STEM," *Ultramicroscopy*, vol. 229, p. 113363, 2021.

[239] H. Yang, L. Jones, H. Ryll, M. Simson, H. Soltau, Y. Kondo, R. Sagawa, H. Banba, I. MacLaren, and P. Nellist, "4D STEM: High efficiency phase contrast imaging using a fast pixelated detector," in *Journal of Physics: Conference Series*, vol. 644, p. 012032, IOP Publishing, 2015.

[240] C. O'Leary, C. Allen, C. Huang, J. Kim, E. Liberti, P. Nellist, and A. Kirkland, "Phase

reconstruction using fast binary 4D STEM data," *Applied Physics Letters*, vol. 116, no. 12, p. 124101, 2020.

[241] T. J. Pennycook, G. T. Martinez, P. D. Nellist, and J. C. Meyer, "High dose efficiency atomic resolution imaging via electron ptychography," *Ultramicroscopy*, vol. 196, pp. 131–135, 2019.

[242] P. Binev, W. Dahmen, R. DeVore, P. Lamby, D. Savu, and R. Sharpley, "Compressed sensing and electron microscopy," in *Modelling Nanoscale Imaging in Electron Microscopy*, pp. 73–126, Springer, 2012.

[243] H. S. Anderson, J. Ilic-Helms, B. Rohrer, J. Wheeler, and K. Larson, "Sparse imaging for fast electron microscopy," in *Computational Imaging XI*, vol. 8657, pp. 94–105, SPIE, 2013.

[244] A. Stevens, H. Yang, L. Carin, I. Arslan, and N. D. Browning, "The potential for Bayesian compressive sensing to significantly reduce electron dose in high-resolution STEM images," *Microscopy*, vol. 63, no. 1, pp. 41–51, 2014.

[245] A. Stevens, H. Yang, W. Hao, L. Jones, C. Ophus, P. D. Nellist, and N. D. Browning, "Subsampled STEM-ptychography," *Applied Physics Letters*, vol. 113, no. 3, p. 033104, 2018.

[246] D. Nicholls, J. Lee, H. Amari, A. J. Stevens, B. L. Mehdi, and N. D. Browning, "Minimising damage in high resolution scanning transmission electron microscope images of nanoscale structures and processes," *Nanoscale*, vol. 12, no. 41, pp. 21248–21254, 2020.

[247] B. L. Mehdi, A. Stevens, L. Kovarik, N. Jiang, H. Mehta, A. Liyu, S. Reehl, B. Stanfill, L. Luzi, W. Hao, L. Bramer, and N. D. Browning, "Controlling the spatio-temporal dose distribution during STEM imaging by subsampled acquisition: In-situ observations of kinetic processes in liquids," *Applied Physics Letters*, vol. 115, no. 6, p. 063102, 2019.

[248] D. Nicholls, J. Wells, A. Stevens, Y. Zheng, J. Castagna, and N. D. Browning, "Subsampled imaging for STEM: Maximising image speed, resolution and precision through reconstruction parameter refinement," *Submitted*, 2021.

[249] A. Robinson, D. Nicholls, J. Wells, A. Moshtaghpour, A. Kirkland, and N. D. Browning, "SIM-STEM Lab: Incorporating compressed sensing theory for fast STEM simulation," *Ultramicroscopy*, p. 113625, 2022.

[250] A. W. Robinson, D. Nicholls, J. Wells, A. Moshtaghpour, A. I. Kirkland, and N. D. Browning, "Compressed STEM simulations," *Microscopy and Microanalysis*, vol. 28, no. S1, p. 3116–3117, 2022.

[251] A. W. Robinson, J. Wells, D. Nicholls, A. Moshtaghpour, M. Chi, A. I. Kirkland, and N. D. Browning, "Towards real-time STEM simulations through targeted subsampling strategies," *Journal of microscopy*, vol. 290, no. 1, pp. 53–66, 2023.

[252] A. Stevens, Y. Pu, Y. Sun, G. Spell, and L. Carin, "Tensor-dictionary learning with Deep Kruskal-Factor Analysis," in *Artificial Intelligence and Statistics*, pp. 121–129, PMLR, 2017.

[253] X. Zhang, Z. Chen, and D. Muller, "How many detector pixels do we need for super-resolution ptychography?," *Microscopy and Microanalysis*, vol. 27, no. S1, pp. 620–622, 2021.

[254] J. M. Rodenburg and H. M. Faulkner, "A phase retrieval algorithm for shifting illumination," *Applied physics letters*, vol. 85, no. 20, pp. 4795–4797, 2004.

[255] A. M. Maiden and J. M. Rodenburg, "An improved ptychographical phase retrieval algorithm for diffractive imaging," *Ultramicroscopy*, vol. 109, no. 10, pp. 1256–1262, 2009.

[256] A. Maiden, M. Humphry, M. Sarahan, B. Kraus, and J. Rodenburg, "An annealing algorithm to correct positioning errors in ptychography," *Ultramicroscopy*, vol. 120, pp. 64–72, 2012.

[257] A. M. Maiden, M. J. Humphry, and J. M. Rodenburg, "Ptychographic transmission microscopy in three dimensions using a multi-slice approach," *JOSA A*, vol. 29, no. 8, pp. 1606–1614, 2012.

[258] V. Elser, "Phase retrieval by iterated projections," *JOSA A*, vol. 20, no. 1, pp. 40–55, 2003.

[259] A. D'alfonso, A. Morgan, A. Yan, P. Wang, H. Sawada, A. Kirkland, and L. Allen, "Deterministic electron ptychography at atomic resolution," *Physical Review B*, vol. 89, no. 6, p. 064101, 2014.

[260] R. Bates and J. Rodenburg, "Sub-Ångström transmission microscopy: a Fourier transform algorithm for microdiffraction plane intensity information," *Ultramicroscopy*, vol. 31, no. 3, pp. 303–307, 1989.

[261] J. Rodenburg and R. Bates, "The theory of super-resolution electron microscopy via Wigner-distribution deconvolution," *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, vol. 339, no. 1655, pp. 521–553, 1992.

[262] H. Yang, I. MacLaren, L. Jones, G. T. Martinez, M. Simson, M. Huth, H. Ryll, H. Soltau, R. Sagawa, Y. Kondo, C. Ophus, P. Ercius, L. Jin, A. Kovács, and P. D. Nellist, "Electron ptychographic phase imaging of light elements in crystalline materials using Wigner distribution deconvolution," *Ultramicroscopy*, vol. 180, pp. 173–179, 2017.

[263] G. Martinez, M. Humphry, and P. Nellist, "A comparison of phase-retrieval algorithms for focused-probe electron ptychography," *Microscopy and Microanalysis*, vol. 23, no. S1, pp. 476–477, 2017.

[264] C. M. O'Leary, G. T. Martinez, E. Liberti, M. J. Humphry, A. I. Kirkland, and P. D. Nellist, "Contrast transfer and noise considerations in focused-probe electron ptychography," *Ultramicroscopy*, vol. 221, p. 113189, 2021.

[265] E. Okunishi, H. Sawada, and Y. Kondo, "Experimental study of annular bright field (ABF) imaging using aberration-corrected scanning transmission electron microscopy (STEM)," *Micron*, vol. 43, no. 4, pp. 538–544, 2012.

[266] W. Hoppe, "Beugung im inhomogenen Primärstrahlwellenfeld. I. Prinzip einer Phasenmessung von Elektronenbeugungsinterferenzen," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 25, no. 4, pp. 495–501, 1969.

[267] R. W. Gerchberg, "Phase determination from image and diffraction plane pictures in the electron microscope," *Optik*, vol. 34, pp. 275–284, 1971.

[268] H. N. Chapman, "Phase-retrieval X-ray microscopy by Wigner-distribution deconvolution," *Ultramicroscopy*, vol. 66, no. 3-4, pp. 153–172, 1996.

[269] Z. Hu, Y. Zhang, P. Li, D. Batey, and A. Maiden, "Near-field multi-slice ptychography: quantitative phase imaging of optically thick samples with visible light and X-rays," *Optics Express*, vol. 31, no. 10, pp. 15791–15809, 2023.

[270] R. W. Gerchberg, "A practical algorithm for the determination of plane from image and diffraction pictures," *Optik*, vol. 35, no. 2, pp. 237–246, 1972.

[271] J. R. Fienup, "Reconstruction of an object from the modulus of its fourier transform," *Optics letters*, vol. 3, no. 1, pp. 27–29, 1978.

[272] J. R. Fienup, "Phase retrieval algorithms: a comparison," *Applied optics*, vol. 21, no. 15, pp. 2758–2769, 1982.

[273] P. Thibault, M. Dierolf, O. Bunk, A. Menzel, and F. Pfeiffer, "Probe retrieval in ptychographic coherent diffractive imaging," *Ultramicroscopy*, vol. 109, no. 4, pp. 338–343, 2009.

[274] P. Thibault and M. Guizar-Sicairos, "Maximum-likelihood refinement for coherent diffractive imaging," *New Journal of Physics*, vol. 14, no. 6, p. 063004, 2012.

[275] D. R. Luke, "Relaxed averaged alternating reflections for diffraction imaging," *Inverse problems*, vol. 21, no. 1, p. 37, 2004.

[276] M. Guizar-Sicairos and J. R. Fienup, "Phase retrieval with transverse translation diversity: a nonlinear optimization approach," *Optics express*, vol. 16, no. 10, pp. 7264–7278, 2008.

[277] M. Pham, A. Rana, J. Miao, and S. Osher, "Semi-implicit relaxed douglas-rachford algorithm (sdr) for ptychography," *Optics Express*, vol. 27, no. 22, pp. 31246–31260, 2019.

[278] A. Tripathi, I. McNulty, and O. G. Shpyrko, "Ptychographic overlap constraint errors and the limits of their numerical recovery using conjugate gradient descent methods," *Optics Express*, vol. 22, no. 2, pp. 1452–1466, 2014.

[279] O. Bunk, M. Dierolf, S. Kynde, I. Johnson, O. Marti, and F. Pfeiffer, "Influence of the overlap parameter on the convergence of the ptychographical iterative engine," *Ultramicroscopy*, vol. 108, no. 5, pp. 481–487, 2008.

[280] Z. Zhang, A. Chatterjee, C. Grein, A. J. Ciani, and P. W. Chung, "Molecular dynamics simulation of mbe growth of cdte/znte/si," *Journal of electronic materials*, vol. 40, pp. 109–121, 2011.

[281] B. H. Savitzky, L. Hughes, K. C. Bustillo, H. D. Deng, N. L. Jin, E. G. Lomeli, W. C. Chueh, P. Herring, A. Minor, and C. Ophus, "py4DSTEM: Open source software for 4D-STEM data analysis," *Microscopy and Microanalysis*, vol. 25, no. S2, pp. 124–125, 2019.

[282] B. H. Savitzky, S. E. Zeltmann, L. A. Hughes, H. G. Brown, S. Zhao, P. M. Pelz, T. C. Pekin, E. S. Barnard, J. Donohue, L. R. DaCosta, E. Kennedy, Y. Xie, M. T. Janish, M. M. Schneider, P. Herring, C. Gopal, A. Anapolsky, R. Dhall, K. C. Bustillo, P. Ercius, M. C. Scott, J. Ciston, A. M. Minor, and C. Ophus, "py4DSTEM: A Software Package for Four-Dimensional Scanning Transmission Electron Microscopy Data Analysis," *Microscopy and Microanalysis*, vol. 27, 5 2021.

[283] A. Clausen, D. Weber, K. Ruzaeva, V. Migunov, A. Baburajan, A. Bahuleyan, J. Caron, R. Chandra, S. Halder, M. Nord, *et al.*, "LiberTEM: Software platform for scalable multi-dimensional data processing in transmission electron microscopy," *Journal of Open Source Software*, vol. 5, no. 50, p. 2006, 2020.

[284] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and Y. Akbari, "Image inpainting: A review," *Neural Processing Letters*, vol. 51, pp. 2007–2028, 2020.

A1 | Supplemental Materials

A1.1 Chapter 5

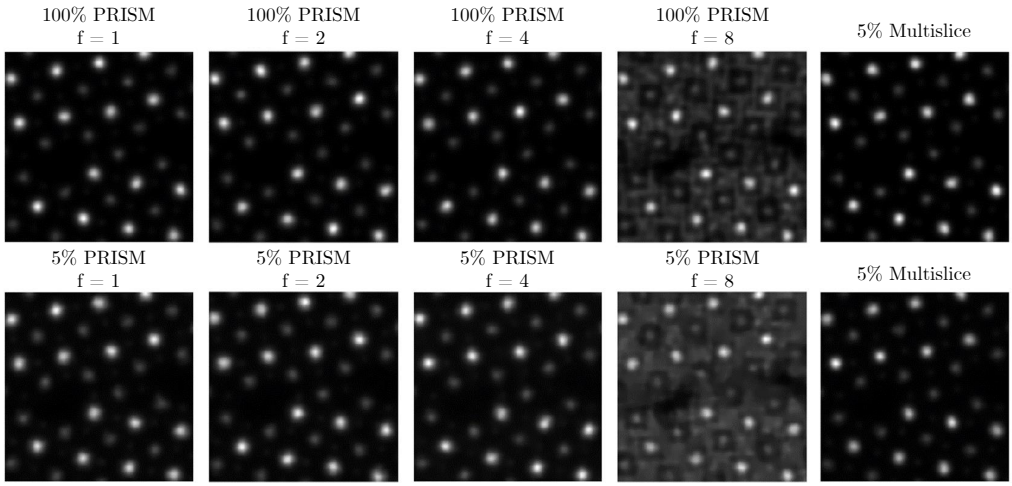

Figure A1.1: **Reconstructions of the $SrTiO_3$ grain boundary simulation using BPFA-EM.** The title of each image corresponds to the sampling ratio used and simulation method respectively.

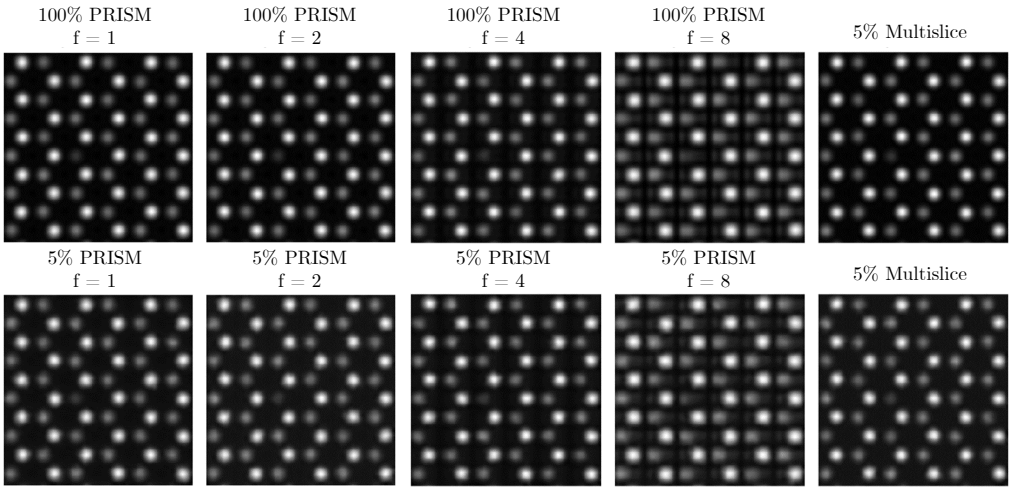

Figure A1.2: **Reconstructions of the 2H-MoS_2 monolayer simulation using BPFA-EM.** The title of each image corresponds to the sampling ratio used and simulation method respectively.

A1.2 Chapter 6

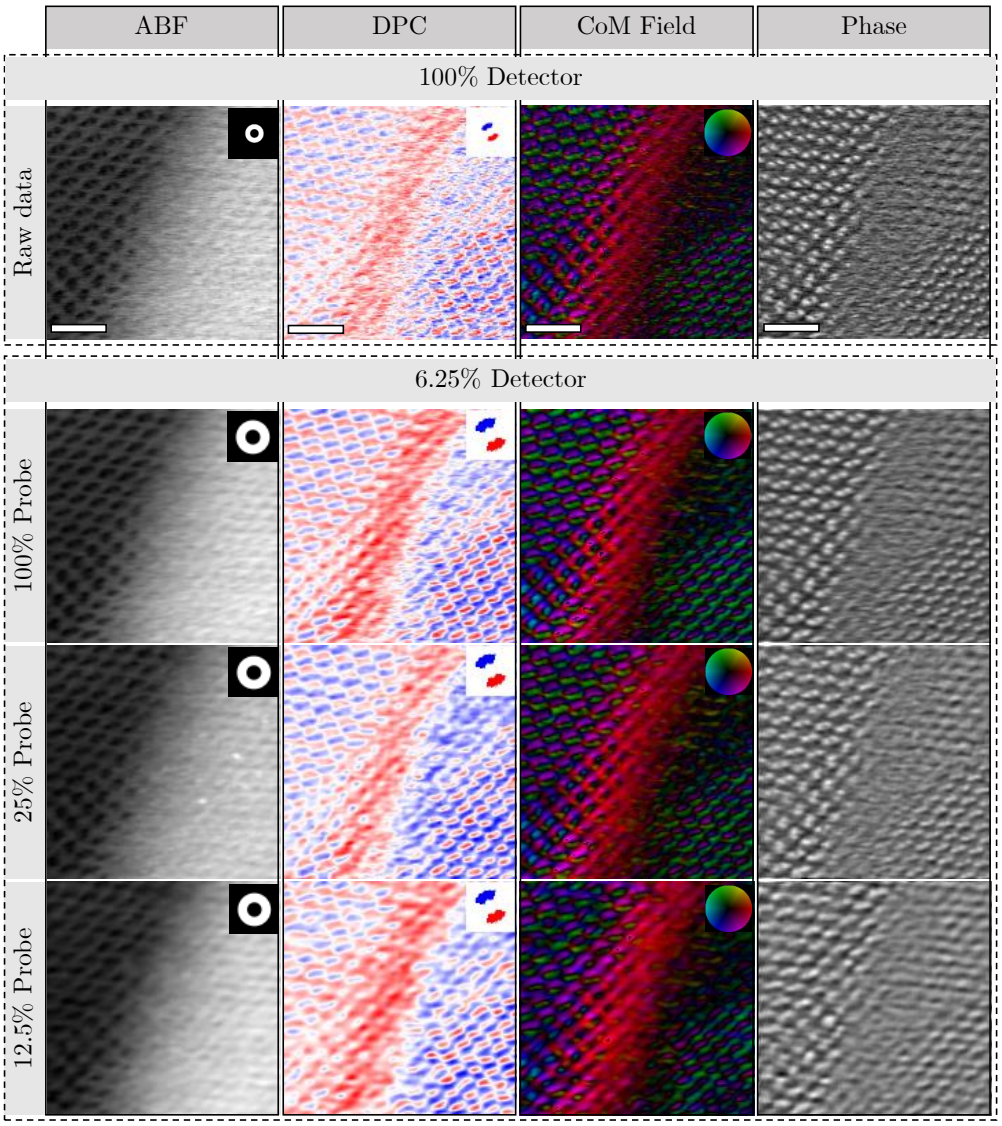

Figure A1.3: **Simulation of sub-sampled 4-D STEM using experimentally acquired 4-D STEM data of a CdTe-Si interface.** Top row shows the ABF, DPC, CoM field, and object phase reconstruction using WDD (from left to right) for the fully sampled, raw data. The remaining rows are then down-sampled on the detector (6.25%) and probe sub-sampled, with the recovery of the data being performed using the BPFA. Scale bar indicates 1nm.

A1.3 Chapter 7

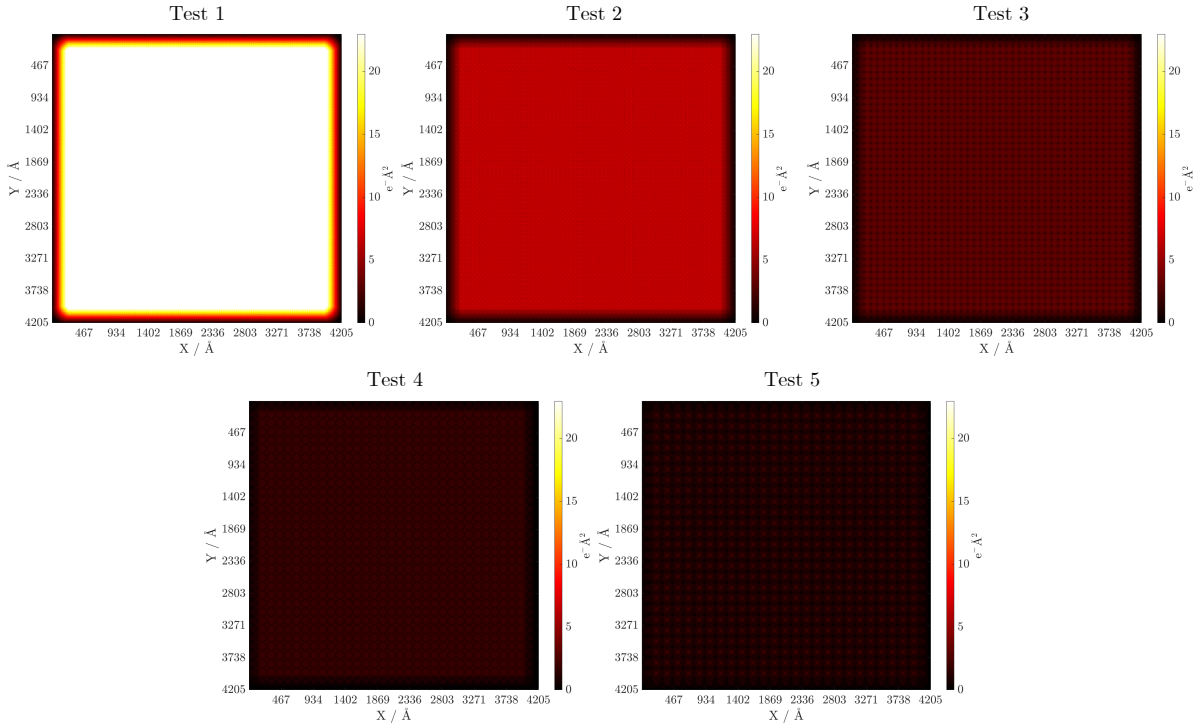

Figure A1.4: **Simulation of the dose distribution for the parameters given in table 7.1.** Title for each dose-distribution map corresponds to a different column in table 7.1, where the estimated fluence is the average intensity across the map. The distributions are computed using the code given in the appendix section A3.2.

A2 | Derivations

A2.1 Expectation and variance of discrete random variables

Let X be a random variable, and let a be a constant. Let $\mu = \mathbb{E}[X]$ be the expectation value for X. Let $f(X) = aX$ be a linear function on X, such that the expectation of $f(X)$ is calculated as,

$$\begin{aligned} \mathbb{E}[f(X)] &= \mathbb{E}[aX] \\ &= a\mathbb{E}[X] \\ &= a\mu \ . \end{aligned} \tag{1}$$

The variance of X, $\mathrm{Var}[X]$ is defined as,

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mu^2 \ , \tag{2}$$

such that the variance of $f(X)$ is calculated as,

$$\begin{aligned} \mathrm{Var}[f(X)] &= \mathbb{E}[f(X)^2] - \mathbb{E}[f(X)]^2 \\ &= a^2 E[X^2] - a^2\mu^2 \\ &= a^2\big(E[X^2] - \mu^2\big) \\ &= a^2 \mathrm{Var}[X] \ . \end{aligned} \tag{3}$$

A2.2 Fourier and real domain constraints in iterative ptychography

This derivations follow from the results given in section 7.2.

A2.2.1 Fourier domain constraints

Let $\Psi^{\mathrm{u}}(\boldsymbol{k}, \boldsymbol{r}_{\mathbf{p}})$ and $\Psi^{\mathrm{c}}(\boldsymbol{k}, \boldsymbol{r}_{\mathbf{p}})$ be the uncorrected and corrected detector waves respectively. The Fourier domain constraint follows the definition of the following convex minimisation problem,

$$\hat{\Psi} = \underset{\boldsymbol{u}}{\operatorname{argmin}} ||\boldsymbol{u} - \Psi||^2 \text{ s.t. } |\boldsymbol{u}| = \sqrt{D} \ , \tag{4}$$

which has a solution given as,

$$\Psi^{\mathrm{c}}(\boldsymbol{k}, \boldsymbol{r}_{\mathbf{p}}) = \sqrt{D(\boldsymbol{k}, \boldsymbol{r}_{\mathbf{p}})} \exp\left(\angle \Psi^{\mathrm{u}}(\boldsymbol{k}, \boldsymbol{r}_{\mathbf{p}})\right) \ . \tag{5}$$

A2.2.2 Real domain constaints

Let $\psi^{\mathrm{u}}(\boldsymbol{r}, \boldsymbol{r}_{\mathbf{p}})$ and $\psi^{\mathrm{c}}(\boldsymbol{r}, \boldsymbol{r}_{\mathbf{p}})$ be the uncorrected and corrected exit waves respectively. Following the forward model given in section 7.2 Eq. 1, the uncorrected and correct exit waves are,

$$\psi^{\mathrm{u}}(\boldsymbol{r}, \boldsymbol{r}_{\mathbf{p}}) = P^{\mathrm{old}}(\boldsymbol{r} - \boldsymbol{r}_{\mathbf{p}}) \cdot O^{\mathrm{old}}(\boldsymbol{r}) \tag{6}$$

$$\psi^{\mathrm{c}}(\boldsymbol{r}, \boldsymbol{r}_{\mathbf{p}}) = P^{\mathrm{new}}(\boldsymbol{r} - \boldsymbol{r}_{\mathbf{p}}) \cdot O^{\mathrm{new}}(\boldsymbol{r}) \ , \tag{7}$$

and by subtracting the uncorrected wave from the corrected wave yields,

$$\psi^{\mathrm{c}}(\boldsymbol{r}, \boldsymbol{r}_{\mathbf{p}}) - \psi^{\mathrm{u}}(\boldsymbol{r}, \boldsymbol{r}_{\mathbf{p}}) = P^{\mathrm{new}}(\boldsymbol{r} - \boldsymbol{r}_{\mathbf{p}}) \cdot O^{\mathrm{new}}(\boldsymbol{r}) - P^{\mathrm{old}}(\boldsymbol{r} - \boldsymbol{r}_{\mathbf{p}}) \cdot O^{\mathrm{old}}(\boldsymbol{r}) \ . \tag{8}$$

From here, two separate assumptions can be made. It can be assumed that *(i)* the new probe is the same as the old probe, or *(ii)* the new probe is a scalar transformation old probe. Considering *(i)* at the object update step, Eq. 8 reduces to,

$$\psi^{\mathrm{c}}(\boldsymbol{r}, \boldsymbol{r}_{\mathbf{p}}) - \psi^{\mathrm{u}}(\boldsymbol{r}, \boldsymbol{r}_{\mathbf{p}}) = P^{\mathrm{new}}(\boldsymbol{r} - \boldsymbol{r}_{\mathbf{p}}) \cdot \left[O^{\mathrm{new}}(\boldsymbol{r}) - O^{\mathrm{old}}(\boldsymbol{r})\right] \ , \tag{9}$$

i.e., fix the probe for object updates. This then generates the new guess for the object as,

$$O^{\text{new}}(\boldsymbol{r}) = \frac{(\psi^{\text{c}}(\boldsymbol{r}, \boldsymbol{r}_{\mathbf{p}}) - \psi^{\text{u}}(\boldsymbol{r}, \boldsymbol{r}_{\mathbf{p}}))}{P^{\text{new}}(\boldsymbol{r} - \boldsymbol{r}_{\mathbf{p}})} + O^{\text{old}}(\boldsymbol{r}) \ , \tag{10}$$

which can be controlled through a learning rate L_α such that,

$$O^{\text{new}}(\boldsymbol{r}) = L_\alpha \frac{(\psi^{\text{c}}(\boldsymbol{r}, \boldsymbol{r}_{\mathbf{p}}) - \psi^{\text{u}}(\boldsymbol{r}, \boldsymbol{r}_{\mathbf{p}}))}{P^{\text{new}}(\boldsymbol{r} - \boldsymbol{r}_{\mathbf{p}})} + O^{\text{old}}(\boldsymbol{r}) \ . \tag{11}$$

The same argument can be used to derive the probe update step. Following this, it is clear from Eq. 11 that the solution is unstable for zero values in $P^{\text{new}}(\boldsymbol{r} - \boldsymbol{r}_{\mathbf{p}})$. To account for this, the ePIE multiplies the numerator and denominator by the complex conjugate of $P^{\text{new}}(\boldsymbol{r} - \boldsymbol{r}_{\mathbf{p}})$, and sets the denominator to the maximum of the value of $|P^{\text{new}}(\boldsymbol{r} - \boldsymbol{r}_{\text{p}})|^2$. A small value, ϵ, can also be added to the denominator to prevent instability, typically at the expense of convergence.

Finally, this results in the ePIE real domain constraints where the probe update is derived by fixing the object as above,

$$P^{\text{new}}(\boldsymbol{r}) = P^{\text{old}}(\boldsymbol{r}) + L_\beta \frac{O^{\text{old}*}(\boldsymbol{r} - \boldsymbol{r}_{\text{p}})}{\max_{\boldsymbol{r}} |O^{\text{old}}(\boldsymbol{r} - \boldsymbol{r}_{\text{p}})|^2} (\psi^{\text{c}}(\boldsymbol{r}, \boldsymbol{r}_{\mathbf{p}}) - \psi^{\text{u}}(\boldsymbol{r}, \boldsymbol{r}_{\mathbf{p}})) \tag{12}$$

$$O^{\text{new}}(\boldsymbol{r}) = O^{\text{old}}(\boldsymbol{r}) + L_\alpha \frac{P^{\text{new}*}(\boldsymbol{r} - \boldsymbol{r}_{\text{p}})}{\max_{\boldsymbol{r}} |P^{\text{new}}(\boldsymbol{r} - \boldsymbol{r}_{\text{p}})|^2} (\psi^{\text{c}}(\boldsymbol{r}, \boldsymbol{r}_{\mathbf{p}}) - \psi^{\text{u}}(\boldsymbol{r}, \boldsymbol{r}_{\mathbf{p}})) \ . \tag{13}$$

A3 | Code Embed

Embedded codes which are discussed in the thesis that are open source.

A3.1 Line hop mask

```
function [mask]  = line_hop_base(row_height,padding,height,width)
    % calculate number of lanes/rows to divide region into
    num_rows = ceil(height / (row_height + padding));
    % reshapes region to this size
    track = zeros([num_rows * (row_height + padding) width]);
    % Initialise the mask
    mask = ones([height width]);
    % Initialise the mask updater
    im = zeros([height width]);
    % Catch if row_height is <=1
    if row_height>1
        % Loop through the lanes/rows
        for row = 0 : num_rows-1
            % First set the upper and lower bounds for row
   position
            row_start = (row) * (row_height + padding);
            row_end = row_start + row_height - 1;
            % If row>0, pick a random start in the bounds
            if row > 0
                row_random = row_start + randi(row_height)-1;
```

```
            else
                % if the first row, set to middle of row
                row_random = round((row_start + row_end) / 2);
            end
            % Stops position being out of range
            if row_random < height
                ypos = row_random;
            else
                ypos = height;
            end
            % loops through columns
            for x = 1:width
                track(ypos+1, x) = 1;
                jump = randi(3);
                % Allows a jump up or down from previous height
                if jump == 1
                    if ypos < row_end && ypos < height-1
                        ypos = ypos + 1;
                    else
                        ypos = ypos - 1;
                    end
                elseif jump == 2
                    if ypos > row_start
                        ypos = ypos - 1;
                    else
                        ypos = ypos + 1;
                    end
                end
            end
            % updates the mask updater
            im = mask.* track(1:height, 1:width);
        end
```

```
    else
        for i=1:num_rows
            r = randi(row_height+padding);
            rin = (i-1)*(row_height+padding) + r;
            im(rin,:) = 1;
        end
    end
    % ensures mask is same size as the target
    mask = im(1:height,1:width);
end
```

A3.2 Dose distribution maps

```
%% Dose estimator for STEM
% Author: Alex W. Robinson, University of Liverpool
% Written: 18/09/2023, Last Modified: 27/09/2023
%-----------------------------------------------------------------%
%%%%%%%%%%%%%% Initialize %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%-----------------------------------------------------------------%
clc,clear,close all

list_factory = fieldnames(get(groot,'factory'));
index_interpreter = find(contains(list_factory,'Interpreter'));
for i = 1:length(index_interpreter)
    default_name = strrep(list_factory{index_interpreter(i)},'
   factory','default');
    set(groot, default_name,'latex');
end

%-----------------------------------------------------------------%
%%%%%%%%%%%%%% Set parameters %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%-----------------------------------------------------------------%
```

```
19
20 res = 1; % resolution of simulation in pixels / angstrom
21 percentage = 1; % fraction of maximum intensity values to carry
      into dose estimate
22 probe_current = 4e-12; % Amperes
23 dwell_time = 1e-3; % seconds
24 scan_step = 31.25; % angstroms
25 scan_dim_x = 127; % number of scans in x
26 scan_dim_y = 127; % number of scans in y
27 convergence_semi_angle = 1.034e-3; % radians
28 defocus = -130000; % angstroms (assume atleast Scherzer defocus)
29 dose_limit = 0; % changes the maximum limit on colorbar e/A^{2}
30 uniform  = false; % unifrom density probe if true
31 sigma_probe = 0.5; % standard deviation of a gaussian probe as
      percentage of probe radius
32
33 %-------------------------------------------------------------%
34 %%%%%%%%%%%%% Create mask? %%%%%%%%%%%%%%%%%%%%%%%%
35 %-------------------------------------------------------------%
36
37 % Uncomment if random sampling
38 % g = 0.25; % sampling ratio if random sampling
39 % % sampling ratio ~ g
40 % mask = rand([scan_dim_y scan_dim_x]);
41 % mask(mask>g) = 0;
42 % mask(mask>0) = 1;
43
44 % Uncomment if down-sampling
45 DS_factor  = 1;
46 %sampling_ratio ~ 1/(DS_factor^2)
47 mask = zeros([scan_dim_y scan_dim_x]);
48 maskx = 1:DS_factor:scan_dim_x;
```

```
masky = 1:DS_factor:scan_dim_y;
for i=1:length(masky)
    y = masky(i);
    for j=1:length(maskx)
        x = maskx(j);
        mask(y,x) = 1;
    end
end


% Uncomment below if using LineHop mask

% addpath('linehop\');
% row_height = 4;
% row_padding = 0;
% % sampling ratio ~ 1/(row_height+row_padding)
% mask = line_hop_main(row_height,row_padding,scan_dim_y,
    scan_dim_x);

%-----------------------------------------------------------------%
%%%%%%%%%%%% Do not change from here below %%%%%%%%%%%
%-----------------------------------------------------------------%
% check for a mask
try
    mask_check = mask(1,1);
catch
    mask = ones([scan_dim_y scan_dim_x]);
end
clear mask_check;
charge = 1.6e-19; % elementary charge, Coulombs
electrons_per_second = probe_current/charge; % number of electrons
    per second
```

```
electrons_per_probe = electrons_per_second.*dwell_time; % electrons per probe
probe_radius = res*abs(defocus)*tan(convergence_semi_angle); % probe radius estimated using defocus value and CSA
fluence_probe = (res^2)*electrons_per_second*dwell_time/(pi*probe_radius^2); % fluence of the probe

% Initialise the output
map = zeros([round((scan_dim_y-1)*res*scan_step+2*probe_radius) round((scan_dim_x-1)*res*scan_step+2*probe_radius)]);

% Initialise the probe box to be added at each acquired location
box = zeros(2*round(probe_radius));
if uniform
    for  i=1:size(box,1)
        for j=1:size(box,2)
            r = sqrt((i-probe_radius-0.5)^2 + (j-probe_radius-0.5)^2);
            if r<=probe_radius
                box(i,j) = fluence_probe;
            end
        end
    end
else
    sigma = probe_radius*sigma_probe; % standard deviation of a gaussian probe
    for  i=1:size(box,1)
        for j=1:size(box,2)
            r = sqrt((i-probe_radius-0.5)^2 + (j-probe_radius-0.5)^2);
            if r<=probe_radius
                box(i,j) = exp(-(r/sigma)^2);
```

```
            end
        end
    end
    box = box./sum(box,"all");
    box = box.*electrons_per_probe; % renormalise probe
end
sbdy = floor(size(box,1)/2);
sbdx = floor(size(box,2)/2);

% For loop over the scanned region and scanned probe locations

for i=1:scan_dim_y
    cr = (i-1)*res*scan_step+probe_radius;
    cr = round(cr);
    for j=1:scan_dim_x
        cc = (j-1)*res*scan_step+probe_radius;
        cc = round(cc);
        if mask(i,j) == 1
            try
                map(cr-sbdy+1:cr+sbdy,cc-sbdx+1:cc+sbdx) = ...
                    map(cr-sbdy+1:cr+sbdy,cc-sbdx+1:cc+sbdx)+ box;
            catch
                maphat = zeros([size(map,1)+1 size(map,2)+1]);
                maphat(1:end-1,1:end-1) = map;
                map = maphat;
                clear maphat;
                map(cr-sbdy+1:cr+sbdy,cc-sbdx+1:cc+sbdx) = ...
                map(cr-sbdy+1:cr+sbdy,cc-sbdx+1:cc+sbdx)+ box;
            end
        else
        end
    end
```

```
end

% Produces output figure, pop-up, and CW output

TF = map > (1-percentage)*max(map,[],'all');
M = mean(map(TF));
max_fluence = max(map,[],"all"); % maximum fluence
truemap = map(1:end-1,1:end-1); % adjusted map to discard edges
mean_fluence = mean(truemap,"all"); % average fluence
disp(append('Average fluence estimation: ',num2str(mean_fluence)))
disp(append('Maximum fluence estimation: ',num2str(max_fluence)))
xticks_in = round(linspace(0,size(truemap,2),10))';
yticks_in = round(linspace(0,size(truemap,1),10))';
f = figure(100);
f.Color = 'w';
imagesc(truemap);axis image;colormap hot;
xlabel('X / \AA','Interpreter','latex');
ylabel('Y / \AA','Interpreter','latex');
xticks(xticks_in);
yticks(yticks_in);
xticklabels(num2str(xticks_in./res));
yticklabels(num2str(yticks_in./res));
dose_estimation = mean(truemap,"all");
set(gca,'FontSize',24);
title('Dose distribution map');
if dose_limit>0
    clim([0 dose_limit]);
end
a = colorbar;
a.Label.Interpreter = 'latex';
a.Label.String = 'e$^{-}$\AA$^{2}$';
a.FontSize = 24;
```

```
168 f = msgbox(append("Average: ",num2str(mean_fluence),"e/A^2.
      Maximum: ",num2str(max_fluence),"e/A^2"),"Fluence");
```