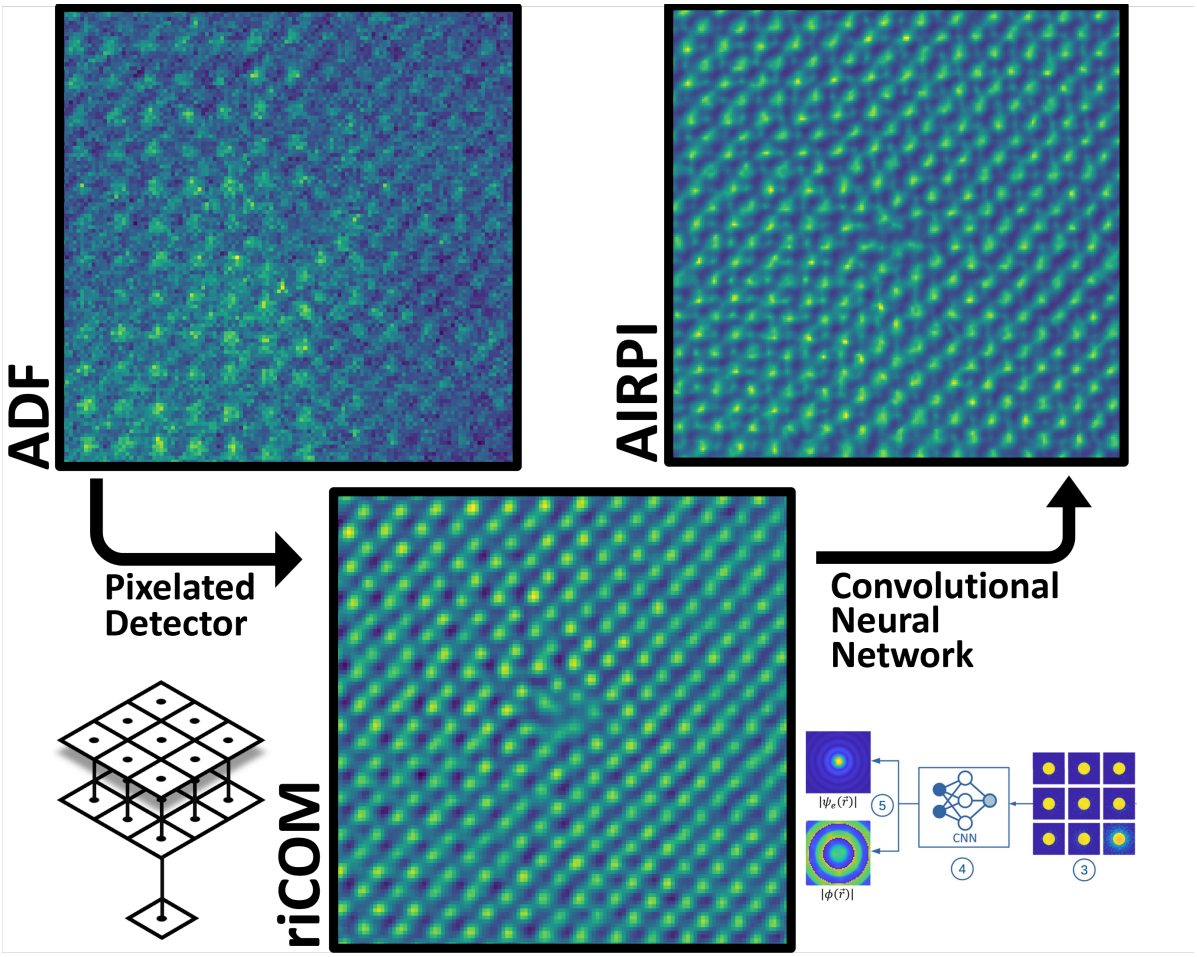
Novel Imaging Methods of Transmission Electron Microscopy Based on Electron Beam Scattering and Modulation

Chu-Ping Yu



Supervisors **prof. dr. Johan Verbeeck | prof. dr. Sandra Van Aert**

Thesis submitted for the degree of Doctor of Science: Physics
Faculty of Science | Antwerp, 2023

University of Antwerp

Faculty of Science

Novel Imaging Methods of Transmission Electron Microscopy Based on Electron Beam Scattering and Modulation

---

Vernieuwende Beeldvormingsmethoden voor Transmissie Elektronen Microscopie Gebaseerd op Verstrooiing en Modulatie van de Elektronenbundel

Thesis submitted for the degree of
Doctor of Science: Physics
at the University of Antwerp

Chu-Ping Yu

Antwerp, 2023

Supervisors
prof. dr. Johan Verbeeck
prof. dr. Sandra Van Aert

Jury

Chairman
prof. dr. Bart Partoens, University of Antwerp, Belgium

Supervisors
prof. dr. Johan Verbeeck, University of Antwerp, Belgium
prof. dr. Sandra Van Aert, University of Antwerp, Belgium

Members
prof. dr. Jan De Beenhouwer, University of Antwerp, Belgium
prof. dr. Peter Nellist, University of Oxford, U.K.
prof. dr. Knut Müller-Caspary, Ludwig Maximilian University of Munich, Germany

Contact
Chu-Ping Yu
University of Antwerp
Faculty of Science
Department of Physics
EMAT
Groenenborgerlaan 171, 2020 Antwerpen, België
M: chu-ping.yu@uantwerpen.be

ISBN 987-90-57285-34-7
Wettelijk depot D/2022/12.293/03

Summary

Transmission Electron microscopy (TEM) encompasses a diverse range of techniques that employ an electron beam to investigate materials, exploring their morphology, shape, and intricate details at a spatial resolution as small as the size of an atom. From the perspective of the probing mechanism, TEM can also be regarded as the examination of the scattering of the fast electrons and the subsequent products as they interact with the studied matter. Through the utilization of various hardware and analyzing methods, the outcomes of scattering can be recorded and analyzed, yielding valuable data that characterize the investigated objects and even facilitating the quantification of certain properties.

The progress in the hardware has enhanced the efficiency of data acquisition, resulting in larger volumes of data with increased complexity. Despite these advancements, the phase information of the transmitted electron wave, which is related to the elastic scattering of the fast electron, is not easily accessible, even with state-of-the-art equipment. This information, nevertheless, can be extracted from the recorded electron scattering pattern by specific algorithms or imaging modes.

Ptychography, phase contrast imaging, and phase plates are among some of the methods or tools that leverage the phase information carried by the electron wave to achieve a deeper understanding of the investigated materials. These approaches involve either reconstructing the object that introduces the phase shift or forcing the phase information to be reflected on the amplitude variations of the wave, which can be directly recorded. While the aforementioned methods are widely recognized for their high dose efficiency and enhanced contrast for weak scatterers, it is important to highlight two aspects that have received comparatively less attention within the field of phase-related methodologies.

The first aspect concerns real-time imaging. Traditional imaging modes of TEM allow data from each acquisition step to be interpreted separately and independently of each other and thus enable live feedback during the data acquisition. On the other hand, the majority of these phase-related imaging methods, with very few exceptions, lack the ability to update the reconstruction results during the data collection process. This limitation is due to either extensive processing time or the absence of ways to restate the algorithm in a streaming workflow where partial incoming data gets processed as soon as it is available, and successive updates of the result are made. Since many operations on the microscope, such as the navigation on the sample plane and the adjustment of the optical condition, rely on live feedback, these phase-related methods cannot fully replace the traditional imaging modes despite their high dose efficiency. As a consequence, TEMs are often operated with less dose-efficient imaging modes, and only the region of interest (ROI) is "recorded" at lower doses when it comes to beam-sensitive materials, which is clearly a non-ideal solution. As studies in the field of dose-sensitive materials gain significant awareness over the years, there is a pressing need for a compatible imaging

method that combines good dose efficiency, process efficiency, and strong contrast from light elements.

The second aspect pertains to the design of the probing waveform. While the phase shift of the transmitted wave is typically the focus of the imaging or measuring purposes of the TEM, it is also important to investigate the design of the incident wave that can efficiently carry the desired information. Indeed, there is a surge of interest in waveform manipulation in TEM that leads to the creation of fulfilling physical devices. Novel measurement techniques based on the designed waveforms have also been proposed. However, these complex wave modulators have seen rather little improvements in their adaptability, making them suitable only for specific purposes or the creation of certain waveforms. Sample inspection or TEM operation that utilizes information delivered by multiple waveforms has not yet been demonstrated. The potential and new applications that can be realized with adaptive probing electron waves need to be explored further.

This thesis is thus dedicated to progress in these aspects:

1. The first chapter gives a comprehensive overview of the fundamental principles of TEM that are relevant to the topics. It includes a description of the elastic scattering of fast electrons and elucidates how the phase shift of the wave is introduced as a result. Additionally, a concise introduction to various TEM setups is presented, including the different operation modes and parameters that can be tuned to optimize the outcomes of an experiment. A discussion on imaging and data acquisition, as well as the corresponding hardware, are included, involving different types of detectors and their advantages and disadvantages.

2. The second chapter describes various methods that are employed or referenced in this thesis. First of all, the simulation package and the model describing the scattering event and wave propagation are introduced. Subsequently, descriptions of a series of imaging methods that utilize the phase information of the electron wave are presented. The chapter also includes an introduction to statistical approaches that are used to analyze data and obtain quantitative information that cannot be intuitively acquired. Besides, the concept of the hypothesis test and how the design of a TEM experiment can benefit from such tests is elaborated.

3. In the third chapter, an imaging method that exhibits efficiency in both dose usage and data processing is introduced. The processing of the data and how real-time imaging is achieved by decomposing the algorithm are discussed in detail. The contrast transfer function of the imaging method is explained, including its connection with the choice of parameter and warnings of possible misinterpretation of the contrast. The efficiency of the algorithm is demonstrated through the processing of pre-recorded data, as well as from data streamed directly from the detector.

4. The fourth chapter builds upon the concept of real-time imaging proposed in the third chapter with a more intricate procedure. A novel reconstruction method is proposed to handle the diffraction patterns obtained from a TEM experiment. Inspired by ptychography, patterns acquired from proximity are processed to retrieve phase images of the inspected object with the help of AI. The AI model is trained to work on finite data size, enabling the initiation of image reconstruction once a small amount of data has been collected, thus enabling live feedback during the data acquisition.

5. The fifth chapter explores the idea of modulating the electron wave through the implementation of a programmable phase plate. The phase plate is composed of multiple openings, and in each opening, a confined electrostatic potential can be created with a tunable voltage source. As a consequence, the phase shift can be programmed for each partial wave. Such ability is crucial for the creation of orthogonal wave sets for sample inspection. Furthermore, the potential of the device in various research fields is discussed and demonstrated.

6. In the sixth chapter, the possibility to distinguish atom species based on the elastic scattering of the fast electrons is investigated. As each type of element establishes a different potential profile, the scattering result of the passing wave would also be different, and thus atom differentiation can be achieved by observing the acquired wave intensity. By formulating the process of distinguishing waveforms as a specific measurement of quantum states, criteria to maximize the difference between the states can be derived. With that, designs of probing electron waves are proposed and tested with simulation. The benefit of such designs for imaging larger features is also discussed.

7. The seventh chapter summarizes the contents of the thesis and provides an outlook on future developments in this domain.

Samenvatting

Transmissie elektronen microscopie (TEM) omvat een breed scala aan technieken die een elektronenbundel gebruiken om materialen te onderzoeken en hun morfologie, vorm en details te verkennen met een ruimtelijke resolutie zo klein als de grootte van een atoom. Vanuit het perspectief van het sondering mechanisme kan TEM ook worden beschouwd als het onderzoek naar de verstrooiing van de snelle elektronen en de daaropvolgende producten waneer ze interageren met de bestudeerde materie. Door gebruik te maken van verschillende hardware en analysemethoden kunnen de uitkomsten van verstrooiing worden geregistreerd en geanalyseerd. Dat levert waardevolle gegevens op die de onderzochte objecten karakteriseren en zelfs de kwantificering van bepaalde eigenschappen mogelijk maken.

De vooruitgang in de hardware heeft de efficiëntie van gegevensverzameling verbeterd, wat resulteert in grotere hoeveelheden gegevens met verhoogde complexiteit. Ondanks deze vooruitgang is de fase-informatie van de transmissie elektronengolf, die gerelateerd is aan de elastische verstrooiing van het snelle elektron, zelfs met state-of-the-art apparatuur niet eenvoudig toeganklijk. Deze informatie kan echter worden geëxtraheerd uit het geregistreerde elektronnenverstrooiingspatroon door middel van specifieke algoritmen of beeldvormingsmodi.

Ptychography, fasecontrastbeeldvorming en faseplaten behoren tot enkele van de methoden of tools die een dieper inzicht in de onderzochte materialen aan te bieden door de fase-informatie die door de elektronengolf wordt gedragen. Deze benaderingen betreffen ofwel het reconstrueren van het object dat de faseverschuiving introduceert, of het forceren van de fase-informatie om te worden weerspiegeld op de amplitudevariaties van de golf, die direct kunnen worden geregistreerd. Hoewel de bovengenoemde methoden algemeen worden erkend voor hun hoge dosisefficiëntie en verhoogd contrast voor zwakke verstrooiers, is het belangrijk om twee aspecten te benadrukken die in het veld van fase-gerelateerde methodologieën relatief minder aandacht hebben gekregen.

Het eerste aspect betreft real-time beeldvorming. Traditionele beeldvormingsmodi van TEM laten gegevens uit elke acquisitiestap toe om afzonderlijk en onafhankelijk van elkaar te worden geïnterpreteerd en bieden zo live feedback tijdens de gegevensacquisitie. Aan de andere kant hebben de meeste van deze fase-gerelateerde beeldvormingsmethoden, op enkele uitzonderingen na, niet de mogelijkheid om de reconstructieresultaten bij te werken tijdens het gegevensverzamelingsproces. Deze beperking is te wijten aan een te lange verwerkingstijd of het ontbreken van manieren om het algoritme te herschrijven in een "streaming workflow" waarbij gedeeltelijke inkomende gegevens worden verwerkt zodra ze beschikbaar zijn en opeenvolgende updates van het resultaat worden gemaakt. Aangezien veel bewerkingen op de microscoop, zoals navigatie op het monsteroppervlak en aanpassing van de optische toestand, afhankelijk zijn van live feedback, kunnen deze fase-gerelateerde methoden de traditionele beeldvormingsmodi niet volledig vervangen,

ondanks hun hoge dosisefficiëntie. Als gevolg daarvan worden TEM's vaak bediend met minder dosisefficiënte beeldvormingsmodi, en wordt alleen het interessegebied "geregistreerd" bij lagere doses als het gaat om stralingsgevoelige materialen, wat duidelijk een niet-ideale oplossing is. Aangezien studies op het gebied van dosisgevoelige materialen in de loop der jaren aanzienlijke aandacht hebben gekregen, is er een dringende behoefte aan een compatibele beeldvormingsmethode die goede dosisefficiëntie, procesefficiëntie en sterk contrast op de lichte elementen combineert.

1. Het eerste hoofdstuk geeft een uitgebreid overzicht van de fundamentele principes van TEM die relevant zijn voor de onderwerpen. Het bevat een beschrijving van de elastische verstrooiing van snelle elektronen en legt uit hoe de faseverschuiving van de golf hierdoor wordt geïntroduceerd. Daarnaast wordt een beknopte introductie van verschillende TEM-opstellingen gepresenteerd, inclusief de verschillende werkingsmodi en parameters die kunnen worden ingesteld om de resultaten van een experiment te optimaliseren. Er wordt ingegaan op beeldvorming en gegevensverwerving en de bijbehorende hardware, waarbij verschillende soorten detectoren en hun voor- en nadelen aan bod komen.

2. Het tweede hoofdstuk beschrijft verschillende methoden die in dit proefschrift worden gebruikt of waarnaar wordt verwezen. Allereerst worden het simulatiepakket en het model dat de verstrooiing en golfvoortplanting beschrijft geïntroduceerd. Vervolgens worden beschrijvingen gegeven van een reeks beeldvormingsmethoden die gebruik maken van de fase-informatie van de elektronengolf. Het hoofdstuk bevat ook een inleiding tot statistische benaderingen die worden gebruikt om gegevens te analyseren en kwantitatieve informatie te verkrijgen die niet intuïtief kan worden bekomen. Daarnaast wordt het concept uitgewerkt die gaat over de hypothesetest en hoe het ontwerp van een TEM-experiment dergelijke tests kan benutten.

3. In het derde hoofdstuk wordt een beeldvormingsmethode geïntroduceerd die efficiënt is in zowel dosisgebruik als gegevensverwerking. De verwerking van de gegevens en hoe real-time beeldvorming wordt bereikt door het algoritme te ontleden, worden in detail besproken. De "contrast transfer function" van de beeldvormingsmethode wordt uitgelegd, inclusief het verband met de parameterkeuze en waarschuwingen voor mogelijke misinterpretatie van het contrast. De efficiëntie van het algoritme wordt aangetoond aan de hand van de verwerking van vooraf opgenomen gegevens en gegevens die rechtstreeks van de detector worden gestreamd.

4. Het vierde hoofdstuk bouwt voort op het concept van real-time beeldvorming zoals voorgesteld in het derde hoofdstuk met een meer ingewikkelde procedure. Er wordt een nieuwe reconstructiemethode voorgesteld om de diffractiepatronen van een TEM-experiment te verwerken. Geïnspireerd door ptychography, worden patronen verkregen uit nabijheid verwerkt om fasebeelden van het geïnspecteerde object op te halen met behulp van AI. Het AI-model is getraind om te werken met een beperkte hoeveelheid gegevens, wat het mogelijk maakt om de beeldreconstructie te starten zodra een kleine hoeveelheid gegevens is verzameld, waardoor live feedback tijdens de gegevensverzameling mogelijk is.

5. Het vijfde hoofdstuk onderzoekt het idee om de elektronengolf te moduleren door middel van een programmeerbare faseplaat. De faseplaat bestaat uit meerdere

openingen en in elke opening kan een begrensde elektrostatische potentiaal gecreëerd worden met een afstembare spanningsbron. Hierdoor kan de faseverschuiving voor elke partiële golf geprogrammeerd worden. Deze mogelijkheid is cruciaal voor het creëren van orthogonale golfsets voor monsterinspectie. Verder wordt het potentieel van het apparaat in verschillende onderzoeksgebieden besproken en gedemonstreerd.

6. In het zesde hoofdstuk wordt de mogelijkheid onderzocht om atoomsoorten te onderscheiden op basis van de elastische verstrooiing van de snelle elektronen. Aangezien elk type element een ander potentiaalprofiel vaststelt, zou het verstrooiingsresultaat van de passerende golf ook verschillend zijn, en dus kan atoomdifferentiatie worden bereikt door de verworven golfintensiteit te observeren. Door het proces van het onderscheiden van golfvormen te formuleren als een specifieke meting van kwantumtoestanden, kunnen criteria worden afgeleid om het verschil tussen de toestanden te maximaliseren. Daarmee worden ontwerpen voor het sondering van elektronengolven voorgesteld en getest met simulatie. Het voordeel van dergelijke ontwerpen voor het afbeelden van grotere eigenschappen wordt ook besproken.

7. Het zevende hoofdstuk vat de inhoud van het proefschrift samen en geeft een vooruitblik op toekomstige ontwikkelingen in dit domein.

Contents

Chapter 1

Introduction to Transmission Electron Microscopy

Humans use their senses to perceive the world. They hear and see objects that emit or interfere with the transmission of sound or light, and build primitive knowledge about things that can be observed in these ways. Humans are enlightened by their senses but also limited by them. We do not possess the receptors to magnetism, humidity, and not even light with wavelengths outside a narrow band that can trigger a signal in the eyes. We do not naturally possess the ability to observe things that are far away or much smaller than us.

To perceive these objects, we rely on the use of tools. Telescopes, microscopes, etc., are tools that provide access to information that was once impossible to acquire. Celestial bodies lightyears away from us can be recorded with telescopes, and their motions, transformations, and interactions with each other are great teachers that reveal phenomena that cannot be observed on Earth. In contrast, on the micro- and nanoscale, Tissues, cells, viruses, and proteins can be imaged with microscopes, which leads to tremendous advances in life science and medicine. New theories, hypotheses, and inventions as a result of breakthroughs in the technology of observation gradually change our lives and allow the creation of even more sophisticated instruments to probe the unknowns.

Electric fields or potentials provide rich knowledge about nature since most of the substances that we know are composed of atoms, and atoms create electric fields. Despite being another example of things that humans cannot naturally perceive, an electric field can be probed by charged particles such as electrons. The short wavelength of high-energy electrons is beneficial for the probing of minuscule objects, as the shortest resolvable distance is proportional to the wavelength, according to the Abbe diffraction limit. Besides, electrons can be generated, controlled, and detected relatively easily, compared to other charged particles, such as protons or helium nuclei. With all these advantages, the development of electron microscopes (EM) has seen great success in achieving imaging abilities unmatched by other contemporary microscopic techniques. Such a powerful tool naturally attracts applications in multiple fields of research. Various types of EM with different sets of abilities and operating conditions are thus created for different purposes, such as scanning electron microscopes (SEM), and transmission electron microscopes (TEM). Each of them focuses on the generation and collection of some kind of signal, produced by the interaction between the probing electrons and the sample.

In general the results of the interaction or scattering can be categorized as elastic and inelastic by whether energy exchange is involved. While the results of the latter are commonly used across various types of EMs, for instance, the emission of characteristic X-ray [1] and the energy loss of the fast electron [2] corresponding to the excitation of specific atomic orbitals are used to characterize the composing elements of the material, what separates TEM from other electron microscopy is the utilization of the coherent wave nature of the elastically scattered electrons, which is preserved as a result of the limited sample thickness.

Elastic scattering comes from interactions that do not involve energy exchange, but the momentum of the fast electrons, however, can be altered by the scattering, thus changing their propagation direction. Since the energy and wavelength of the electrons after the scattering remain unchanged, the temporal coherence of the electron wave is preserved, and the result of elastic scattering can thus be described as changes in the coherent waveform, making use of the particle-wave duality. As a result, the recorded intensity of the wave contains information about the interacting sample and can be processed to form an image of the illuminated sample region with spatial resolution varying from sub-angstrom level to tens of micrometers.

This chapter is dedicated to providing a basic description of elastic scattering and describes how the sample influences the electron waveform. The essential components of a TEM and its different setups are discussed, including their influence on the properties of the probing electron wave. Additionally, a few common imaging modes are introduced, as well as various types of detectors to record the result of scattering.

1.1 Elastic Scattering

The kinetic energy of an electron wave is influenced by the potential variations along its path. However, in a typical electron microscopy setup, the kinetic energy is preserved as the electrostatic potential before and after the sample is kept at zero, and therefore the only permanent effect of this passage through the potential landscape of the sample is a phase modulation, which is found to be [3]:

$$\Delta\varphi(\vec{r}) = \frac{\pi}{\lambda E}\int_{-\infty}^{\infty} V(\vec{r},z)\,dz - \frac{e}{\hbar}\int_{-\infty}^{\infty} A_z(\vec{r},z)\,dz \tag{1.1}$$

where λ and E are the wavelength and energy of the electron wave, e is the electric charge carried by a single electron, $\hbar$ the reduced Planck's constant, and $\Delta\varphi(\vec{r})$ describes the resulting 2D phase profile with the assumption that the electrons travel straight in the z direction. This phase profile is a function of the spatial distribution of the electric potential $V(\vec{r},z)$ and the magnetic vector potential $A_z(\vec{r},z)$.

The first term in equation 1.1 corresponds to the influence of the local electrostatic potential. Despite that the positive potential of the nuclei is almost completely shielded by the electrons on a macroscopic scale, the combined potential landscape on the microscopic scale as a result of the charges of both signs can be experienced by the electron as it passes through the material, introducing phase shift to the passing wave. The second

term describes the effect of the magnetic vector potential, which creates contrasts that originate from several magnetic effects in the sample, such as domain walls [4, 5], magnetic vortices [6], and skyrmions [7]. Since this acquired phase is a direct result of the potential distribution in the material, one would be able to extract valuable information, such as the locations of the atoms in the sample, if the phase shift can be measured in some manner.

1.2 Creating an Electron Beam

In a TEM, the probing electron wave that can carry such phase information is created in the form of certain beam shapes. This beam is created by selecting the electrons emitted from the electron source with an aperture, limiting it to contain only electrons that are almost perfectly parallel to the optical axis to ensure the beam is formed with electrons from a spatially coherent source [8]. This beam is further processed by the condenser system into a desired shape and size before it propagates to the sample plane and interacts with the investigated material. The process of creating an electron beam is summarized in figure 1.1.

Other parameters of the TEM setup also influence the properties of the electron beam. To maximize the information one can obtain from a TEM experiment, as well as minimize unwanted side effects, these parameters need to be selected carefully. The following discussion covers factors that determine the beam's properties, including the operation modes of the TEM, the angle, the energy, and the current of the beam. Additionally, the corresponding TEM components that allow the adjustment of these parameters are also introduced.

Operation modes

Depending on whether a parallel beam or a convergent beam is desired, the condenser system creates the image of the aperture or the diffraction pattern of the aperture at the sample plane, respectively, as the probing electron wave.

The creation of a convergent beam can be described by Fraunhofer diffraction or far-field diffraction. When the convergence angle is sufficiently small to fulfill the paraxial approximation, which is indeed the case for most of the TEM, the electron wave on the aperture plane $\psi_{ap}(\vec{k})$ and the incident wave on the sample plane $\psi_i(\vec{r})$, are connected by an inverse Fourier transform IFT:

$$\begin{aligned}\psi_i(\vec{r}) &= \int_{-\infty}^{\infty} \psi_{ap}(\vec{k}) e^{i2\pi\vec{k}\cdot\vec{r}}\, d^2\vec{k} \\ &= \int_{-\infty}^{\infty} A_{con}(\vec{k}) e^{i\chi(\vec{k})+i2\pi\vec{k}\cdot\vec{r}}\, d^2\vec{k} \\ &= \mathrm{IFT}\left[\psi_{ap}(\vec{k})\right]\end{aligned} \tag{1.2}$$

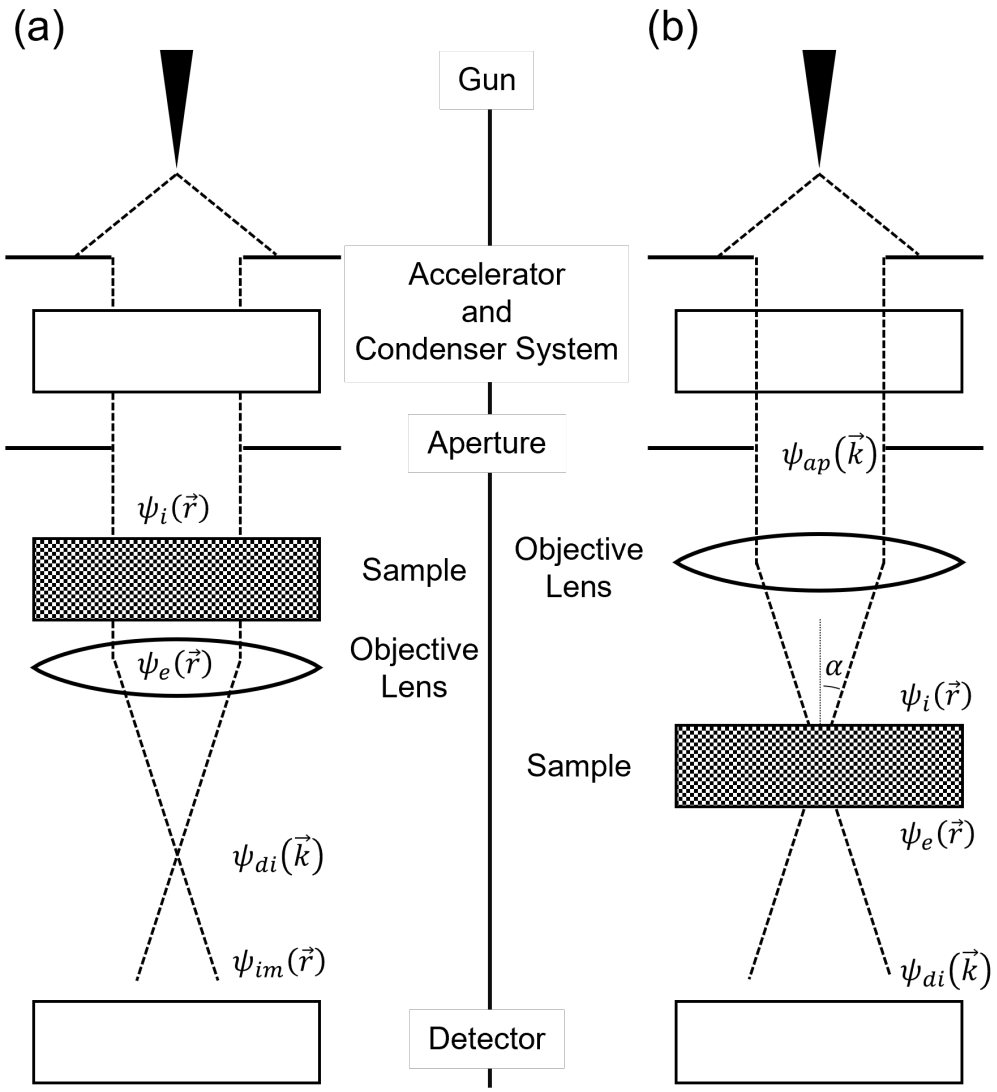


Figure 1.1: Common operation modes and components of a TEM. (a) Parallel illumination utilizes a broad beam to interact with a relatively large area of the sample. (b) A convergent beam with a convergence angle α creates a sharp probe at the sample plane. The wave at the condenser aperture plane $\psi_{ap}(\vec{k})$, incident wave $\psi_i(\vec{r})$ and exit wave $\psi_e(\vec{r})$ at the sample plane, and the transmitted wave at diffraction plane $\psi_{di}(\vec{k})$ and image plane $\psi_{im}(\vec{r})$ are indicated.

where $A_{con}(\vec{k})$ is the condenser lens aperture that crops the passing wave, and $\chi(\vec{k})$ describes the aberrations in the condenser system.

The waveform $\psi_i(\vec{r})$ of a convergent beam or a parallel beam on the sample plane thus describes the probing wave that directly interacts with the sample. The result of this interaction is directly written onto the passing wave, creating a different waveform on the exit surface of the sample $\psi_e(\vec{r})$. The effect of this interaction will be covered in more detail in the next chapter, where the concept of multislice simulation is introduced. After further propagation, the exit wave forms a diffraction pattern on the back focal plane of the lower objective lens. The process is again described by Fraunhofer diffraction, which can be simplified to a Fourier transform FT of the wave when the outgoing angle is small:

$$\begin{aligned}\psi_{di}(\vec{k}) &= \int_{-\infty}^{\infty} \psi_e(\vec{r}) e^{-i2\pi\vec{k}\cdot\vec{r}}\, d^2\vec{r} \\ &= \mathrm{FT}\left[\psi_e(\vec{r})\right]\end{aligned} \tag{1.3}$$

In equation 1.3, the transmitted wave $\psi_e(\vec{r})$ and $\psi_{di}(\vec{k})$ are found on the sample plane and the diffraction plane, respectively. The waveform of the exit wave can be recreated on the image plane of the objective lens by taking another Fourier transform of the waveform on the diffraction plane:

$$\begin{aligned}\psi_{im}(\vec{r}) &= \int_{-\infty}^{\infty} \psi_{di}(\vec{k}) \exp\left[i2\pi\vec{k}\cdot\vec{r}\right]\, d^2\vec{k} \\ &= \int_{-\infty}^{\infty} A_{ob}(\vec{k}) \exp\left[i\chi(\vec{k}) + i2\pi\vec{k}\vec{r}\right]\, d^2\vec{k} \\ &= \mathrm{IFT}\left[\psi_{di}(\vec{k})\right]\end{aligned} \tag{1.4}$$

The term $A_{ob}(\vec{k})$ only exists when an objective aperture is used, $\chi(\vec{k})$ describes the aberration from the objective lens, and the vector $\vec{r'}$ describes the plane where the copy of the exit wave is recorded.

Convergence angle

The convergence angle α of a beam is defined by the size of the probe forming aperture and also how strong the condenser lens is excited (see figure 1.1). The convergent angle is related to the distance from the optical axis on the condenser aperture plane $|k|$ by equation 1.5:

$$\alpha = |k|\lambda \tag{1.5}$$

Assuming a round aperture and ignoring lens aberrations, the resulting beam can be described by a cylinder function on the aperture plane, and its diffraction on the sample plane is found by Fourier transform to be an Airy disc (figure 1.2 and equation 1.6).

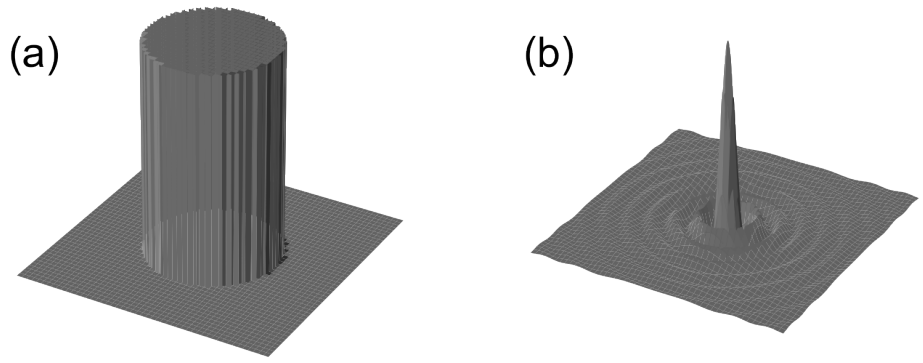


Figure 1.2: (a) The cropped wave by a round aperture can be described by a cylinder function with uniform intensity inside the opening of the aperture and zero intensity outside. (b) The diffraction of the aperture image forms the probe, which can be described by an Airy disc.

$$\begin{aligned} \psi_a p(k) &= \begin{cases} 1, & k < k_{max} \\ 0, & \text{else} \end{cases} \\ \psi_i(r) &= \frac{J_1(\pi r k_{max})}{\pi r k_{max}} \end{aligned} \tag{1.6}$$

where k_{max} is the largest angle allowed by the aperture and J_1 is the first order Bessel function of the first kind. It is seen from equation 1.6 that the width of the cylinder is inversely related to the width of its transformation. It describes the diffraction limit created by the finite aperture size, which also defines the sharpness of the probe at a given convergence angle. However, electrons at larger angles are more susceptible to a stronger aberration effect from the imperfect optical system, resulting in a spread of the focused electron probe. Therefore, the sharpest electron probe is normally not created by maximizing the convergence angle but rather by finding the best trade-off between the opening angle and the aberrations [9].

Energy

The acceleration voltage of the TEM determines the energy of the electron beam, and its wavelength is then found to be:

$$\lambda = \frac{h}{\sqrt{2m_e E\left(1 + \frac{E}{2m_e e^2}\right)}} \tag{1.7}$$

where m_e is the rest mass of the electron, c is the speed of light, and h the Planck's constant. The kinetic energy E is equal to eV_a where e is the electron charge and V_a is the acceleration voltage.

From equations 1.2 and 1.5, it shows that one could create a sharper probe by increasing the energy of the electron beam to reduce the wavelength and achieve a larger q_{max} with a given convergence angle. Besides, higher kinetic energy also gives the electron beam a smaller cross-section to elastic and certain inelastic scattering, such as radiolysis of organic material [10]. However, for certain materials where dislodging of the atoms is prone to happen under the illumination of high-energy electron beam, a lower acceleration voltage may be favored to protect the sample[11]. Lower scattering cross-section also gives high energy electron beam better penetration depth, which benefits the studies of samples that cannot easily be made thin.

The lower scattering cross-section for the elastic scattering is manifested by equation 1.1, which states that the phase shift resulting from the projected potential is inversely proportional to the product of λE, which is positively related to the electron energy E, indicating that when interacting with the same material, an electron beam of higher energy would result in less phase shift. When imaging a sample that is entirely composed of light elements, such as tissue sections and other biomaterials, a low-energy electron beam can be more effective by creating stronger contrast between the weak scatterers [12, 13]. Interests in exploiting such advantage of low energy electron imaging contribute to the driving force in designing and making TEMs that operate at voltages as low as 5 keV [14].

Current

The electron beam is generated by the emission of electrons from the tip of an electron gun. The most commonly used electron gun for high-resolution imaging purposes is the field emission gun (FEG). The FEG emits electrons with the help of an external electric field, which extracts currents through tunneling effect [15]. The emission of the FEG sometimes exploits additional help from the thermal excitation of valence electrons [16] which populate higher energy bands and increase the tunneling current. Therefore, the current in the beam is directly affected by the extraction voltage and the gun temperature, which are the two factors that determine the tunneling probability.

Since the beam is the result of electron selection by the aperture, the aperture size is also a factor that determines the current. Apart from using a smaller aperture to directly reduce the number of passing electrons, the current can also be tuned by adjusting the strength of the condenser lens preceding the probe-forming aperture. When this lens is strongly excited, a cross-over of the beam is created at a longer distance above the condenser lens aperture, resulting in a larger spread of the beam at the aperture plane, and thus more electrons to be cropped out from the wave.

The benefit of a larger current is easily understood, as it leads to a stronger signal and reduces statistical errors such as shot noise. For inelastic scattering signals, a larger current helps to locate and separate characteristic peaks, for example in energy loss spectroscopy or a spectrum formed by characteristic X-rays, and for imaging with elastically scattered electrons, it makes features that only create minimum contrast variation more distinguishable.

On the other hand, operating the TEM with a lower current has its advantages. When the extraction voltage is lowered, the emission of electrons from the gun occurs predominantly from a limited region with significant curvature (the tip), and electrons from spatially incoherent sources are thus reduced. This decrease in the source size strongly increases the coherence length in the condenser aperture plane [17]. Improvement of the beam coherency enhances phase contrast since the phase contrast is the result of interference between coherent partial waves. Current reduction is also one of the main approaches to control beam damage, although finding a balance between a tolerable level of damage and obtaining enough signal for meaningful measurements is not easy or even impossible in some cases.

1.3 Imaging

Parallel beam imaging

When the TEM is operated with a parallel beam (sometimes referred to as the conventional TEM or CTEM), it illuminates a relatively large area at once with a broad plane wave (figure 1.3-a). The plane wave with an original flat wavefront acquires a phase profile corresponding to the local potential due to the interaction with the sample. Since the phase cannot be directly measured, some form of aberration is often deliberately introduced to create the contrast related to the phase profile. This process can be understood by considering the effect of aberration in equation 1.3 and 1.4. In cases where no objective lens aperture is used, the exit wave on the detector plane $\psi_e(\vec{r'})$ is essentially the exit wave on the sample plane convolved with the Fourier transform of the aberration function $\psi_e(\vec{r}) \otimes \mathrm{FT}\left[\exp\left(i\chi\right)\right](\vec{r})$. While $\psi_e(\vec{r})$ contains purely phase variation, its convolution with the aberration function forces a coherent sum among wavelets in a small region and therefore produces phase contrast in the exit wave image $|\psi_e(\vec{r'})|^2$.

The resolution of the formed image is determined by the aberrations of the objective lens, as high spatial frequency features of the sample can be faithfully transferred to the image only if the wave at large scattering angles is not severely distorted by geometrical aberration [18]. Since the transmitted beam also contains inelastically scattered electrons, whose wavelength is different from the elastically scattered and the unscattered electrons, the chromatic aberration, which describes the variation in focusing ability for electrons of different energy, will also limit the resolution of the TEM [19]. Besides aberrations, the highest spatial frequency presented in the image is also determined by the characteristic of the detector in use. To capture the intensity variation of the waveform, a detector comprising an array of pixels is required, and the density of the pixel array hence determines the maximal frequency of the recorded image. Given a fixed number of pixels, the operator of the microscope thus has to make the necessary decision between

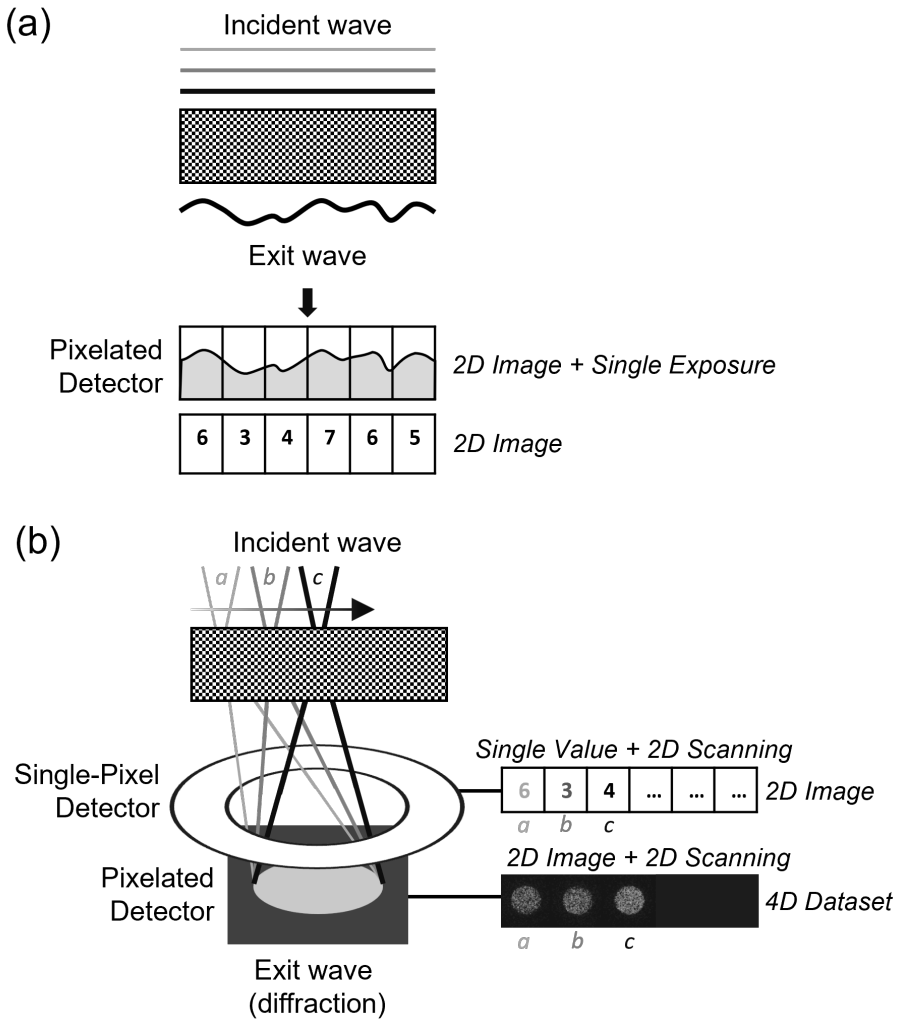


Figure 1.3: Common imaging modes. (a) In CTEM, the intensity of the exit wave is recorded by the detector, which outputs an array of pixel values at once. (b) In STEM, the diffraction pattern of the scanning probe is formed on the detector plane. The recording of the data is done sequentially and synchronized with the scanning of the electron probe, resulting in a 4D dataset if the diffraction pattern is recorded with a pixelated detector or a 2D image if electrons within a specific angular range are counted with a single-pixel detector.

a larger field of view or higher sampling frequency by choosing a specific magnification.

Convergent beam imaging

Convergent beams are often created to illuminate a limited area that can sometimes be approximated as a point. Scanning coils can be employed to perform a raster scan of the beam on the sample plane so that data can be acquired from a large area consisting of a 2D array of probe positions and the name "scanning transmission electron microscopy (STEM)" is thus given to such operation mode. With a specific setting of the projection system of a TEM, the scanning convergent beam creates the convergent beam electron diffraction (CBED) patterns on the detector plane. By recording these 2D diffraction patterns, with e.g. a pixelated detector, from the 2D scanning pattern, one could acquire datasets with 4 dimensions which can then be further processed to learn information about the sample (figure 1.3-b).

However, with technical difficulties, such as response of the pixelated detectors (see discussion in the next section), the most common imaging mode for STEM utilizes a single-pixel detector that outputs a value from each probe position. These values collected from the 2D raster scan effectively render a 2D image, representing the intensity fluctuation as a function of the probe position on the sample plane (figure 1.3-b).

For imaging in STEM mode, the detector plane is usually conjugate with the diffraction plane. From equations 1.2 and 1.3, it can be seen that the recorded diffraction pattern would be an aperture image if the beam propagates in a vacuum. In other words, if a none-zero intensity is picked up by the single-pixel detector placed in a region that is supposedly blocked by the aperture (the dark field, or DF), it indicates the presence of a scatterer, such as an atomic column, at the current probe position. Coupled with the scanning process, the generated 2D image thus captures the position and the relative scattering power of the atoms in the examined material.

This concept serves as the cornerstone of the most common imaging mode in STEM, which is the annular-dark-field (ADF) imaging mode. Depending on the scattering angle that is covered by the detector, this imaging mode further branches into high-angle and low-angle ADF (HAADF and LAADF). HAADF imaging normally employs a detector that covers a wide range on the detector plane at high angles. The large integration area has been shown effective in causing lateral incoherency [20, 21], indicating the intensity acquired at one probe position can be seen simply as the result of local scattering, without the need to consider the contribution from the environment of the sampled probe position. This simplicity encourages many statistical analyses to be made based on HAADF images, as the contrast is tightly linked to atom species [22, 23] and thickness [24]. On the contrary, LAADF utilizes electrons at low scattering angles that reflect local crystal periodicity and strain [25]. Besides, for a thin sample of weak scatterers, such as graphene [11] and other 2D materials [26], LAADF can be more efficient in imaging as the scattering angle resulting from such materials is typically low.

The resolution of ADF imaging is related to the sharpness of the electron probe [27], and the density of the scanned probe positions. However, a dense raster scanning process translates to a long exposure time for forming an image, and thus greater dose usage and potential artifacts induced by sample drift. The operator, therefore, needs to find a

balance between these factors to acquire data under optimal conditions.

Detectors

In the previous section, two types of detectors are mentioned: pixelated detectors which are composed of arrays of detection units, and single-pixel detectors which are essentially large pieces of detecting unit covering a certain area in the detector space (figure 1.1).

Traditionally the pixelated detectors are made of charge-coupled devices (CCD) [28]. This type of detector is composed of a scintillator layer and a MOS layer responsible for charge storage and readout. When the incident electron enters the detector, it deposits its energy in the scintillator layer and generates photons. The photons are then captured in the channel of the MOS, generating electron-hole pairs that can be stored in the charge wells of individual pixels. The signals of the photon-induced charge carriers can be read from the address of the corresponding pixels, and the distribution of the captured electrons is consequently delivered to the user.

Although prevalent in many TEM models, the CCD detectors come with some performance drawbacks. The signals of stored charges need amplification before they are read out, and this process takes place in a serial manner, which can be time-consuming and limits the maximum frame rate of such devices. Furthermore, since the detection of electrons relies on photons of much lower energy, the noise threshold is set low, making the readout susceptible to thermal agitations and electronics noise and hence a low signal-to-noise ratio, especially for low-dose imaging as these noises dominate at low count rates over Poisson noise. Besides, the photons can be detected at a distance from the point of impact of the incident electron on the scintillator layer, leading to a signal distribution spanning several pixels and introducing a blurring effect in the output image.

Pixelated detectors are in theory compatible with the STEM mode, which results in a frame of the diffraction pattern from each visited probe position, rather than a single value. However, the CCD detectors suffer from a long dead time due to the readout procedure, and this delay is especially significant when compared to the scanning rate of the electron beam in the STEM mode.

Single-pixel detectors are traditionally made with a scintillator layer coupled with a photomultiplier tube (PMT) that consequently amplifies the signal. With the charge storage no longer needed and a reduced volume of the output data, the response of such detectors can be much faster and is typically adequate to be synchronized with the scanning speed of a STEM experiment. While the scintillation effect is fast, the decay rate of the PMT, however, causes prolonged signal output and thus holds the response time of such detector to be more than 1μs [29, 30], resulting in noticeable imaging artifact when the scanning is set faster.

The development of semiconductor sensing layers and readout electronics has seen great success over the years. The latest sensing layer eliminates the need for the scintillator and the PMT, as the energy is directly deposited into the semiconductor and generates electron-hole pairs. A bias voltage is applied to this sensing layer to avoid recombination of the charge carriers and to guide the induced holes to the electronics layer for further signal processing. The direct energy deposition creates many more charge car-

Detector Type	Data Format	Data Volume
Single Pixel	electron count	$S_x \times S_y$
Pixelated		
Frame-based	electron distribution	$S_x \times S_y \times D_x \times D_y$
Data-driven	electron position, time of arrival, time over threshold	number of detected electrons

Table 1.1: The data format and data volume produced by each detector type in a STEM experiment. S_x and S_y refer to the size of the scanning, and D_x and D_y to the size of the detector.

riers compared to the amount of photoelectrons generated with the extra scintillation step, allowing the threshold for noise discrimination to be set much higher. This effectively reduces any noise from thermal excitation while maintaining the sensitivity to the high-energy fast electrons.

Since these detectors no longer require an intermediate step of photon generation, they are often called direct electron detectors (DED), and both pixelated and single-pixel detectors benefit greatly from adopting such technology. For the single-pixel detectors, direct electron detection allows faster response and elimination of the artifacts stemming from the PMT [29]. For pixelated detectors, several designs are made based on DED and each shows different advantages.

Two commonly seen series of pixelated DEDs are the Medipix and the Timepix families. A brief introduction to the Medipix and Timepix detectors is provided below, the description of detectors from both series is based on the design of the chips and functionalities used for some of the works in this thesis (Medipix3 [31] and Timepix3 [32]).

Medipix is a pixelated detector that provides a frame-based readout, which means that whenever the readout is triggered, all of the pixels contribute to the output. Each pixel in a Medipix detector owns two counters capable of storing signals of multiple incident electrons. The signal can be converted to digital at the pixel level, and the number of bits for digitization defines the depth or the dynamic range of the detector, which describes the range of electron counts that can be distinguished. The design of the Medipix chip enables several useful features. The charge-summing mode is effective for reducing the excessive counts due to the large diffusion range of the incident electrons within the sensing layer. When operated in this mode, the energies distributed among a cluster of pixels due to a single incident electron are immediately compared, and their contribution would then be allocated to the counter that received the most energy. As a consequence, a single incident electron would be counted only once even though the energy is indeed spread in a range of pixels. Another feature utilizing the counters is the continuous readout. Although the readout of each frame takes time, the two counters can be used in turn, so that one counter can record signals while the other one is going through the readout stage. This feature allows Medipix to be operated without dead time and a maximum time resolution equal to the readout time of a full frame. In conclusion, Medipix is a very powerful detector that provides a maximum dynamic range of 2^{24} (when both counters are used), and a much faster frame rate (depending on

the depth of the pixel, maximum ~14 kHz at 1-bit mode), compared to the CCD detectors. However, for low-dose imaging in STEM mode where the frames are mostly populated with zeros, the frame-based readout is very inefficient in terms of the processing time and the required amount of data storage. Besides, the frame rate of the Medipix detector is around 10 to 1000 times slower than the common dwell time for STEM, making the tolerable level of sample drift drastically lowered. It forces the beam current to be set exceptionally low in order to compensate for the damage and contaminations that are induced due to the long dwell time.

Timepix is a pixelated detector providing both frame-based and data-driven readout. Data-driven readout refers to the operation mode where the readout is triggered by pixels with a signal exceeding the threshold. This significantly reduces the average readout time if the "frame" consists of mostly inactive pixels, and thus could be an effective solution to the aforementioned problem. Another important feature of the Timepix camera is its time resolution. When the signal of an incident electron exceeds the threshold, it triggers the readout of a data pack that contains time-related information, from which the arrival time of the electron can be calculated. When synchronized with the scanning engine, this time of arrival can be used to retrieve the corresponding probe position from which the electron scattering pattern is recorded, and therefore "frames" of CBED patterns can be reconstructed if needed. Despite the fact that the volume of the data can be significantly reduced under low-dose working conditions, it is calculated that when more than 50 % of the pixels are activated within the readout time of a single pixel, the data-driven mode is slower than frame-based. Besides, the electronics behind the sensing layer are designed to measure the "time over threshold", which is the duration of the electron-induced signal, and therefore the dead time before the pixel can be used to count the next incident electron is not negligible. Both factors limit the current of the electron beam, or the dose rate, making the Timepix detector not yet able to completely replace the functions of annular detectors.

Although pixelated DED such as Medipix and Timepix show several advantages over traditional CCD detectors, such as low noise, faster readout, and larger dynamic range, there are also drawbacks making the CCD detectors not yet fully replaceable for TEM experiments. For example, to protect the electronics underneath the sensing layer, the sensing layer is made rather thick to prevent penetration. This causes side effects such as large scattering distribution, and as a consequence, the pixels are made large (55 μm for both Medipix3 and Timepix3) so that the energy deposited by the fast electron can be limited within only a few pixels. The large pixel size allows fewer pixels to be fitted into the chip. Both Timepix and Medipix chips are composed of 256×256 pixels, or 512×512 if tiled, which is still much fewer than most of the CCD detectors that possess a pixel number ranging from 1024×1024 to beyond 4096×4096, making them more effective and efficient for recording high-resolution images from a large illumination field [33, 34].

Chapter 2

Methods

In the previous chapter, a few basics of TEM were discussed, for example, the forming of an electron probe and the imaging mechanisms under various TEM operation modes. Upon these basics, methods that assist microscopists in designing experiments, analyzing data, or even improving the operation of the microscopes, are developed. Several of these methods that inspire and complement the work of this thesis will be introduced in this chapter, including concepts and procedures to simulate TEM experiment results, advanced imaging modes utilizing phase contrast, and ways of analyzing results with statistical approaches.

2.1 Simulating TEM Experiments

Simulation plays an important role in the methodology development of TEM. It can create an environment mimicking the experiment, including arbitrary choice of sample type and geometry, as well as the imaging conditions. Besides, the ability to generate synthetic data according to a wide range of possible experimental conditions makes simulation a crucial part in, for example, evaluating imaging methods [35, 36, 37], validating hypothesis built upon observation [38], and understanding effects of various types of imperfections in the optical system [39, 40]. While simulations can be convenient, one should remain cautious to make sure that the model adequately represents all factors that are present in an experiment.

For the work in this thesis, the simulation package MULTEM [41, 42] is used. MULTEM is a TEM simulation package that allows the user to freely program microscope parameters, incident waveform, and sample configuration. The creation of a simulated electron beam, as well as its propagation and scattering, follow closely the equations and concepts discussed in the previous chapter. In addition, one important approximation, usually referred to as multislice, is utilized to simplify the calculation of the combined process of scattering and wave propagation within the sample. This approximation states that the interaction of a fast electron with a thin slice of the sample can be simplified with the assumption that the electron first acquires some phase resulting from its interaction with the projected potential of the slice, followed by unperturbed propagation through the sample until the next slice is reached. Such approximation only holds when the exact z positions of the atoms within the slice make negligible differences, as the potential profile

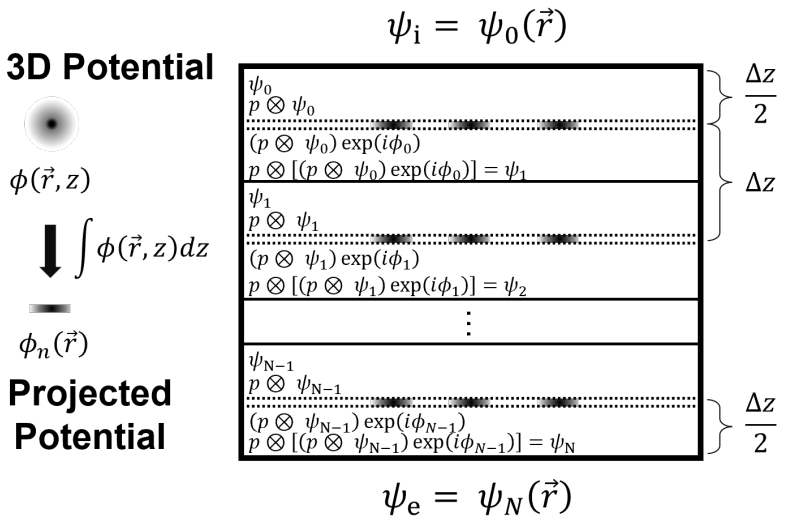


Figure 2.1: Computing the transmitted waveform using multislice simulation. The projected potential of each slice is first calculated by integrating the potential in the direction of wave propagation. The waveform evolution as a result of scattering and propagation can be calculated by computing the cascaded effect of all the slices.

of the atoms is squeezed into a flat slice. Making a thick sample into a single slice would eliminate any information in the z direction in the transmitted wave, such as high-order Laue zone diffraction and thus results in a wrong outcome of the computed transmitted waveform. In that case, a thick sample can be decomposed into multiple thin slices that fulfill the approximation, and the transmitted wave can thus be computed by calculating the effect of each slice, one slice at a time, hence the name "multislice".

This approximation can be formulated mathematically. From equation 1.1 the corresponding phase shift $\phi_{z_i}(\vec{r})$ induced by a slice can be written as:

$$\phi_{z_i}(\vec{r}) = \frac{\pi}{\lambda E} \int_{z_i}^{z_{i+1}} V(\vec{r}, z)\, dz \tag{2.1}$$

where z_i describes the depth of the entrance surface of the slice i. The evolution of the waveform in a short distance can be described by Fresnel diffraction or near-field diffraction [43]:

$$\begin{aligned}
\psi(\vec{r}, \Delta z) &= \left[p(\vec{r}, \Delta z) \otimes \psi(\vec{r}, z)\right] \exp\left(2\pi \Delta z/\lambda\right) \\
&= \mathrm{IFT}\left\{\mathrm{FT}\left[\psi(\vec{r}, z)\right] \tilde{p}(\vec{k}, \Delta z)\right\} \\
p(\vec{r}, \Delta z) &= \frac{1}{i\lambda \Delta z} \exp\left(\frac{i\pi}{\lambda \Delta z}|r|^2\right) \\
\tilde{p}(\vec{k}, \Delta z) &= \exp\left(-i\pi\lambda k^2 \Delta z\right).
\end{aligned} \tag{2.2}$$

Equation 2.2 describes the effect of propagation of a distance Δz on the beam as a convolution between the wave and a Fresnel propagator $p(\vec{r}, \Delta z)$, or a multiplication with the Fourier transform of the propagator $\tilde{p}(\vec{k}, \Delta z)$ in the reciprocal space. In combination with the phase shift created by a slice of the sample and by ignoring the homogeneous phase shift created by propagation, the waveform at the exit plane of that slice is found as:

$$\begin{aligned} \psi_{z_{n+1}}(\vec{r}) &= p(\vec{r}, \Delta z) \otimes \left[\psi(\vec{r}, z) \exp\left(i\phi_n(\vec{r})\right)\right] \\ &= \text{IFT}\left\{\text{FT}\left[\exp\left(i\phi_n(\vec{r})\right) \psi_{z_n}(\vec{r})\right] \tilde{p}(\vec{k}, \Delta z)\right\} \end{aligned} \tag{2.3}$$

The aforementioned approximation is explicitly expressed in equation 2.3, as the incident wave first acquires the phase induced by the projected potential of the slice and then propagates by a distance corresponding to the thickness of the slice. The precision of this solution can be improved by assuming the projected potential of each slice can be found as a thin slab at the center, which would require extra half-slice propagators at the top and bottom surfaces of the simulated sample [43, 44]. The evolution of the waveform after each slice with the improved precision can be shown as:

$$\psi_{z_{n+1}}(\vec{r}) = p(\vec{r}, \Delta z/2) \otimes \left[p(\vec{r}, \Delta z/2) \otimes \psi(\vec{r}, z) \exp\left(i\phi_n(\vec{r})\right)\right] \tag{2.4}$$

When a thicker sample composed of multiple slices is considered, the transmitted wave can be found simply by cascading the effect of each slice, as shown in figure 2.1.

2.2 Advanced Imaging Methods

In section 1.3 a few common imaging methods for the CTEM and STEM modes were introduced. These methods are undeniably valuable and have supported innumerable studies of materials with their high-resolution imaging ability, but as the demand for dose-limited imaging surges due to the interest in biomaterials and other beam-sensitive substances, great efforts have been invested in the development of novel imaging methods for both operation modes. In this section, a few of these advanced methods to achieve higher dose efficiency and better utilization of phase information of the transmitted wave are discussed. These methods exploit specific algorithms and hardware to obtain or enhance contrast that stems from elastic scattering.

Advanced imaging in CTEM

As discussed in section 1.3, a parallel beam transmits through the sample and acquires a phase profile with information about the object, which can result in detectable contrast variation with the help of aberrations. However, inducing aberration is not the only way to force phase contrast in the recorded intensity. Other approaches that lead to

interference between the scattered and the unscattered electrons can also produce the desired contrast.

Consider a case of an extremely thin sample that fulfills the phase object approximation (POA), which states that the sample can be represented by the phase profile created by its projected potential and the propagation within the sample can be ignored, the exit wave is thus found to be:

$$\psi_e(\vec{r}) = \psi_i(\vec{r}) \exp\left(i\phi(\vec{r})\right) \tag{2.5}$$

where $\psi_i(\vec{r})$ and $\psi_e(\vec{r})$ stand for the incident and the exit waves, respectively, and the phase profile related to the sample projected potential is written as $\phi(\vec{r})$, which is acquired by the electron wave in the process of transmission. Assuming the sample is composed of weak scatterers so that the induced phase shift is also low (weak phase object approximation, WPOA), the exponential can thus be approximated by only its zeroth and first-order term:

$$\psi_e(\vec{r}) = \psi_i(\vec{r}) \left(1 + i\phi(\vec{r})\right). \tag{2.6}$$

Equation 2.6 states that the exit wave $\psi_e(\vec{r})$ can be separated into two parts: one containing the unscattered electrons, thus being the same as the incident wave $\psi_i(\vec{r})$, and at the same time, the other has information related to the scattering (in this case, phase shift) experienced by the electrons interacting with the phase object. The second part, written as $i\psi_i(\vec{r})\phi(\vec{r})$, is purely imaginary, or has a pure phase of $\frac{\pi}{2}$.

One way to enhance the contrast would be applying a potential corresponding to a $-\frac{\pi}{2}$ phase shift to the unscattered electrons, so that a total phase difference of π can be created between the two parts of the beam, enabling stronger contrast variation as a result of destructive interference. To achieve that, such potential can be created in a confined area containing only the central diffraction spot on the diffraction plane. Since the scattered electrons acquire a distribution of scattering angle and are no longer strictly parallel to the optical axis, they are spread on the diffraction plane. The unscattered paraxial beam, on the other hand, is focused into a single spot at the center of the diffraction plane that coincides with where the potential is applied and thus acquires the desired phase shift (figure 2.2). Such device that introduces a designed phase profile to the passing wave is referred to as a phase plate, and this scheme to create the condition for destructive interference between the scattered and the unscattered partial waves describes the concept behind the Boersch phase plate [45] and a Zernike phase plate [46], which creates an extra phase shift of $\frac{\pi}{2}$ to the scattered electrons instead. These phase plates are frequently used for imaging biological samples which mostly consist of light elements and thus suffice the WPOA.

Advanced imaging in STEM

The imaging methods discussed in section 1.3, such as HAADF and LAADF, output a single value for each probe position, which has a one-to-one correspondence with the intensity registered in each pixel of the resulting 2D image. These rather traditional

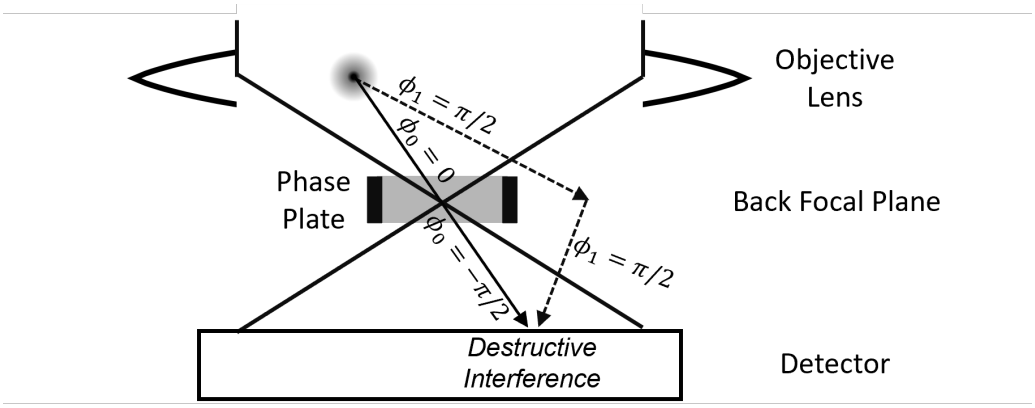


Figure 2.2: A phase plate that introduces a phase shift of $-\frac{\pi}{2}$ to the unscattered wave (solid lines) has the effect of creating destructive interference with the scattered electrons (dashed lines), thus enhancing the phase contrast in the final image.

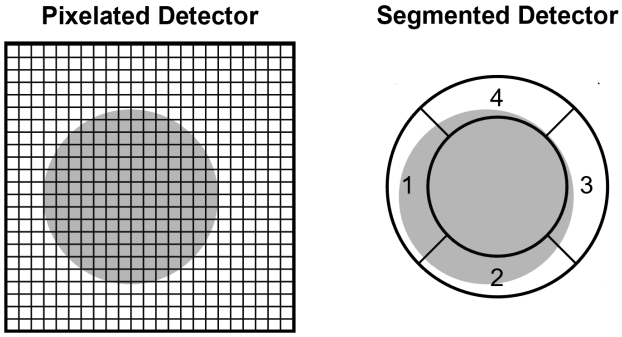
STEM imaging methods do not utilize all the information from a STEM experiment, and hence offer a suboptimal dose usage. First of all, the annular detectors do not utilize all the electrons transmitted through the sample, thus throwing away valuable information. Secondly, because the detector integrates the signal from a large area, the transmitted waveform as a function of scattering angle cannot be found and used. Finally, the one-to-one correspondence between the pixel intensity and the probe position means that the value of each pixel in the image is not calculated by taking into account information and contribution from data collected elsewhere. In this section, several advanced STEM imaging methods that utilize these three aspects of information are introduced, and their unique advantage will be discussed.

A rather simple approach is demonstrated by the center-of-mass (COM) related imaging methods. These are a series of methods that are based on the shift of intensity distribution of the CBEDs, which can either be calculated from the results of a pixelated detector or a segmented detector [47, 48], which produces data often referred to as the differential phase contrast (DPC) signal (figure 2.3). Both the COM and DPC signals reflect, to some extent, changes in the waveform, and also these methods could utilize a larger percentage of the transmitted electrons (if the signal is computed based on the readout of a pixelated detector), compared to ADF.

The COM/DPC methods produce a 2D vector at each probe position, corresponding to the calculated intensity shift. In the case of COM, this shift can be calculated as the first moment of the CBED intensity distribution [49, 50], and is shown related to the local in-plane electric field ($\vec{E}_\perp$):

$$\vec{I}^{\mathrm{COM}}(\vec{r}_p) = \int \vec{k}\mathrm{I}(\vec{k})\,dk \propto \vec{E}_\perp \tag{2.7}$$

where the vector $\vec{k}$ describes a 2D diffraction plane, and $\mathrm{I}(\vec{k})$ is the normalized intensity



$$\mathrm{COM}_x = \frac{\sum_{x=1}^{X}\sum_{y=1}^{Y} xI(x,y)}{\sum_{x=1}^{X}\sum_{y=1}^{Y} I(x,y)} \qquad \mathrm{DPC}_x = \frac{I_3 - I_1}{\sum_{i=1}^{4} I_i}$$

$$\mathrm{COM}_y = \frac{\sum_{x=1}^{X}\sum_{y=1}^{Y} yI(x,y)}{\sum_{x=1}^{X}\sum_{y=1}^{Y} I(x,y)} \qquad \mathrm{DPC}_y = \frac{I_4 - I_2}{\sum_{i=1}^{4} I_i}$$

Figure 2.3: Computation of the COM and DPC signal. The gray disk represents the CBED pattern on the detector plane. The COM of the CBED is found by taking the first moment of the intensity distribution recorded by a pixelated detector. The DPC signal from a segmented detector is calculated by taking the difference between intensity recorded by opposing segments.

measured thereon. With the POA, the COM shift $\vec{I}^{\mathrm{COM}}(\vec{r}_p)$ is shown to be a function of the projected potential of the sample and the probe function [51]:

$$\begin{aligned}\vec{I}^{\mathrm{COM}}(\vec{r}_p) &= \frac{1}{2}|\psi_i(\vec{r},\vec{r}_p)|^2 \star \nabla\phi(\vec{r}) \\ &= \frac{1}{2}\nabla(|\psi_i(\vec{r},\vec{r}_p)|^2 \star \phi(\vec{r}))\end{aligned} \tag{2.8}$$

where $\psi_i(\vec{r},\vec{r}_p)$ is the probe function when situated at probe position $\vec{r}_p$, $\phi(\vec{r})$ is the amount of phase shift corresponding to the projected potential of the sample, and ∇ is the gradient operator. The vector array can be further integrated to produce the integrated COM (iCOM) image (and iDPC with the same integration procedure), essentially removing the gradient operator and producing a result directly linked to the object function sampled by the probe:

$$\begin{aligned}I^{\mathrm{iCOM}}(\vec{r}_p) &= \frac{1}{2}\left(|\psi_i(\vec{r},\vec{r}_p)|^2 \star \phi(\vec{r})\right). \\ &= \mathrm{IFT}\left\{-i\frac{\vec{k}_p}{|k|^2}\mathrm{FT}\left[\vec{I}^{\mathrm{COM}}\right](\vec{k}_p)\right\}(\vec{r}_p)\end{aligned} \tag{2.9}$$

In equation 2.9, the integration is performed in the Fourier space, where the Fourier

transform of the COM shift vector array is multiplied with the factor $1/i\vec{k}_p$. The one-to-one correspondence is lifted due to the integration, as it takes all probe positions into account. With these properties, namely, the utilization of the intensity distribution of the wave, minimizing the transmitted electron loss, and the sharing of information between probe positions during the reconstruction of the object, iCOM and iDPC are shown to be very effective for low-dose imaging [52] and imaging of weak scatterers [53]. Their relatively low computational demand, compared to the other advanced imaging methods, also makes them ideal for a rapid inspection of the quality of the acquired 4D datasets.

Despite being computationally efficient, there are some drawbacks to the method. Formulated by Rose [54], the phase contrast transfer function of the differential signal peaks at the spatial frequency of half of the convergence angle, and signals at lower / higher frequencies are strongly suppressed. Such characteristic makes both COM and DPC imaging methods only sensitive to features of similar size of the probe size and not suitable for imaging features of large scale, or anything beyond the diffraction limit defined by the convergence angle. The attenuation at low frequency can be mitigated to some extent with the integration process, as the multiplication with the $1/\vec{k}$ term in equation 2.9 amplifies low-frequency signal. However, the ultimate loss at the high frequency beyond the diffraction limit is not retrievable.

Besides iCOM, many more STEM imaging methods utilize the waveform recorded from each probe position, for example, the ptychography methods [55]. These methods retrieve the exit wave and then use it to reconstruct an image or even a 3D object of the examined sample [56, 57]. The richer information contained in the waveform gives ptychography methods the potential to achieve resolution beyond the diffraction limit and possibly even higher dose efficiency, compared to iCOM/iDPC.

Many ptychography methods are derived from the same exit wave phase retrieval approach, which is the Gerchberg-Saxton (GS) algorithm [58]. The GS algorithm is a simple and powerful tool to solve for the missing phase when only the amplitude of the wave can be recorded. It runs on the condition that an image of the wave on both planes, i.e. $I(\vec{k})$ and $I(\vec{r})$, can be recorded as the squared amplitude of the wave. The process of the GS algorithm is described by the following pseudocode:

Algorithm 1 Gerchberg-Saxton Algorithm

$\tilde{\psi}(\vec{k}) := \sqrt{I(\vec{k})}$

$\psi(\vec{r}) := \sqrt{I(\vec{r})}$

for $n := 0$ to N_{ite} **do**

$$\psi_{n+1}(\vec{r}) := \left|\psi_n(\vec{r})\right| \frac{\mathrm{IFT}\left[\tilde{\psi}_n(\vec{k})\right]}{\left|\mathrm{IFT}\left[\tilde{\psi}_n(\vec{k})\right]\right|}$$

$$\tilde{\psi}_{n+1}(\vec{k}) := \left|\tilde{\psi}_{n+1}(\vec{k})\right| \frac{\mathrm{FT}\left[\psi_{n+1}(\vec{r})\right]}{\left|\mathrm{FT}\left[\psi_{n+1}(\vec{r})\right]\right|}$$

end for

At the beginning of the process, only the amplitude of the wave for both planes is known, which is represented by $\left|\tilde{\psi}(\vec{k})\right|$ and $\left|\psi(\vec{r})\right|$. The algorithm starts with an initial guess that the wave is completely real on the diffraction plane, i.e. $\tilde{\psi}(\vec{k}) = \left|\psi(\vec{k})\right|$. The wave on the sample plane is found to be the inverse Fourier transform of the assumed wave on the diffraction plane, according to Fraunhofer diffraction (equation 1.4). Since the phase term of $\psi(\vec{k})$ is missing, the calculated waveform $\psi(\vec{r})$ is most likely inconsistent with the observed amplitude of the wave. This is then resolved by swapping the amplitude part of the newly constructed waveform with the one recorded from the experiment, while the phase part is kept as it currently is. By going back and forth between the r-space and k-space, repeating such projection of the waveform between the two spaces one can gradually retrieve the right phase of the wave that was lost during the data acquisition.

The GS algorithm can be applied in many optical experiments, including STEM. However, simultaneous recording of the wave in both real and reciprocal spaces is not possible for a TEM. While the wave amplitude in the Fourier space $\left|\psi(\vec{k})\right|$ can be taken as the square root of the recorded CBED intensity, some assumptions are usually made to fulfill the missing wave image in the real space. For example, the POA (equation 2.5) is often adapted so that the transmitted wave amplitude on the sample plane can be assumed the same as the incident wave, i.e. $\left|\psi_i(\vec{r})\right| = \left|\psi_e(\vec{r})\right|$, and that the incident waveform can be constructed with the information of the aperture size and the measured aberration function from the aberration correction process (note that the exit wave $\psi_e(\vec{r})$ in the picture of TEM optics is the real-space wave $\psi(\vec{r})$ in the GS algorithm).

In addition to the missing information, other factors can hinder the convergence of the algorithm and even result in inaccurate solutions. One of the factors that influences profoundly the performance of GS is the inversion symmetry of the illuminating wave. This problem stems from the fact that the Fourier transform of a wave $\psi(\vec{r})$ and its inverted complex conjugate $\psi^*(-\vec{r})$ has exactly the same amplitude. In other words, if only the intensity of the wave is recorded, none of the two possible solutions can be reasonably rejected. For an exit wave with the POA $\psi_e(\vec{r}) = \psi_i(\vec{r}) \exp\left[i\phi(\vec{r})\right]$, by its recorded intensity one cannot distinguish whether the interacting object should be $\phi(\vec{r})$ or $-\phi(-\vec{r})$ if the incident wave $\psi_i(\vec{r})$ is central symmetric. This incapability originates from the nature of the Fourier transform or the creation of a diffraction pattern and is independent of the choice of the phase retrieval method. For the GS algorithm, it has been shown to cause the twin-image problem in the retrieved phase of the wave [59]. Therefore, to successfully reconstruct an image of the object, it is necessary to break the symmetry, or to reduce its effect, so that the phase retrieval algorithm can be applied to the STEM data.

Luckily for STEM, this detrimental effect from the symmetry can be naturally alleviated by the translation of the probe and the overlap between the probe intensities from different scan points [60]. The overlap of the probe functions leads to repeated sampling of the same region of the object. This can be used to solve the twin problem since only one of the twins in the retrieved waveform can be illuminated twice at the same position by the two probes that lie some distance from each other, as shown in figure 2.4.

As the twin problem is alleviated and the transmitted waves are accurately retrieved, the object can be reconstructed with the knowledge of the incident waveform and the

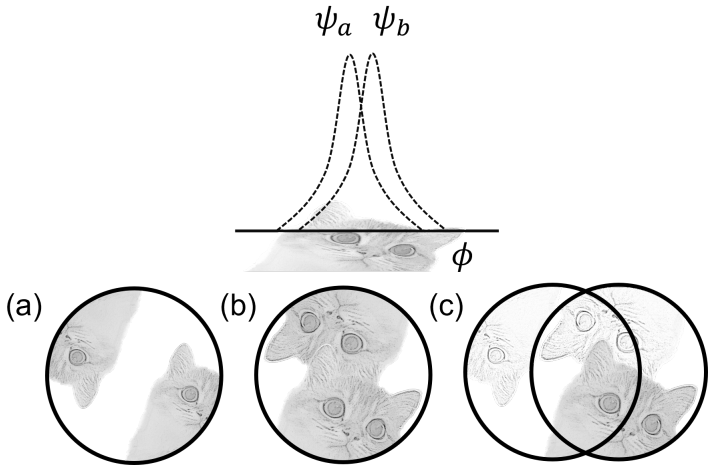


Figure 2.4: Solving the twin problem with overlapping probe functions. (a,b) show the retrieved object with probe functions ψ_a and ψ_b individually. Since the probe functions possess central symmetry, and thus their individually retrieved object phase suffers from the twin problem. (c) shows the solution when both probe functions are considered, and the false twin is suppressed, as only one of the twins is supported by both illuminations.

retrieved waveform from each probe position in combination with the POA.

The retrieval process of the wave and the reconstruction procedure may be further refined to reduce the effect of possible errors in the data acquisition and wrong assumptions for the algorithm. For example, the estimation of the incident wave can be corrected during the iterative process [61], partial coherency of the electron wave can be addressed by treating the probe as a mixed state of waves [62], mis-tilt of the sample can be addressed by using a tilted propagator [63], and the mispositioning of the probe can be fixed by simultaneous retrieval of the right position [64].

There are also ptychography methods that do not rely on the GS or any kind of iterative algorithm. An example can be found in the single sideband (SSB) reconstruction [27]. By applying a 2D Fourier transform to the 4D-STEM dataset with respect to the sample plane, the result is found to be connected to the frequency components of the imaged object with the WPOA:

$$\begin{aligned} G(\vec{k},\vec{k}_p) &= \mathrm{FT}_{\vec{r}_p}\left[I(\vec{k},\vec{r}_p)\right] \\ &= |\psi_i(\vec{k})|^2\delta(\vec{k}_p) + \psi_i(\vec{k})\psi_i^\star(\vec{k}+\vec{k}_p)\exp(-i\phi(-\vec{k}_p)) \\ &+ \psi_i^\star(\vec{k})\psi_i(\vec{k}-\vec{k}_p)\exp(i\phi(-\vec{k}_p)) \end{aligned} \tag{2.10}$$

where $I(\vec{k},\vec{r}_p)$ is the electron count at pixel position $\vec{k}$ while the probe is sitting at $\vec{r}_p$, $G(\vec{k},\vec{k}_p)$ is the mentioned Fourier transform of the dataset, where $\vec{k}_p$ represents the sampled frequency as a result of the scanning process. In equation 2.10, the imaged weak phase object $\phi(\vec{k}_p)$ is a function of the sampling, so the step size for the scanning

needs to be carefully chosen in order to achieve the desired resolution while maintaining the dose usage. Additionally, the resolution of the reconstructed image is further limited by the convergence angle, as no overlap between function $\psi_i(\vec{k})$ and $\psi_i^\star(\vec{k}+\vec{k}_p)$ can be found when $\vec{k}_p$ is larger than twice the convergence angle. It is also worth noticing that each frequency component is sampled with a different weighting. As pointed out by [65], the frequency transfer function of SSB peaks approximately at half of the maximum $\vec{k}_p$. Such behavior brings pros and cons to the imaging method. The transfer function depends on frequency instead of only on the projected potential, implying that the light element can result in strong contrast as well. This alleviates the problem of light elements being hidden by the much stronger contrast created by heavier elements. On the other hand, the ability to distinguish element type, local thickness, and density is reduced due to the weakened correlation between the projected potential and the contrast.

2.3 Statistical Model and Hypotheses Test

Following the discussion in section 1.3, the HAADF contrast exhibits a strong correlation with sample thickness and atom species, making it a suitable imaging method for characterizing the morphology and composition of the sample. When the image is taken under a zone axis (atoms are aligned in the direction of the optical axis), the individual atomic columns can be resolved and their positions and contrast can be statistically quantified. For example, the location of the atomic columns provides information on local strain [66], and their contrast provides information on the relative density [21].

Besides analyzing acquired data, these statistical models can be combined with simulations to provide insight into how to optimize a STEM experiment. In this section, a few related statistical methods based on the HAADF images are introduced. The assumptions and prerequisites necessary to build the relevant statistical models are emphasized in the section, highlighting the condition under which these methods can be applied.

Statistical model based analysis

To analyze the investigated material with TEM experiment results, one must first find a connection between the property of the imaged material and the acquired data. In the special cases where the sample is crystalline and oriented in one of the zone axes so that the atoms are aligned in columns, the total change in the electron wave can be seen as a combined contribution of the scattering with each atomic column. This "scattering property" of the atomic column can be represented by the percentage of electrons scattered to the high angle due to the specific atomic column [67]. Therefore, the HAADF image of the column can be analyzed to extract interesting information about individual columns, as well as the sample composed of them.

The intensity of a pixel in a HAADF image is proportional to the number of electrons received by the detector while the probe is situated at the corresponding position. By normalizing the electron count with the dose per probe position, the percentage of the electrons that are scattered to the angle of the detector can be found, and this percentage thus characterizes the local scattering power in the sample. To formulate a connection

between the columns and the normalized HAADF intensity, the variation in the intensity due to the distance between the electron probe and the atomic column needs to be properly removed. The remaining factors that determine the scattering would then be the atomic column and the probe function. Since the probe function can be assumed invariant during the acquisition of a single image, the effect of the probe does not necessarily needs to be removed before the data can be analyzed to extract information purely associated with the columns.

One of the ways to remove the effect of the distance between the probe and the column is by only using the intensity recorded at the center of the columns, also referred to as the peak intensity. However, this approach brings several disadvantages. First of all, the scanned position is very likely to miss the center of the column, and in order to have a consistent sampling of the center for every column in the image, the step size of the scanning must be set very small. Secondly, using only the values at the center of the column means that the information acquired from the experiment is largely wasted. With limitations imposed by damage and sample drift, a sufficient dose to achieve an acceptable signal-to-noise ratio for accurate interpretation of the data is often not possible.

E et al. proposed to use the sum of the intensity of a cluster of pixels to represent each column [68] instead of peak intensity. Besides the higher data usage rate, this approach also shows several advantages. Demonstrated with a simplified model describing HAADF imaging, it is shown that the sum from the cluster is less sensitive to defocus and source size broadening, while the sensitivity to column thickness is preserved. The idea is then further explored by De Backer et al.. The Voronoi cell, which defines an area around an atomic column, formed by the perpendicular bisectors of the direct connections to the neighboring columns, is used to determine to which atomic column a certain pixel value should contribute. The sum of pixel intensity from a Voronoi cell is named the scattering cross-section (SCS) of the corresponding column.

The SCS can also be calculated as the volume of the fitted Gaussian based on the intensity cluster around the corresponding atomic column. [69, 70]. This approach takes into account the joint contribution from multiple atom columns to the pixel intensity by allowing overlap between the fitted Gaussians, and a homogeneous background can also be fitted in the process.

With these approaches, the scattering ability of an atomic column is simplified to a single value, and the value of each column can be analyzed together statistically. For example, the thickness of the atomic columns can be learned with such information [38]. When the considered atomic columns are all of the same composition, with the assumption that the column's height and environment do not affect the HAADF intensity, the remaining variant that affects the SCS of a column is its thickness, if the background is homogeneous or vacuum. Since the thickness of an atomic column is a discrete function (an integer number of atoms), so should be the corresponding SCS. By the number of different SCSs found in an image, ideally one should be able to determine the number of different column thicknesses in the acquired image.

In reality, however, the discrete nature of SCSs from experimental data is implicit due to shot noise and noise produced by the ignored factors: the height and environment differences of each column, uneven substrate, and inhomogeneous contaminant distribution. Shot noise, also known as Poisson noise, is the only remaining noise of a perfect system. In a case where some events may be measured after a trial, the distribution of the

measured count follows a binomial distribution which is a function of the probability of successful measurement and the number of trials. When this probability is significantly low, the distribution can be simplified as the Poisson distribution:

$$P(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{2.11}$$

where k is the count of the event, and λ is the expected count which is equal to the probability of acquiring a signal multiplied by the number of trials. Assuming the effect of the ignored factors, in combination with Poisson noise, is unbiased and follows a Gaussian distribution, the distribution of the SCSs from all the considered columns can be seen as the combined result of Gaussian components representing each possible thickness (figure 2.5).

$$\begin{aligned} f(x) &= \sum_i C_i(x) \\ &= \sum_i \frac{w_i}{\sigma_i \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right] \end{aligned} \tag{2.12}$$

where $f(x)$ is the expected amount of times to acquire an SCS at the value x, C_i is the i^{th} Gaussian component, and w_i, σ_i, μ_i are the weight, standard deviation, and the mean of the i^{th} components. The number of components can then be found by fitting the distribution with an increasing amount of components. The amount of the components resulting in the highest "goodness" of fitting (not likelihood) hence determines the number of different thicknesses that can be expected from the HAADF image.

The goodness of the fitting result is determined by the likelihood that the observation can be explained by the model and a punishment term to avoid over-fitting. A suitable design of the punishment and its balance with the likelihood changes from case to case [71, 72, 73], and for the discussed case here it was found that the integrated classification likelihood (ICL) criterion is a good indicator of the right component number [38]. The ICL contains the likelihood, the entropy of the model composition, and a term $d \log N$, which is the product of the degree of freedom d and the sample size N [74]. When the ICL calculated at different amounts of Gaussians is plotted, the local minimum in the curve indicates the right amount of components. Once the component number is found, as well as the shape and position of their distribution, the corresponding thickness is then assigned to each component (see example in figure 2.6).

Assuming that the column thicknesses found in the image are continuous integers, the components are assigned a thickness according to the order of their expected SCSs, starting from the lowest expected thickness and with an increment of 1. With that, the thickness of an atomic column with an SCS x can be determined by the component of the highest likelihood, which can be expressed as:

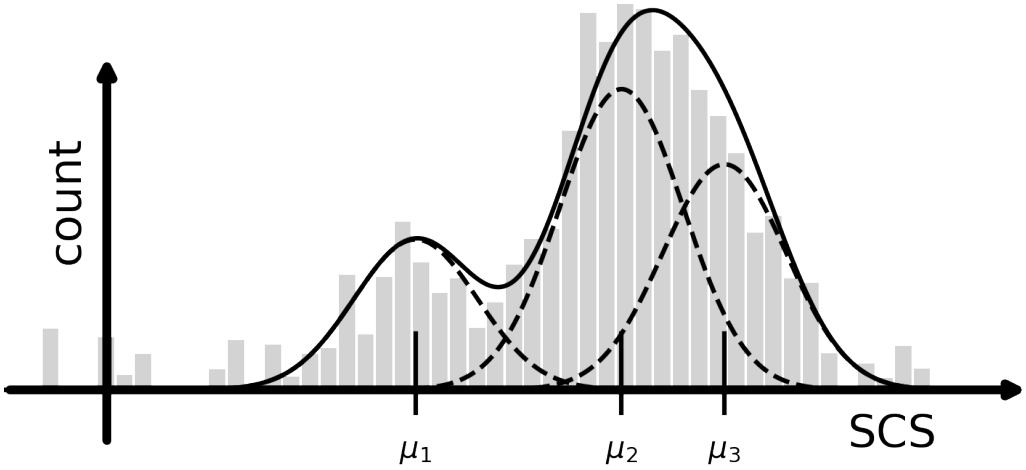


Figure 2.5: Scattering cross-sections and distribution of the components. The measured SCSs from the HAADF image are shown in a histogram. Three components corresponding to column thicknesses of 1, 2, and 3 are responsible for the measured distribution. The distributions of individual components are drawn in dashed curves, with their mean indicated by the black vertical lines, and the combined distribution in the solid curve.

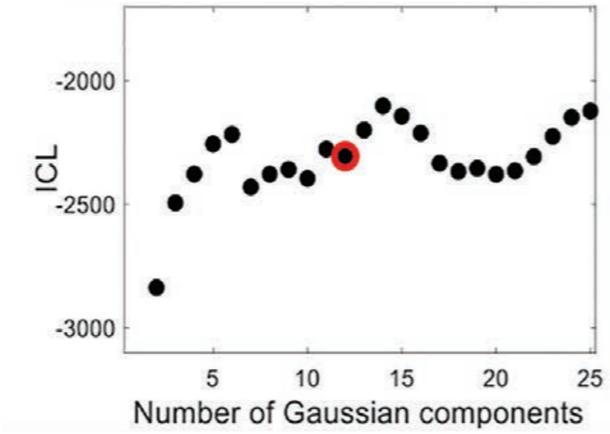


Figure 2.6: Example of an ICL curve calculated from experimental data. The selected local minimum is highlighted in red. It determines the number of different atomic column thicknesses captured in the HAADF image. Image taken from [75].

$$\begin{aligned} &\text{Component Index } i \\ &= \arg\max_i C_i(x) \\ &= \arg\max_i \frac{w_i}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right] \end{aligned} \tag{2.13}$$

As a consequence of this statistical method, a semi-3D sample model can be constructed with the information acquired from a 2D raster scanning. From this model, the thickness of the atomic columns can be determined but not their position in the Z direction, which requires extra analysis. For instance, molecular dynamic calculations [76] can be employed to determine the most stable structure as a function of the height of each atomic column, while the column thickness is fixed to the calculated value based on the experimental result and statistical analysis.

Hypothesis test

Consider the result of an experiment as an observation of the system on which the experiment is conducted, based on the observation and the prior knowledge of the system, one could further deduce possible configurations of the system that may lead to the observation and form a set of hypotheses. The likelihood of a hypothesis being true can be computed if the probability of acquiring the observation under a specific system configuration described by the hypothesis can be found. A hypothesis test can then be performed to compare the likelihood of each hypothesis, and the one found with the highest likelihood is then deemed the one that is most likely to describe the system accurately.

For example, consider an experiment that produces a single value result k, and that a set of possible hypotheses $\mathbb{H}$ exists, the selection of the most likely accurate hypothesis can be formulated as:

$$\begin{aligned} &\mathrm{H}_i \text{ is true if } \; P(k \mid \mathrm{H}_i) > P(X = k \mid \mathrm{H}_j), \\ &\text{with } \mathrm{H}_j \in \mathbb{H} \text{ but } j \neq i \end{aligned} \tag{2.14}$$

where $P(k \mid \mathrm{H}_i)$ is the probability to acquire k given the condition described by hypothesis H_i, which is the likelihood of H_i being true. Once the likelihood of each hypothesis is computed, the most likely accurate hypothesis is then determined.

From this simple formulation of a hypothesis test, the criteria for applying such a test are also implied. First of all, the causality between the hypotheses and the observation should be clear, and the probability $P(X = k \mid \mathrm{H}_i)$, or the likelihood of a hypothesis, needs to be computable. Secondly, the number of hypotheses with a significant probability to be true must be limited. In other words, if the parameter space is too wide or unbounded, it would be impossible to compute all the likelihood for each hypothesis, and the test would thus fail.

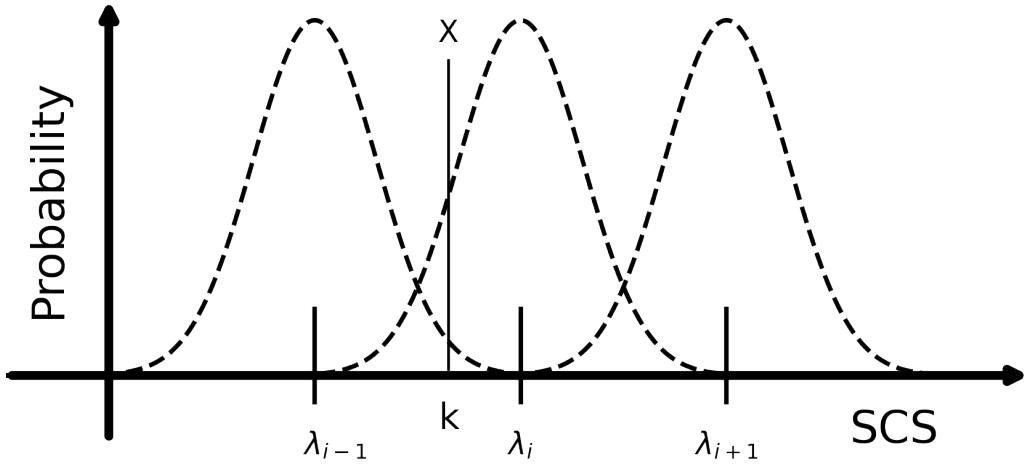


Figure 2.7: Selecting process of a hypothesis test. The observation results in a measured SCS of k. The probability $P(k \mid \mathrm{H}_i)$ for each component can be found at the intersection of the vertical line at k with the probability distribution function. The component resulting in the highest probability, i in this case, is then chosen.

The atomic column thickness estimation mentioned in the last section is an example of where the hypothesis test may be applied. The observation in such an experiment can be a certain observation related to thickness, such as SCS [67], and possible configurations of the system that form the hypotheses are thus the different thicknesses of the columns. Ideally, an accurate hypothesis can only be built if the environment of that column is perfectly known, but since the sample is mostly unknown in advance, some assumptions are thus necessary to support the construction of these hypotheses. Assuming that the lateral incoherence nature of HAADF imaging and the channeling effect of the electron probe [77] completely eliminate the influence of the neighboring columns to the contrast of a single column, the SCS should also be invariant to the environment. With that assumption, and the extra assumption that the crystal surface is flat so that every atomic column starts at the same height, the hypotheses can thus be built with the SCSs calculated from simulated HAADF images of perfectly orthorhombic crystals with a range of crystal thicknesses.

The criteria for running a hypothesis test can now be checked. First, as scattering events can be simulated, the SCS or other measurements related to the column thickness can be computed. If the probability density function of the distribution of the measured value can also be found, for example, the Poisson distribution may be used to describe the distribution of SCS, the likelihood of the hypotheses can then be computed and thus fulfill the first criterion. Second, if an estimation of the maximal thickness of the sample can be made in advance, the range of tested thicknesses can be bounded, for example, by testing up to twice or three times the maximum. With that, the number of hypotheses is also limited and therefore the second criterion is fulfilled. As both criteria matched, the hypothesis test (equation 2.14) can be performed, and the selection of the hypothesis is shown in figure 2.7.

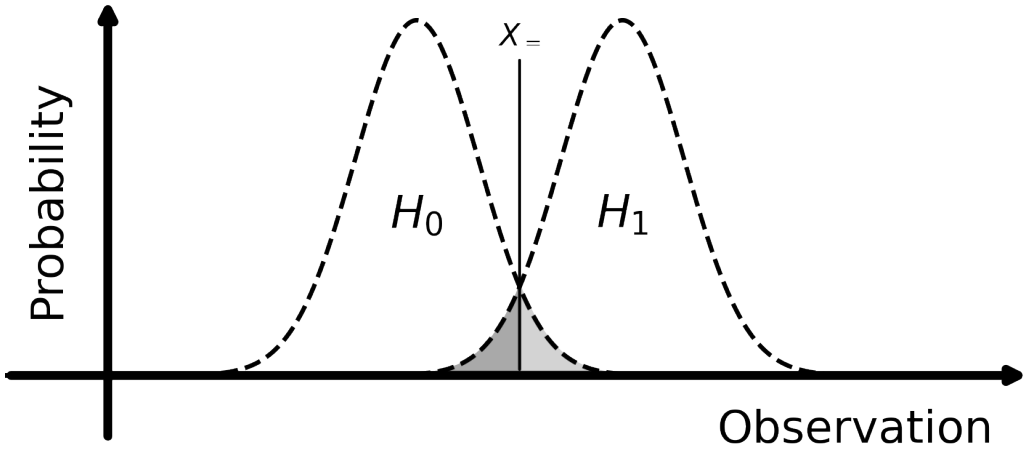


Figure 2.8: Visual representation of the PE can be found as half of the overlapped region of the hypotheses. The intersection point in between is marked by $X_=$.

This method is not restricted to a single-dimensional observation, such as an SCS calculated for each atomic column. The dimensionality of the parameter describing the atomic columns can be expanded on the sample plane as well as on the detector plane. On the sample plane, instead of using the sum of the intensities or use the volume of the fitted Gaussians as a single-value representation of the atomic column, one can opt to use the intensity distribution over a certain number of pixels directly as the observation. To be more precise, given a pixel cluster containing I pixels, instead of calculating the SCS by summing the value from the cluster, all I values form a set that describes a single atom column. The expansion of dimensionality on the detector plane is achieved by using multiple detectors, such as acquiring simultaneously the images with HAADF and LAADF or using a pixelated detector to acquire 4D datasets. Assuming from each probe position J channels of signals are collected from independent detecting units, the resulting observation is hence J-dimensional and can be presented as a $\vec{k}_J$. Combining the two possible expansions, the probability of obtaining a specific observation $\vec{k}_{I,J}$ can be found by the joint probability distribution that describes the probability distribution of a $(I \times J)$-dimensional vector:

$$P(\vec{k}_{I,J}) = \Pi_{i=1}^{I}\Pi_{j=1}^{J}\frac{\lambda_{i,j}^{k_{i,j}}e^{-\lambda_{i,j}}}{k_{i,j}!}, \tag{2.15}$$

Probability of error and experiment design

Besides the analysis of experimental data, the hypothesis test provides insight into the optimization of the experiment setup. In cases where the goal of an experiment can be formulated as a hypothesis test, the performance of the condition or the setup with

which the experiment is run can be characterized by the probability of choosing the wrong hypothesis based on the resulting observation, also known as the probability of error (PE). The probability of choosing a wrong hypothesis can be expressed as $P(\mathrm{H}_i \mid \mathrm{H}_j)$ with $i \neq j$, which means the probability of H_i is chosen while H_j is the ground truth. Take the simplified two-candidates hypothesis test as an example, the PE is shown as follows:

$$\begin{aligned} P_e(\mathrm{H}_0, \mathrm{H}_1) &= \frac{1}{2}P(\mathrm{H}_0 \mid \mathrm{H}_1) + \frac{1}{2}P(\mathrm{H}_1 \mid \mathrm{H}_0) \\ &= \frac{1}{2}\mathrm{CDF}_1(X_=) + \frac{1}{2}\left(1 - \mathrm{CDF}_0(X_=)\right) \end{aligned} \tag{2.16}$$

Equation 2.16 states that the probability of error is the sum of the probability that H_0 is picked given that H_1 is the ground truth and vice versa, divided by the corresponding prior probability of each hypothesis, which is assumed to be equal for both hypotheses and thus 1/2 in this case. The PE can be shown as the sum of the overlapped regions between the probability distribution functions of the hypotheses (figure 2.8), as an observation found in this region leads to a wrong prediction by the hypothesis test. The overlapped region can be calculated using the cumulative distribution function (CDF) of the hypothesis, which is the integral of the probability distribution function from the left or the right to the intersection between the two distributions $X_=$.

The optimization of an experiment can hence be done by observing the relationship between the parameters and the resulting PE to find the optimal condition where the PE is minimized. For example, De Backer et al. [67] have demonstrated detector range optimization for atomic column thickness estimation, where the PE for determining the right amount of atoms in the inspected column is found to be minimized when the detector occupies the entire dark field.

2.4 Conclusion

In this chapter, methods that enable the study and the utilization of information carried by the elastically scattered electrons are introduced. The concept and theory behind multislice TEM simulation are first introduced. Such simulation is largely utilized in the works included in this thesis, for example, the verification of the imaging methods in chapters 3, 4, and 5, and also the generation of training data for the convoluted neural network in chapter 4. Following the introduction to the TEM simulation, some advanced imaging methods that utilize the phase contrast in the recorded image of electron wave intensity are then covered. With these imaging methods, important background knowledge about wave modulation, which is the key concept for chapters 5 and 6, and 4D-STEM reconstruction, which is compared with the proposed methods in chapters 3 and 4, are introduced. In the final section, a few methods of statistical analysis based on TEM experiment results are discussed. Among the methods, the PE and the optimization of experiments by minimizing the PE form the essentials of the discussion about wave design in chapter 6.

Chapter 3

Real-Time Integration Center of Mass Reconstruction for 4D-STEM

This chapter is based on:
***Yu, C.P.**, Friedrich, T., Jannis, D., Van Aert, S. and Verbeeck, J., 2022. Real-time integration center of mass (riCOM) reconstruction for 4D STEM. Microscopy and Microanalysis, 28(5), pp.1526-1537.*

The author has major contribution to the conceptualization of the paper, conducting TEM experiments, and creating the software for data processing and imaging.

3.1 Introduction

The higher response rate and shorter dead time of the DED pixelated detectors lower the difficulty to acquire 4D-STEM data, compared to the time when CCD detectors were the only option [49]. With the availability of such detectors increasing in different research centers, the interest in developing methods to process 4D-STEM datasets is also growing These datasets capture the outcome of intricate scattering events, and if the embedded information can be extracted and utilized, the 4D-STEM technique has the potential to push the limit of electron imaging to work under even more critical conditions, for example, lower dose and faster sampling [78], or to exhibit higher resolution [79].

Despite showing great advantages over traditional imaging methods, the accompanying demand on the processing power and time to perform these reconstructions should not be underestimated. In most cases, accelerators (e.g. GPU), powerful CPUs, and large amounts of computer memory to store the recorded data or some reduced version of it, are necessary. Even though other reconstruction methods exist that are less computationally demanding, such as iCOM, iDPC, and SSB, they are performed after the acquisition, and thus the reconstruction result can only be updated after the raster scan, and not during.

Due to these limitations, the 4D-STEM reconstruction methods are mostly used as post-processing methods. Operations on a TEM, such as sample search, adjusgting focus, and compensating astigmatism, whether they are done by a machine or a human operator, depend on frequent feedback so that the operations can be adjusted in time for efficient

optimization of the experimental conditions. Therefore, these operations remain to rely on the real-time update capability of HAADF and LAADF imaging, despite the fact that the dose usage in these methods can be relatively high and may lead to significant damage before the condition to record data is met.

To achieve real-time imaging based on 4D-STEM data, the algorithm should be able to render images in parallel with the scanning. It means that the streamed data from the detector needs to be processed individually or in small groups, while the benefit of an advanced imaging technique, such as high dose efficiency and sensitivity to all kinds of atom species should not be sacrificed. Proposed by A. Strauch et al. [80], a dose-efficient reconstruction with live image update can be done based on SSB. By first allocating memory for the dataset and then gradually filling it with collected and processed data during the scanning process, an update of the reconstructed image can be generated anytime by the algorithm of SSB, even before the dataset is complete. However, it also indicates that the number of probe positions in a dataset is limited by the GPU memory, as the data needs to be stored for later processing. At the current state of technology, this approach is limited in terms of processing rate to about 1000 probe positions per second, while the frame rate of direct electron detectors is approaching 100 kHz [81] and even the MHz range for data-driven cameras at suitable conditions [82].

To overcome these hardware and speed limitations, we hereby propose a new live reconstruction method based on iCOM, which does not rely on storing the entire 4D dataset in memory, does not require accelerators of any kind and thus greatly reduces the computational requirements, as well as allowing reconstructions of images of a larger scale. This choice of the core imaging mechanism is based on the fact that iCOM (and iDPC) do not utilize a complicated algorithm, and thus enjoy a lower demand on the processing power and time, while some of the mentioned merits of 4D-STEM imaging methods can be preserved.

In this chapter, the physical formulation of real-time iCOM (riCOM) is first derived, and details of the software implementation of the reconstruction algorithm are discussed. This software implements direct interfaces to DED cameras of various types, and real-time reconstructions with streamed data during acquisition are recorded, from which one can see that the tuning of the imaging conditions is immediately reflected in live-updated images. RiCOM reconstruction from existing experimental datasets is also shown, with both frame-based or data-driven format [83, 82]. Since iCOM and iDPC are both based on the intensity shift of the transmitted electron wave on the diffraction plane, and the only difference lies in how this shift is computed, the changes made to enable real-time imaging for iCOM can also be implemented for iDPC in the exact same manner.

3.2 Methods

Physical Formulation

Following the discussion in section 2.2, the computation of an iCOM image is described by equation 2.7 and 2.9. The former equation shows that the COM shift vector is related to the gradient of an object function $O(\vec{r}_p)$:

$$\mathrm{O}(\vec{r}_p) = |\psi_i(\vec{r}, \vec{r}_p)|^2 \star \phi(\vec{r}) \tag{3.1}$$

Since the integration of the gradient of a scalar function is path independent, the iCOM signal of a point $\vec{r}_p$ ideally can be found along any integration path, as long as the starting point is far enough from the material to assume that the COM shift is zero. However, in realistic cases, the measurement of COM shift contains noise, and thus a better estimation of the noise-free result can be found by taking the average of all possible path integrals. This can be achieved by averaging path integrals at all possible azimuthal angles, from infinity towards the probe position. In order for this concept to work with a 2D grid of probe positions, the averaged integral can be expressed in a discretized form:

$$\begin{aligned} & I^{\mathrm{iCOM}}(\vec{r}_p) \\ &= \int_{\infty}^{\vec{r}_p} \vec{I}^{\mathrm{COM}}(\vec{r}) \cdot d\vec{r} \\ &= \frac{1}{2\pi} \int_0^{2\pi} \int_{\infty}^{r_p} \vec{I}^{\mathrm{COM}}(\vec{r}) \cdot \hat{n} dr d\theta \\ &= \frac{a}{2\pi} \sum_{x=-\infty}^{+\infty} \sum_{y=-\infty}^{+\infty} \frac{\vec{r}_p - \vec{r}_{xy}}{|\vec{r}_p - \vec{r}_{xy}|^2} \cdot \vec{I}^{\mathrm{COM}}(\vec{r}_{xy}). \end{aligned} \tag{3.2}$$

In equation 3.2, the probe position from which the iCOM value is computed is given by $\vec{r}_p$, and the COM shifts are measured from $\vec{r}$, which is then discretized to $\vec{r}_{xy}$. The vector $\hat{n}$ describes a unit vector that always points towards $\vec{r}_p$, and a describes the finite area assigned to each $\vec{r}_{xy}$, as a result of discretization. Notice that the three vectors $\vec{r}_p$, $\vec{r}$, and $\vec{r}_{xy}$ are all on the sample plane, and $\vec{r}_p$ and $\vec{r}_{xy}$ essentially form the same array, which consists of the visited probe positions from the raster scan. This reflects the fact that the sampling of the COM shift directly determines the number of pixels and the pixels' position on the sample plane for the reconstructed iCOM image.

The discrete representation in equation 3.2 states that the summation has to go over an infinite amount of points, or at least all probe positions in the dataset (as for iCOM reconstructions) in order to acquire or to approximate the desired object function, given by equation 2.9. This requires the COM shift array to be filled at every composing position $\vec{r}_{xy}$ first, indicating the reconstruction can start only after the data acquisition step is complete, and thus a live update of the partially filled dataset is impossible. It is therefore implied that the real-time reconstruction based on the COM shift signal can be enabled by limiting the range of the summation, which produces the result of riCOM reconstruction:

$$I^{\mathrm{riCOM}}(\vec{r}_p) = \frac{a}{2\pi} \sum_{x=r_{p,x}-\frac{n-1}{2}}^{r_{p,x}+\frac{n-1}{2}} \sum_{y=r_{p,y}-\frac{n-1}{2}}^{r_{p,y}+\frac{n-1}{2}} \frac{\vec{r}_p - \vec{r}_{xy}}{\left|\vec{r}_p - \vec{r}_{xy}\right|^2} \cdot \vec{I}^{\mathrm{COM}}(\vec{r}_{xy}) \tag{3.3}$$

As mentioned before, $\vec{r}$ and $\vec{r}_{xy}$ come from the same set of visited probe positions, and therefore the process described in equation 3.3 can be written as a cross-correlation

between $\vec{I}^{\mathrm{COM}}(\vec{r}_p)$ and $\frac{\vec{r}_p}{|r_p|^2}$:

$$
\begin{aligned}
I^{\mathrm{riCOM}}(\vec{r}_p) &= \vec{\mathrm{K}}_n(\vec{r}_p) \star \vec{I}^{\mathrm{COM}}(\vec{r}_p) \\
\vec{\mathrm{K}}_n(\vec{r}_p) &= \begin{cases} \frac{a}{2\pi}\frac{\vec{r}_p}{|r_p|^2}, & |r_p| < \frac{n-1}{2} \\ 0, & \text{else} \end{cases}
\end{aligned}
\tag{3.4}
$$

In equation 3.4, the action of summation is replaced by cross-correlation with a kernel function $\vec{\mathrm{K}}_n(\vec{r}_p)$, where n indicates the range of summation. The real-time imaging ability is manifested by the property of the kernel. It is presented as an array mostly filled with zeros, except a $n \times n$ square at the center. This means that the calculation of $I^{\mathrm{riCOM}}(\vec{r}_p)$ only takes into account $n \times n$ values from the array of $\vec{I}^{iCOM}$, which can be significantly fewer than the total amount of probe positions in a dataset.

Kernel Design

To show how the size of the kernel affects the reconstructed image, we start with the Fourier transform of the object function $\mathrm{O}(\vec{r}_p)$, which is the result of iCOM reconstruction.

$$
\begin{aligned}
\mathrm{FT}\left\{O(\vec{r}_p)\right\}(\vec{k}_p) &= \mathrm{FT}\left\{\int_{-\infty}^{\vec{r}_p} \vec{I}^{\mathrm{COM}}(\vec{r}) \cdot d\vec{r}\right\}(\vec{k}_p) \\
&= \mathrm{FT}\left\{\vec{I}^{\mathrm{COM}}(\vec{r}_p)\right\}(\vec{k}_p) \cdot \frac{-i\vec{k}_p}{|k_p|^2}
\end{aligned}
\tag{3.5}
$$

where $\vec{k}_p$ is a vector in the Fourier domain with a counterpart $\vec{r}_p$ in the real domain. Equation 3.5 shows that, for the computation of $O(\vec{r}_p)$, the integration process of the COM shift can be done by multiplication between the Fourier transform of the COM shift array and the transfer function $\frac{1}{i\vec{k}_p}$, reaching the same conclusion as equation 2.9. The change in magnitude of the transfer function with respect to the magnitude of $\vec{k}_p$ can be seen as a weighting difference given to the COM shift array in the frequency domain, which results in much stronger attenuation for the high spatial frequency components.

By integrating over a finite range, an analytical expression for the riCOM result can be obtained as follows:

$$
\begin{aligned}
&\mathrm{FT}\left\{I^{\mathrm{riCOM}}(\vec{r}_p)\right\} \\
&= \mathrm{FT}\left\{\int_{\vec{r}_p-\Delta\vec{r}}^{\vec{r}_p} \vec{I}^{\mathrm{COM}}(\vec{r})\cdot d\vec{r} - \int_{\vec{r}_p}^{\vec{r}_p+\Delta\vec{r}} \vec{I}^{\mathrm{COM}}(\vec{r})\cdot d\vec{r}\right\}/2 \\
&= \mathrm{FT}\left\{\vec{I}^{\mathrm{COM}}(\vec{r}_p)\right\}\cdot\frac{-i\vec{k}_p}{\left|k_p\right|^2}\times[1-cos(\Delta\vec{r}\cdot\vec{k}_p)] \\
&= \mathrm{FT}\left\{O(\vec{r}_p)\right\}\times[1-cos(\Delta\vec{r}\cdot\vec{k}_p)].
\end{aligned}
\tag{3.6}
$$

In equation 3.6, the summation of riCOM reconstruction is approximated with a 1D integration from both sides of the probe position $\vec{r}_p$, within the range of $2\Delta\vec{r}$. The result shows that by limiting the integration range, it reproduces the function $O(\vec{r}_p)$ with an extra weighting function $1-cos(\Delta\vec{r}\cdot\vec{k}_p)$. This function is close to zero when $\Delta\vec{r}\cdot\vec{k}_p$ is small and thus strongly suppresses the low-frequency signal in the retrieved object function. Also, the extra weighting peaks at $\vec{k}_p = \frac{\pi}{\Delta\vec{r}}$, which implies that by choosing smaller $\Delta\vec{r}$, or shorter integration range, weight can be redistributed to higher frequency components. By using kernels with sizes smaller than the real space dimension of the dataset, this effect of limiting integration range can be achieved. Although the actual frequency spectrum of a 2D kernel deviates, the weight of a kernel of size n at each frequency k can be well approximated with the formula derived from line integration:

$$
\frac{N}{2k_p\pi}\times[1-cos(\frac{n-1}{2}\times\frac{2k_p\pi}{N})] \tag{3.7}
$$

where N is the number of pixels of the image in one direction, and the extra factor $2\pi/N$ scales with the pixel size in the Fourier transformed result. This effect is not equal to but can be compared to a high-pass filter, as it emphasizes high-frequency information in the reconstructed image. Other filtering effects, such as low-pass or band-pass filtering, can be created by decorating the kernel with specific filters. A filter can be seen as a mask that reduces or eliminates a certain range of frequency signals of an image. This is done by multiplying the image in the frequency domain with the filter function, which is equivalent to a convolution between their real space counterparts. For riCOM images, which can be seen as a cross-correlation between a COM shift map and a kernel, the application of such a filter can be included in the design of the kernel. For example, a band-pass filter ($\tilde{f}(\vec{k}_p)$ in the frequency domain and $f(\vec{r}_p)$ in the real domain) effect can be added as follows:

$$
\begin{aligned}
I^{\mathrm{riCOM}}(\vec{r}_p) * f(\vec{r}_p) &= [\vec{I}^{\mathrm{COM}}(\vec{r}_p)\star\vec{\mathrm{K}}_n(\vec{r}_p)]\otimes f(\vec{r}_p) \\
&= \vec{I}^{\mathrm{COM}}(\vec{r}_p)\star[\vec{\mathrm{K}}_n(\vec{r}_p)\otimes f(\vec{r}_p)]
\end{aligned}
\tag{3.8}
$$

with

$$
\begin{aligned}
f(\vec{r}_p)) &= \mathrm{FT}^{-1}\left\{\tilde{f}(\vec{k}_p)\right\} \\
\tilde{f}(\vec{k}_p) &= \begin{cases} 1, & k_{max} \geq |k_p| \geq k_{min} \\ 0, & \text{otherwise} \end{cases}
\end{aligned}
\tag{3.9}
$$

In equation 3.8, $f(\vec{r}_p)$ is the filter function and the symbol $\otimes$ indicates convolution. Equation 3.9 writes a possible band-pass filter, with a hard cutoff at two frequency limits k_{max} and k_{min}, i.e. a band-pass filter. The real space counterpart of the filter can be found by performing an inverse Fourier transform FT^{-1} to the filter function in the Fourier domain. This real space filtering effect can be incorporated into the kernel using the associative property of cross-correlation and convolution. It is worth pointing out that to create a sharp cutoff at the frequency domain, one would need a filter matching the size of the COM shift array. In case the kernel is chosen to be truncated, the outcome of the convolution is reduced to keep its size the same as the kernel $\vec{\mathrm{K}}_n$. This would create damping and slope in the filter function at the cutoff and modify the resulting frequency spectrum.

In Fig. 3.1 the frequency spectra of different kernel designs are illustrated. From bottom to top, the curves correspond to the template kernel with a size of 101×101, a smaller kernel with a size of 41×41, template kernel with high-pass filter ($k_{min} = \frac{12.19}{2}\ px^{-1}$), low-pass filter ($k_{max} = 12.19\ px^{-1}$), and band-pass filter (k_{min}, k_{max} same as before). For the bottom two curves, the results of the corresponding line integration approximation (dashed lines), with $\Delta\vec{r}$ equal to half of the kernel size, are also drawn to show their similarity in oscillation frequency and magnitude.

By comparing the blue and gray curves in figure 3.1, it is clear that Kernel 41 peaks at a higher frequency than Kernel 101, as predicted by the analytical formula, and that the cutoff of the lower frequency due to a smaller kernel happens approximately at the inverse of the kernel size, indicated by the gray circle. This value is then used for k_{min} of the high-pass filter incorporated to Kernel 101-HP (orange curve), which indeed shows a similar overall frequency spectrum as the one of Kernel 41. Notice that the size reduction after convolution between kernel and the decorating filter causes a gradual decrease of frequency components below k_{min} and spectrum differences beyond k_{min}, compared to Kernel 101. Similarly, a kernel with a low-pass filter Kernel 101-LP (green curve) and a kernel incorporating a band-pass filter Kernel 101-BP (pink curve) are created. Both of them show suppression of the higher frequency ranges. Kernel 101-BP also shows a shift of the spectrum peak to a higher frequency because of its high-pass characteristic.

Despite the fact that it is not always possible to recreate the exact characteristics of common post-processing filters, the decorating filters and the choice of kernel sizes allow for great flexibility for frequency tuning and yield consistent and predictable reconstructions. Since these filters become a part of the kernel, the real-time imaging ability of riCOM is preserved with their incorporation.

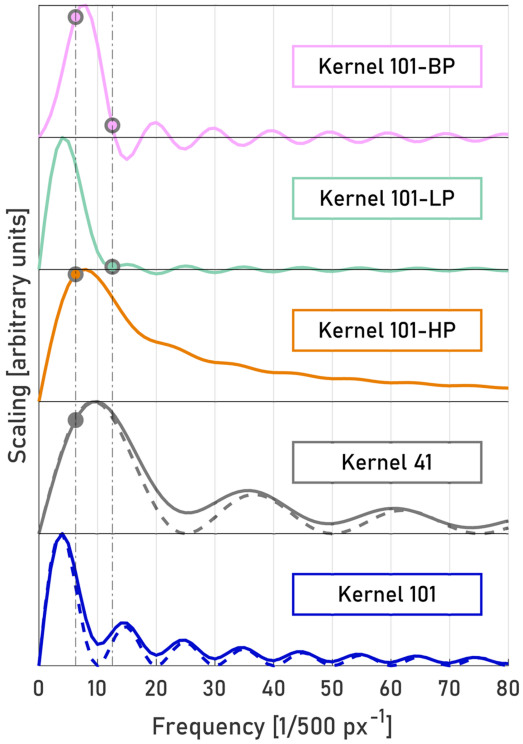


Figure 3.1: Frequency spectra of a set of kernels acting on a COM shift map of size 500×500. The presented examples include, from bottom to top, the template kernel with a size of 101×101, a smaller kernel with size of 41×41, template kernel with a high-pass filter, low-pass filter, and band-pass filter. The dashed line shows the predicted transfer function with line-integration approximation. The two vertical lines indicate the cutoff frequency of the filter or the inverse of the kernel size, and the circles at the intersection of the vertical lines and integral indicate whether a cutoff frequency is applied to the specific design.

Data Processing

In order to achieve real-time imaging on a TEM, software is created to handle the reading, processing, and visualization of the reconstructed result. Interfaces with various server programs of the cameras, provided by Quantum Detector [84] (MerlinEM, based on Medipix3) and Amsterdam Scientific Instrument [85] (Cheetah, based on Timepix3), are implemented, with which the detector output can be delivered to the software via TCP connection. Besides small differences in communication, specific processing procedures are also implemented, meeting the property and output format of each detector.

For real-time imaging with a frame-based camera, the readout time of a frame and the time to calculate the COM of the frame are normally the most time-consuming steps in the whole process instead of performing cross-correlation. As a result, the readout, COM calculation, riCOM construction can be run in a serial manner, since the bottleneck of the process cannot be parallelized (the readout).

On the other hand, data handling for a data-driven camera can be much faster and lighter. Following the discussion in section 1.3, the Timepix detector excels when used for low-dose imaging since the amount of data output is proportional to the number of recorded electrons. For example, if each CBED frame contains 50 events, the amount of output of a data-driven camera is more than 1000 times smaller than a frame-based one ($50/256^2$, also dependent on bit depth). Calculation of the COM is also simpler. Since every piece of data contains information about its location on the detector, the calculation of the COM can simply be done by averaging the x and y coordinates on the detector at each probe position (see table 1.1). The corresponding probe position of each event, however, does take some extra effort to retrieve. If the detector is synchronized with a common clock with the scanning engine, the probe position is simply found by dividing the TOA by the dwell time. In case these two are not synchronized, an extra trigger signal to indicate a certain probe position, such as the beginning and the end of each line, is required. With the timestamps at these signature points, the TOA can be normalized within each line, and the probe position can thus be calculated:

$$\begin{aligned} r_x &= \frac{\mathrm{TOA} - t_0}{t_e - t_0} \times n_{\mathrm{line}} \\ r_y &= \text{trigger count} \end{aligned} \tag{3.10}$$

where r_x, r_y are the x, y coordinates of the probe position, t_0 and t_e are the time in the beginning and the end of the line that contains the corresponding probe position, respectively, and n_{line} is the number of pixels in a line. The y coordinate can be simply found as the number of triggers that either mark the beginning or the end of the line.

The calculation of the cross-correlation between the kernel and the COM shift, as described in equation 3.4, can be done in different approaches. Figure 3.2a shows the conventional operation of cross-correlation with which the reconstruction is delayed by $\frac{n-1}{2}$ lines since the kernel can only be applied once the data of the required probe positions are acquired. This delay can be mitigated by decomposing the COM shift array:

$$\begin{aligned} I_{xy}^{\mathrm{riCOM}}(\vec{r}_p) &= \vec{\mathrm{K}}_n(\vec{r}_p) \star \left[\vec{I}^{\mathrm{COM}}(\vec{r}_p)\delta(\vec{r}_p - \vec{r}_{xy}) \right] \\ I^{\mathrm{riCOM}}(\vec{r}_p) &= \sum_{x=1}^{N_x} \sum_{y=1}^{N_y} I_{xy}^{\mathrm{riCOM}}(\vec{r}_p) \end{aligned} \tag{3.11}$$

The contribution from the COM shift measured at an individual probe position $\vec{r}_{xy}$ can be found by multiplying the COM shift array with a delta function that peaks at the concerned position, which is written as $I_{xy}^{\mathrm{riCOM}}(\vec{r}_p)$. Since the delta function selects only one pixel from the COM shift array to be cross-correlated with the kernel, the result $I_{xy}^{\mathrm{riCOM}}(\vec{r}_p)$ will be of the same size as the kernel, as shown in figure 3.2b. With this approach, the contribution from the measurement at each probe position can be calculated and added to the reconstruction image separately and thus without delay.

The live update with the first approach would unavoidably be delayed for a few lines. However, for a dwell time as short as 1 μs, or a frame rate of more than 1 Mhz, an

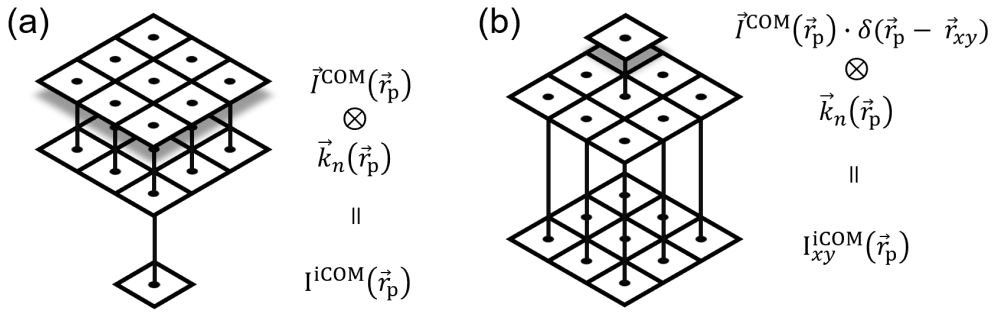


Figure 3.2: The two approaches to perform riCOM reconstruction. (a) The conventional approach where the cross-correlation can be seen as the sum of the products of the COM shift array multiplied by a moving kernel. (b) The decomposed approach where the contribution of the individual COM shift vector is calculated and added to the riCOM reconstruction.

update delay of 100 lines would only make a time difference of 0.1 s for an image of size 1000×1000, and thus the first approach is taken for the Timepix detector since such short dwell times are indeed achievable with it. Despite the delay, the first approach possesses a favored property since the process can be easily parallelized. The summation described in the first approach utilizes multiple data of the COM shift, whose computation is fairly light, to compute a single value in the riCOM array, which can have a much heavier load when the kernel is large. The computation-heavy part can thus be safely handled by multiple threads since each thread worker writes one value to the riCOM array at non-repeated positions, while the simpler process of calculating the COM is done in series. This allows the process to run without the need for an extra lock system to avoid race conditions, which may introduce extra overhead and lower the processing speed. For the Medipix detector, the second approach is taken since the update can be performed at the individual pixel level with a frame rate less than 14 khz, under 1-bit mode.

It is worth mentioning that with the current implementation, the maximal processing speed (not to be limited by the readout of the detector) for a frame-based detector, operated under 1-bit mode, is approximately 35 kHz, and for a data-driven camera, the software can process more than 80 M events per second when the kernel is small ($n < 11$), translated to approximately 2 MHz of frame rate for a low dose imaging condition of 40 electrons per probe position.

The program was developed as a cross-platform application that can be run through a command line interface (CLI) or interactively through a graphical user interface (GUI) as shown in figure 3.3. The core functionality of the algorithm is implemented in a single C++ class object. Visual interfaces interact with an instance of that class across threads through pointers, which allows live updates to be displayed while maintaining a responsive interface without interrupting the reconstruction process. Furthermore, kernel settings for riCOM reconstruction and virtual STEM (vSTEM) settings, such as rotation angle due to Lorentz force, kernel size, filter, and the virtual detector size, can be changed during the process without interruption, which is helpful to find suitable settings interactively while spending the lowest amount of dose on the precious sample area.

The riCOM based class is independent of specific camera models and data types, while additional dedicated classes provide live and file interfaces for given camera types/file formats. This allows for easy extendability of the program by simply including further interface classes. The current implementation is available on GitHub under a GPL license.

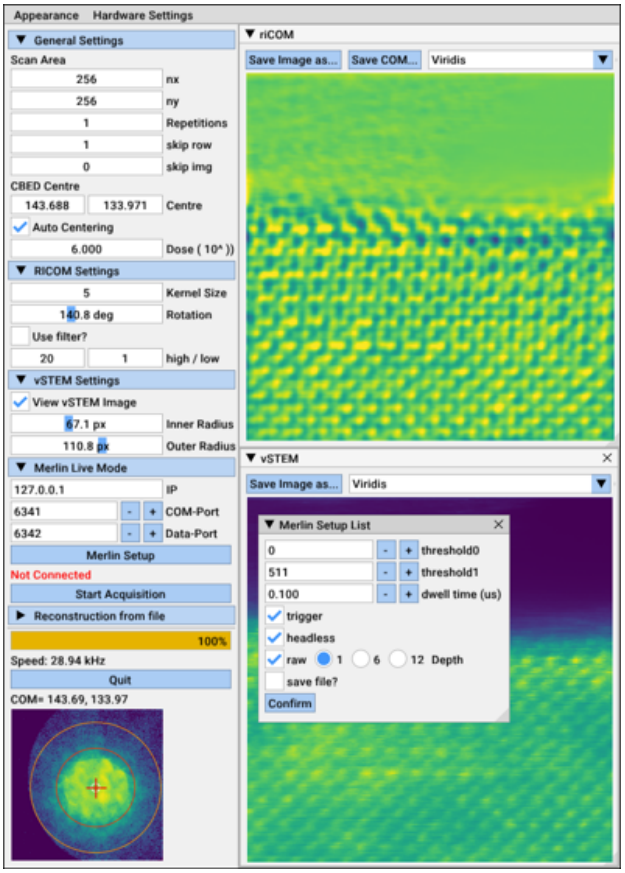


Figure 3.3: Layout of the GUI. The Menu column on the left allows the user to change various settings, such as scan size, riCOM Kernel, filter settings, virtual STEM settings, and interfaces for live mode and file dialogues. During a running reconstruction, a CBED is plotted at the bottom of this menu to visually assist the tuning of the pattern center and integration area for vSTEM interactively. All other windows are floating panels and can be moved and resized.

3.3 Experimental Details

The results presented in this paper are produced from data collected in three experiments. In the first experiment, a $SrTiO_3$ focused ion beam (FIB) lamella is examined with a probe corrected Thermo Fisher Titan^3 (X-Ant-TEM) operated in STEM mode. The resulting CBEDs are collected with a MerlinEM direct electron detector and form 4D datasets for further analysis, as well as movies demonstrating the real-time processing power of the

method. The experiment is performed with a beam energy of 300 keV and a convergence angle of 20 mrad.

The second STEM experiment is performed on a silicalite-1 zeolite sample with a Thermo Fisher Themis Z (Advan-TEM). The data is collected with a custom-made Timepix3 detector based on an Advapix TP3 camera unit and is recorded in an event-based format. The beam energy and convergence angle used in the second experiment are 200 keV and 12 mrad, respectively.

The third experiment ran on a Thermo Fisher Tecnai G2, with a sample of periodic mesoporous organosilica (PMO). The data is collected with a Cheetah camera equipped with a Timepix3 detector. The TEM is operated under 200 keV, and convergence angles of 9 and 3.8 mrad are used.

All the datasets and movies recorded in both experiments, including necessary parameters for the reconstruction, can be found in the online repository [86].

3.4 Results and Discussion

Real Time Reconstruction

To demonstrate riCOM imaging, the software for real-time reconstruction is run directly on incoming data during live experiments. The computer receives frames of CBEDs from the detector, and the software reads the data through a TCP socket. Throughout the process, the only extra prior knowledge to be provided to the algorithm is the COM of an undiffracted pattern in the vacuum so that the relative COM shift at each probe position can be computed. Alternatively, it can also be approximated by averaging the COM from multiple probe positions, thereby omitting any calibration steps, making this method equivalent to more traditional imaging methods regarding ease of use. This step also inherently corrects for systematic shifts of the CBED away from the center of the detector. While scanning, some of the most basic parameters of the microscopic imaging system are tuned, for example changing the defocus, astigmatism, and magnification, as shown in figure 3.4a-c. The live updated results are recorded in movies and can be found in the online repository [86].

Defocus broadens the intensity distribution of the electron probe, and astigmatism has the effect of creating two focal points, making the beam first focused in one direction and then the other when traveling down along the optical axis. This would reduce the electron beam sharpness and make the beam elliptical if out of focus, resulting in stretched atomic features in the images, as can be seen in figure 3.4-b in the region scanned before achieving the right focus.

According to the object function (equation 3.1), the intensity in the iCOM image equals the cross-correlation between the projected electric potential of the material and the probe function, and therefore the reduction in contrast as well as distortions of the atomic features in the riCOM reconstruction is directly related to these beam aberrations. Hence, microscopists can tune optical conditions intuitively to maximize contrast and produce circular atoms with live updated results.

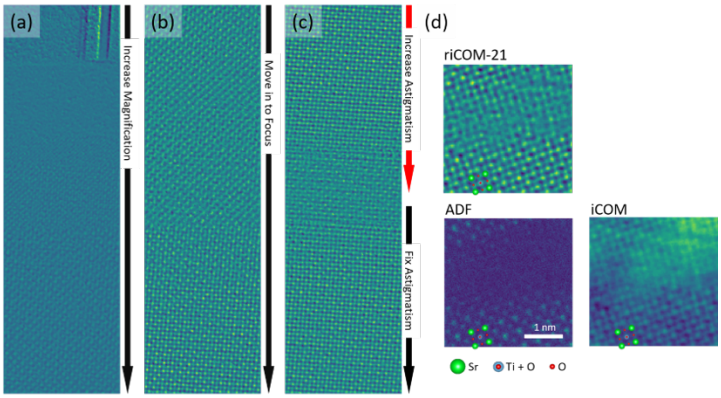


Figure 3.4: Real-time reconstruction of $SrTiO_3$ while tuning the magnification, defocus, and stigmator. (a) The magnification is increased during the scanning. In the topmost part, the contrast reveals a layered structure of the FIB lamella, and with increasing magnification, the atoms can be captured in the image. (b) Tuning focus is reflected by the change in the shape of atomic columns. (c) Tuning the stigmator affects the electron probe's sharpness and also the contrast between atoms and the vacuum. (d) Simultaneous imaging using riCOM, ADF, and iCOM. RiCOM successfully images the crystalline structure in the center of the image and the O columns, which are missing in the ADF image. The small kernel size used in riCOM reconstruction reduces the long-range intensity distribution shown in iCOM.

By changing the magnification during the scanning process the step size is changed accordingly. The live process can still continue, although the intensity needs to be adjusted since a is changed as the scan step size is changed, as shown in equation 3.2. Besides, the optimal kernel size changes with magnification, as the spatial frequency of the desired features will be shifted when the step size is changed. However, since the kernel size can be adjusted during the process, a suitable choice can always be found by tuning the kernel size according to the quality of the live updated reconstruction image.

In figure 3.4-d, a riCOM image rendered with a kernel size of 21 is compared to the ADF image and the iCOM result. Apparent differences can be found in the center of the images, which appears to have a hole according to ADF result but shows some crystalline structure in the riCOM and iCOM images, indicating a possible extension of the crystalline material with lower thickness. ADF gives more significant contrast for differences in scattering ability, making it easier to distinguish Sr columns from $Ti + O$ columns, but also reduces the intensity of weak scatterers, such as thin regions and the pure O columns, to a level that is completely invisible, while riCOM and iCOM successfully image all three types of columns with a trade-off of less distinction between the columns. On the other hand, atomic structures are blurred by the long-range intensity variation in the iCOM result. The origin of this variation could be a local strain, misorientation, contamination, charge accumulation, etc., but it is very difficult to pinpoint the actual cause. RiCOM with an appropriate kernel size suppresses these low-frequency signals and shows a clear image of atomic columns.

The examples in figure 3.4 show how riCOM images can be used to fine-tune optical sys-

tems in a similar manner as using ADF. Moreover, the method is superior to ADF imaging in terms of required electron dose and provides contrast also for the weak scatterers in the object, including thinner regions or atomic columns composed of lighter atoms. The high-pass characteristic of the suitable kernel size has shown to be helpful in highlighting features of higher spatial frequency and reducing low-frequency components, but it also means that the contrast interpretation has to be evaluated carefully, especially for quantitative analysis, as they can be affected by multiple factors unknowingly.

Comparison of Reconstruction Methods

In this section, results from the riCOM reconstruction are compared with other reconstruction methods that have the potential to provide real-time imaging. For 4D datasets, ADF images can be computed using a virtual detector that integrates all electrons in a specified region of the detector. The summing process is independent of the probe position and does not require information beyond the scope of a single diffraction pattern, thus making virtual ADF reconstruction possible for real-time visualization of the dataset. To showcase the performance of riCOM reconstruction, it is compared to both ADF as a traditional imaging mode, and SSB, which is generally considered as a highly dose efficient and quantitative ptychography method. For riCOM reconstruction, three results generated using different kernels are put into comparison, including two kernel sizes and one kernel incorporating a band pass filter.

The dataset used for the comparison is a 4D dataset recorded from a silicalite-1 zeolite specimen. The dataset is recorded in a sparse array, in which the location where electrons hit the detector and the arrival time is recorded. This type of data format has several advantages over more commonly seen frame-by-frame types at suitable experiment conditions. For instance, in the case of low-dose imaging, sparse arrays result in datasets many times smaller than full-frame arrays, since only the pixels of the detector that successfully capture an electron generate data, while other inactive pixels remain silent. For riCOM reconstruction, this format also shows its strength in terms of processing speed. Yet another important feature of this format is that the arrival time can be used to adjust the dose in the post-reconstruction stage. Since the arrival time of each electron is recorded, the amount of dose put into the reconstruction algorithm can be post-adjusted by reducing the acceptance time from each probe position. For example, with a dataset recorded with a beam dwell time of 6000 ns, the dose for the post-reconstruction can be reduced to 1/3 of the original dose if the acceptance time is set to be 2000 ns since any electrons that arrive at the detector after the acceptance time for each probe position will be discarded.

Accordingly, five data treatment algorithms/setups are used for the experimental data at 3 different dose levels. The results are presented in figure 3.5. Comparing the images generated by a virtual ADF detector with other reconstruction methods, it is obvious that even with the maximum dose, it is not enough to generate an interpretable ADF image. The vertical lines in the ADF image are a result of the camera being inactive for unknown reasons, which is discussed in previous work [82]. This however makes an almost unnoticeable difference to other reconstruction methods, since the value of each pixel in the reconstructed image not only depends on the corresponding probe position but also on its surroundings. SSB reconstruction includes a step to integrate specific regions in the CBEDs according to their spatial frequency by performing Fourier

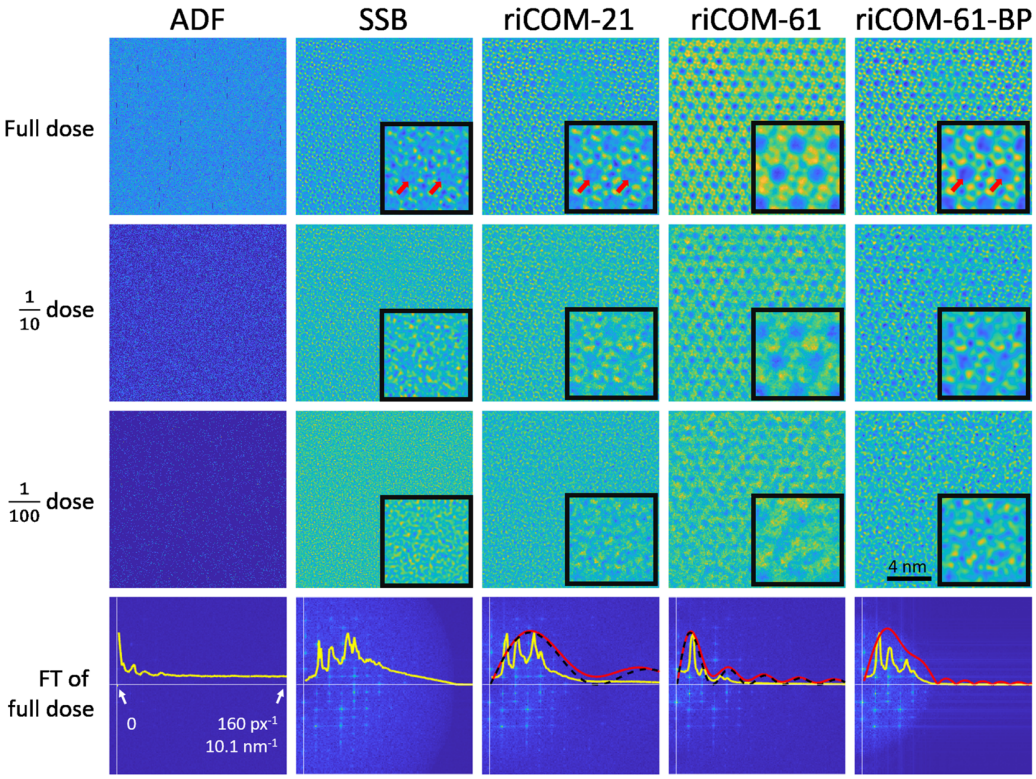


Figure 3.5: Reconstructed image from an experimental zeolite dataset with different doses (Full dose: $1.27e + 4$ e/Å^2). ADF images are generated by integrating the intensities in the detector area beyond the convergence angle at each probe position. For SSB reconstruction, a frame-based dataset is first generated from the event array, with the detector space binned down to 32×32 (8 times smaller). For riCOM reconstruction, three different kernels are used: 21-by-21, 61-by-61, and 61-by-61 with a band-pass filter. The insets show magnified versions of the center of their respective images, and the red arrows point out intensity fluctuations within the holes. The last row shows the Fourier transform of each reconstructed result. The radial averaged frequency spectra are represented with yellow curves, the frequency components of each kernel in red, and the line-integration approximation in black dashed curves.

transformation with respect to probe position. Certain spatial frequencies are weighted more strongly from a larger integration area, and the resulting contrast transfer function thus acts as a band-pass filter[87, 65]. The riCOM images of smaller kernel size (riCOM-21) are shown to be similar to the SSB results, also manifested by the similarity of their frequency spectra, as low-frequency signal is suppressed. For riCOM-61 result, by using a larger kernel size, more components at lower spatial frequencies can be found in the image. These components greatly increase the contrast for the long-range structure in the material, such as the pores and framework of the zeolite crystal but reduce high-frequency components, making the short-range structures such as atomic columns less clear. This is especially highlighted in the result with 1/10 total dose. However, by integrating a bandpass filter to the big kernel (riCOM-61-BP), noise from the high-frequency parts is removed and weights are redistributed to mid-range components from the low-frequency end. It results in a much clearer image of the atomic structure even at 1/10 dose. The filter used for the last column is designed to remove signals from 3.8 nm^{-1} to 1.14 nm^{-1}, with k_{max} = 60 px^{-1} and k_{min} = 18 px^{-1}.

In the third row, only 1/100 of the electrons in the dataset is used for imaging. The insufficient amount of electrons introduces a large amount of noise and hides the atomic structure in the images. Yet for the reconstruction result of riCOM-61, the pores within the zeolite framework are preserved in the image. This is possibly due to the fact that features of a larger scale are reconstructed from more data points, thus averaged over more possible integration paths. This kind of low-frequency components is only supported by kernels of larger size, explaining why other reconstruction methods shown here do not benefit from them and fail to present any meaningful information in the images.

Imaging of zeolites at atomic resolution with iDPC, a similar method as iCOM, has been demonstrated to be successful at a loerw dose between 100 and 1000 electrons per $Å^2$ [88, 89]. In a similar dose range, riCOM is capable of presenting structural features of the sample at different spatial frequencies, showing that the dose efficiency of the method is not sacrificed to enable real-time reconstruction.

While riCOM benefits from amplifying signals at specific frequencies so that clearer images of the lattice structures and atomic features can be captured, one has to bear in mind that the same effect is also applied to the statistical noise present in the experimental data. To study how noise affects the reconstructed images, one could compare results from ideal data with results from data with noise. However, for many reconstruction methods, it does not mean that the effect of noise can be simply acquired by subtracting one from the other since noise is not additive. Luckily, due to the linear independent nature, it is indeed the case for riCOM. In other words, the reconstructed image from a COM shift map with noise is exactly the same as the combination of the reconstructed image from a noise-free COM shift map and the one from pure noise. The latter is thus a suitable candidate for further noise analysis.

To demonstrate how noise is transferred to a reconstructed image at each frequency, a 4D dataset of a 20 nm thick zeolite sample is simulated according to the condition used in the second experiment (see section 3.3). The noise is separated from the dataset to reconstruct an ADF image of pure noise, and the noise-induced COM deviation is calculated by subtracting the COM shift map from the noise-included dataset from the one without noise, which further leads to the reconstruction of riCOM images of the

COM noise. The noise components of the pure noise ADF / riCOM images at different radial frequencies are plotted in figure 3.6. Two major differences between ADF and riCOM images can be found. First, the noise amplitudes of ADF images are higher when the dose is higher, but the opposite for riCOM reconstruction is observed. It is due to the fact that the ADF intensity values follow a Poisson distribution, where the noise increases with the square root of the dose while the signal scales linearly with the dose. The COM shift, on the other hand, is based on the spatial distribution of electrons rather than the cumulative intensity and thus is not directly linked to this kind of shot noise. However, the error of the COM estimation still decreases when more electrons are used. Therefore, despite different noise behaviors, the signal-to-noise ratios of both methods increase with dose. The second difference lies in the distribution of noise at different frequencies. For the ADF noise image, the noise is distributed equally at different frequencies, yet for riCOM, the noise is amplified according to the approximated weighting function based on the kernel size (eq. 3.7). Through this analysis, it is clear that not only the signal from the examined object but also the noise is affected by the weighting in the frequency domain. This greatly changes how noise appears in the reconstructed images compared to traditional imaging methods such as ADF and is worthy of the attention of the microscopists in order not to misinterpret features created by noise.

The different reconstruction results in figure 3.5 show a disagreement about the content inside of the pores that exist in the zeolite framework. Results from methods that give more weight to the high-frequency components, such as SSB and riCOM-21, show some intensity fluctuation inside of the pores, indicating the possible existence of dopants, yet these do not appear in the riCOM-61 image. In order to understand the cause of the difference, another simulation is run with the same condition to compare the reconstructed results with different kernel sizes in figure 3.7. To eliminate the possibility that this difference originates from the presence of noise, the reconstruction is done without adding noise to the dataset. From each reconstructed image, an intensity profile is drawn over the atom framework into the pore (figure 3.7a), which is indeed the vacuum as designed for the simulation. The profile reveals that for riCOM-21, the intensity increases, while riCOM-61 shows a monotonic decay towards the center of the pore (Fig. 3.7b). The intensity increase for riCOM-21 cannot be explained by the projected atomic potential, since it can only decay when moving further away from the atoms.

To investigate the origin of this false intensity, the Fourier-transformed riCOM images are analyzed (Fig. 3.5). The bright spots at the lowest frequency correspond to the periodic structure of the pores and framework. The intensities of these spots are greatly reduced in riCOM-21 but supported in riCOM-61, indicated by the approximated weighting function as red curves. This causes major differences in features that necessarily rely on such low-frequency signals. To illustrate the principle, we simplify the atom framework and the pore using a step function (Fig. 3.7c). By removing the low-frequency components, the step becomes a curve with a concave and a convex segment in the regions of the high and the low step, respectively. This step-function analogy conceptually captures the differences between the zeolite framework and the holes and explains the protruding intensity in the hole for riCOM-21 as the effect of reduced low-frequency components. For riCOM-61, such components are included by the larger kernel size so that no such phantom intensity can be found in the same area.

These examples show that the proposed method, like many other reconstruction methods, is capable of providing extra information compared to traditional imaging methods.

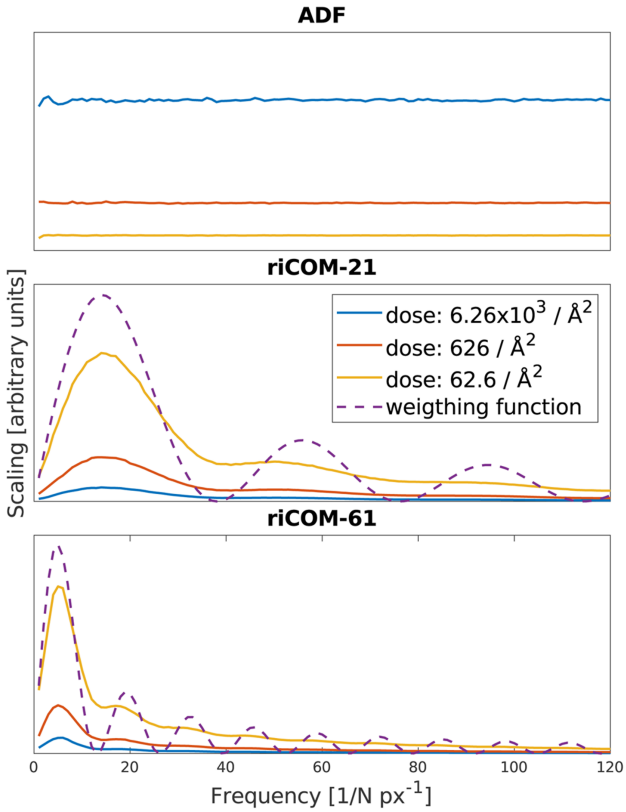


Figure 3.6: Components of the noise images at different frequencies. The noise images are rendered by applying a virtual ADF detector, riCOM with kernel size 21, and riCOM with kernel size 61. Three levels of dose for noise realization are chosen, and the curves are drawn by averaging 30 random noise configurations at each dose. The weighting functions given by the line-integration approximation are also presented for the riCOM results in dashed lines.

RiCOM also shows great dose efficiency, allowing high-quality reconstruction results under low-dose conditions. The freedom to use different kernel sizes grants users the ability to tune the desired spatial frequency range, which is very important in order to avoid misinterpretation of details in the image. Including more low-frequency components enables the reconstruction of long-range structures of the object with even lower amounts of electrons. This could be very useful for microscope operators when imaging objects of larger scale.

Effect of the electron probe size

The COM shift, as shown in equation 2.8, is related to the gradient of the cross-correlation between the probe intensity function and the projected potential. This relationship indicates that the size of the electron probe ultimately determines the resolution of the imaging mode, yet how the low-dose robustness of the imaging methods is affected by

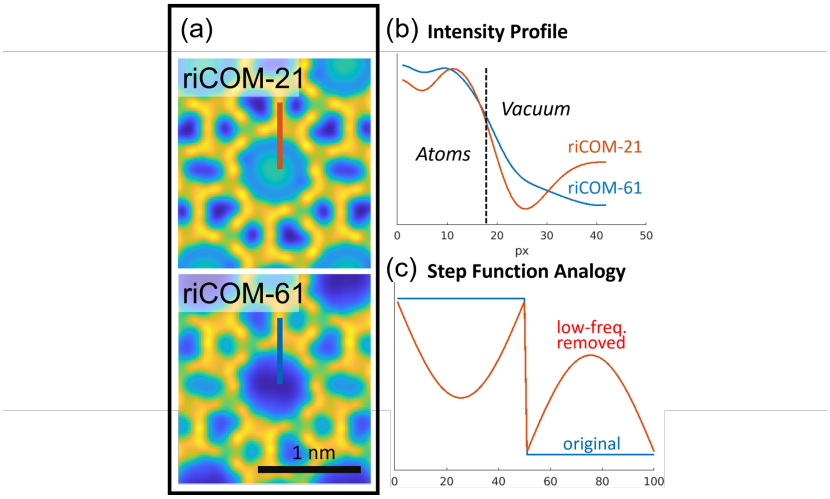


Figure 3.7: (a) Reconstruction results of a simulated zeolite dataset with different kernel sizes. The red and blue lines indicate the locations of intensity line profiles drawn in subplots (b). (b) The intensity profile shows that the intensity inside the hole area increases in riCOM-21 results but decays in riCOM-61 towards the center. (c) A step function for analogy shows that removing low-frequency components may cause imaging artifacts similar to the ones seen in reconstruction results from smaller kernel sizes.

the probe size is unclear.

To understand the effect of probe size on COM-shift signal sampling, we show a schematic of the imaging of two opposite phase ramps with three different probe sizes in figure 3.8. With the sharp probe, a clear COM shift can be acquired when the probe is placed at the ramp since only one ramp can be sampled with the narrow intensity distribution. The other probe position does not register any COM shift since the incident wave interacts with a flat phase. For the broad probe, both ramps are sampled when the probe is placed at the center of the first ramp. The contribution to the COM shift from ramps of opposite slopes cancel each other, and therefore the COM shift would be weaker than the sharp-probe case. Besides, a large proportion of the intensity falls at the flat-phase region and thus results in only weak shifts of the COM. The case where the probe size matches the feature size shows a two-fold benefit. First, the sampling is precise without any mutual cancelation in the COM shift, making the shift more significant than the broad-probe case. Second, COM shifts can be measured in other probe positions due to the broader intensity distribution of the probe compared to the sharp-probe case. Consequently, the phase gradient can be reflected from data collected at multiple positions, which may reduce the noise in the integrated results.

The connection between robustness against low-dose noise and probe-feature size relationship is then tested on a PMO sample. The PMO sample used in this experiment shows a periodic honeycomb structure. The stem of the structure is about 20 nm thick horizontally, and the distance between the pores is approximately 50 nm. For the electron probe scanning the sample, two apertures corresponding to convergence angles of 4.8 and 9 mrad are used, which create electron probes of size 0.68 and 1.61 nm at the minimum (see equation 4.16 in the next chapter for definition). The acquired data are

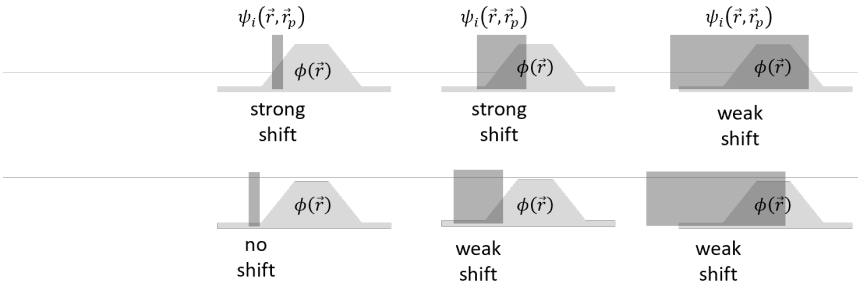


Figure 3.8: Sampling of the phase ramps with three electron probes. The intensity distributions of the probes are approximated with cylindrical functions of different widths for simplicity. Two probe positions are shown in each case for comparison: one at the center of a ramp and the other at the foot of the same ramp where the phase is locally flat.

subjected to the processing of three imaging methods: riCOM, iCOM, and ADF. Here, riCOM is performed with a kernel size of 81, and ADF images are made by a virtual detector matching the DF area in the dataset (varying inner angle matching with the aperture size). Datasets of three dose levels are generated from the raw datasets with the methods mentioned above to highlight the low-dose robustness of the reconstruction.

Figure 3.9 shows the results of the reconstruction. Comparing ADF images at the highest dose, it is seen that a higher resolution can indeed be achieved with a larger convergence angle, as the stems of the PMO appear to be thinner when the 9-mrad aperture is used. Besides that, no apparent difference in the noise level can be found between the two images. The reconstructed images of riCOM and iCOM, on the other hand, show significant differences. The reconstruction results of both methods are much noisier when a sharper probe is used, which has a larger size mismatch with the feature of the imaged object (the stems) in this case. Compared to riCOM, iCOM reconstruction only introduces more low-spatial-frequency contrast instead of improving the feature quality, which indicates that the kernel size of 81 is already sufficient to include the necessary low-frequency components of the mesoporous structure.

At the dose level of 2 e/nm^2, riCOM and iCOM reconstructions at the larger convergence angle have completely lost any periodicity resembling the object. At the same time, the periodic structure is preserved in both ADF images, as well as in the riCOM and iCOM results with the broad probe. By further reducing the dose to 0.3 e/nm^2, ADF images also lose their contrasts. The reconstruction of riCOM is indeed much noisier, yet the structure is not lost, as the position of the pores, the stems, and the edges can be found.

It is worth noting here that the contrast to present these images is not determined with the same standard. The same colormap is used for all the images, but for ADF, the minimum and maximum of the colormap range are set to be the minimal value and 0.3× the maximal value of the image. This effectively eliminates the problem of some exceptionally hot pixels and the following image contrast reduction. For the lowest-dose case, this effectively sets the maximum of the colormap range to 1, and thus the contrast of the image indeed cannot be boosted any further. For riCOM and iCOM, the minimum and maximum of the colormap range are set to be 0.6× the minimal value and 0.6× the maximal value of the image.

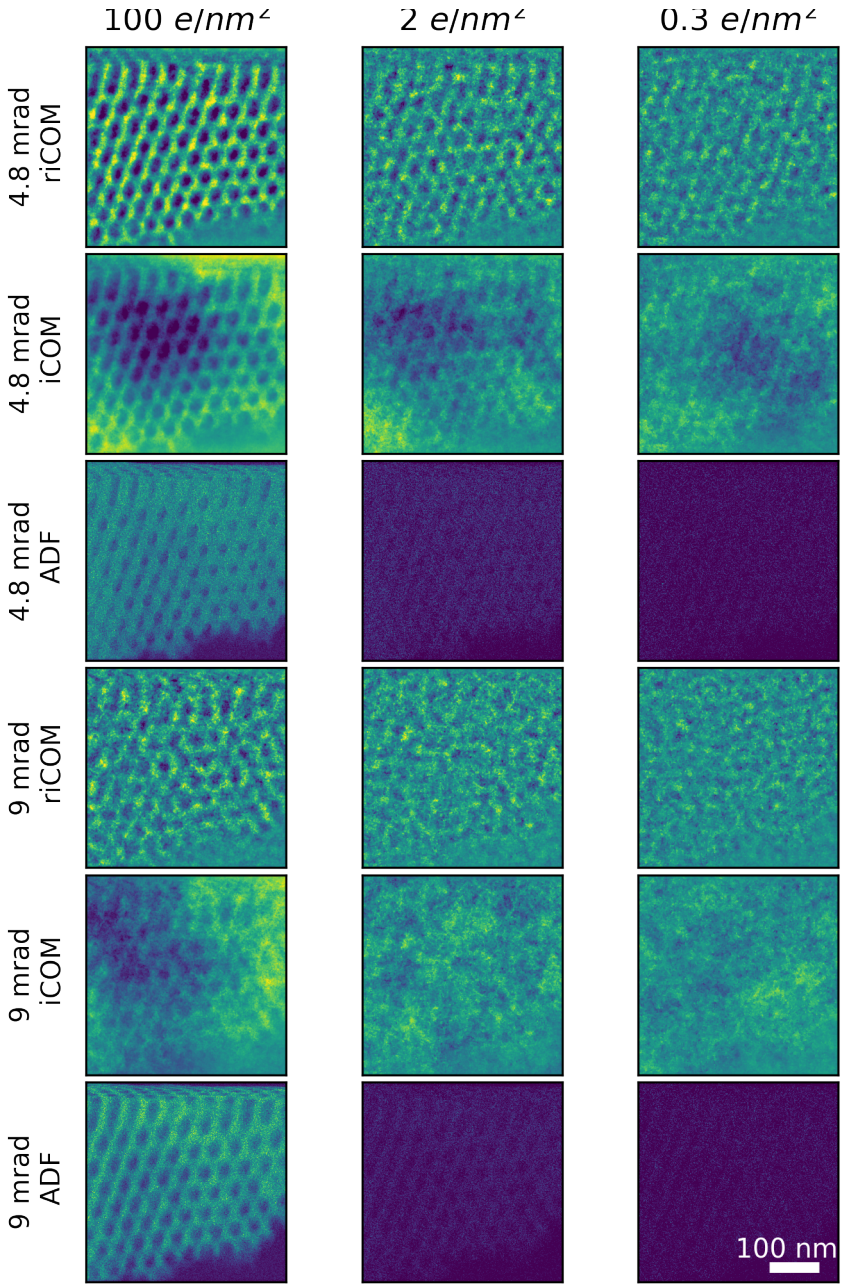


Figure 3.9: RiCOM, iCOM, and ADF reconstruction results at various dose levels and convergence angles.

In summary, this experiment shows that the dose efficiency of the methods can be significantly affected by the choice of probe size. For imaging atoms whose size is much smaller than the probe size formed with a common STEM setup, the sharpest probe can be employed in most cases with a suitable kernel size to include the necessary spatial-frequency components of the object. However, for imaging of features of larger size, for example, the shape of a particle or domain boundaries, a broader probe may be beneficial when dose efficiency is considered. Such an electron probe is not necessarily created by a smaller aperture but can also be made with the natural probe broadening as a result of defocusing, which allows a much more convenient control over the probe size. Additionally, we would also like to point out fundamental differences in the probe-dependency of SSB and riCOM. The CTFs of SSB and COM shift are both shown as functions of the probe size [90, 87, 65]. For SSB, the CTF correlates the phase profile of the object with the final reconstructed image, and therefore the probe size determines the frequency components that can be efficiently transferred to the final image, as well as the maximal resolution. For riCOM, only the COM shift is related to the probe size, while the kernel size determines the weighting of the frequency components. Therefore, for cases of imaging atom columns in a crystal, the largest convergence angle allowed by the optical system can be used to guarantee sharp atomic features, while the spatial frequency and the lattice contrast can be taken care of with an adequately selected kernel size.

3.5 Conclusion

In this work, a real-time imaging method utilizing 4D-STEM data is proposed. The core imaging mechanism is based on a vector describing the intensity shift measured from each probe position, which may be the COM shift of the CBED or DPC signals acquired with specific detectors. Through derivation of the physical formulation, we illustrate the physical relevance and the benefits for numerically efficient implementations of this approach, motivating the application, particularly in real-time imaging scenarios. The freedom to change the size of the kernel or incorporate filters is also discussed, with examples showing their effect.

It is shown that riCOM can effectively reproduce iCOM results, but allows for more flexibility in terms of selecting contributing spatial frequencies. The method, including frequency band pass filtering, depends only on the individual intensity distribution (or CBED) at its corresponding real space location, which in combination with a rather simple algorithm, creates a uniquely flexible and fast reconstruction method that requires very little user input. We further present a well-optimized, interactive GUI implementation, developed in standard C++ and published open source on GitHub.

Demonstrations of the method on an operating microscope show that firstly, the process is fast enough to keep up with the highest frame rate supported by the currently available frame-based detector, as well as the highest beam current allowed by the available data-driven detector, and secondly, providing dynamic feedback to the microscope operator when tuning and optimizing the microscope parameters. This ability enables a swift search of the sample or region of interest, as well as adjustments of the imaging conditions at potentially very low dose conditions.

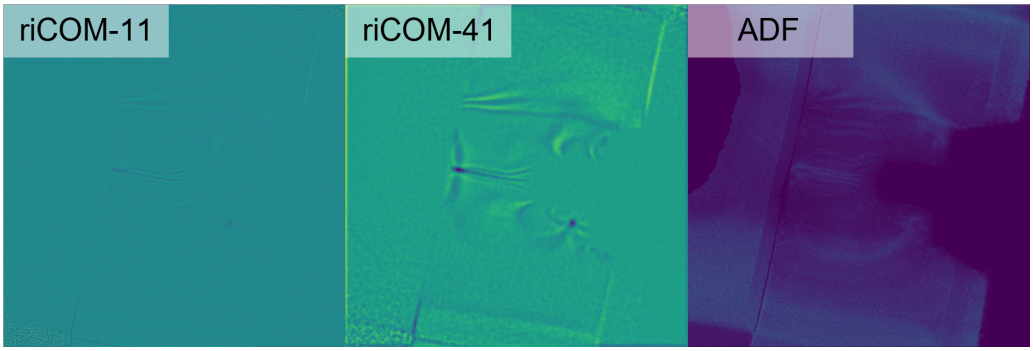


Figure 3.10: Reconstruction results of a FIB lamella under low magnification, riCOM reconstruction with a kernel size of 11, 41, and ADF imaging with a virtual detector are used. The low-frequency contrast plays an important role in imaging samples of larger scale, and the shape of the lamella makes a more pronounced contrast by the reconstruction with a larger kernel size (riCOM-41). The chemical differences between the layers can hardly be noticed by any riCOM reconstruction, due to the insensitivity to scattering angle, compared to the strong contrast created in the ADF image.

Comparisons with results of other non-iterative reconstruction methods show that riCOM renders high-quality images on par with established methods, even at very low doses. The pros and cons of using different frequency components are discussed. Users can accordingly choose the most suitable designs of kernels, and run simultaneously other imaging forming methods, in order to reach the highest dose efficiency or extract the most amount of knowledge from the investigated sample in real-time.

It is also worth discussing some limits of the method. Although already mentioned several times in the course of this thesis, the incapability to distinguish electrons at high and low scattering angles is largely damaging the interpretability of the COM-related methods, especially since the DF electrons are treated in the same way as BF electrons. This gives the (r)iCOM imaging a major drawback in that the contrast has to be built up from intensity shift, and the local scattering strength reflected by the percentage of DF electrons does not directly result in contrast. In other words, unless the transition region from a weak scatterer to a strong scatterer is well sampled, otherwise the two types of materials may result in the same contrast. This drawback can be even more fatal when a small kernel size is used since the low-frequency components are lost due to the transfer function, as shown in figure 3.10.

Chapter 4

Phase Object Reconstruction for 4D-STEM using Deep Learning

This chapter is based on:
*Friedrich, T., **Yu, C.P.**, Verbeeck, J. and Van Aert, S., 2023. Phase Object Reconstruction for 4D-STEM using Deep Learning. Microscopy and Microanalysis, 29(1), pp.395-407.*

The author has major contribution to the conceptualization of the paper, conducting TEM experiments, and designing the workflow for data processing and imaging.

4.1 Introduction

In the last chapter, the speedy update of the reconstruction result of riCOM, as well as its low-dose imaging ability and sensitivity to light elements were demonstrated. The limitation of the riCOM reconstruction, on the other hand, was also explained and manifested with experimental data. Both the advantages and the disadvantages, however, originate in the fact that the waveform is so much simplified that only the intensity shift is utilized for the final reconstruction, which on the one hand allows a high processing frame rate, but on the other limits the usage of information that exists in the acquired data.

The benefits of using the whole CBED patterns for reconstruction, instead of its derivatives, have been demonstrated by the many ptychography examples in section 2.2. However, the nature of the algorithm, which is built upon iterative searching, makes it almost impossible to provide live updates of the reconstruction result during data acquisition. There are two major obstacles to realizing real-time imaging with iterative algorithms: First, waveform retrieval is a computationally intensive task, and therefore the convergence of the search usually takes a longer time than what can be tolerated to keep up with the data acquisition. The necessary time consumption can be magnitudes longer than the time to finish the raster scan, making the synchronization between the two essentially impossible. Second, being iterative means the reuse of data, and thus the acquired CBEDs cannot be processed and dumped, as in the case of riCOM. Instead, data storage to accommodate all of the CBEDs belonging to the same dataset is necessary, so that the data can be pulled repeatedly in the iterative process and improve the reconstruction

result. This demand on memory can be exceptionally large when the amount of probe positions in the dataset grows.

The two obstacles, nonetheless, can be solved if the iterative process can be replaced by something quick and direct. In this work, we thus explore the possibility to use machine learning (ML) to perform waveform retrieval for phase object reconstruction. We show that using a convolutional neural network (CNN) enables fast exit wave retrieval for a given CBED, by using only a 3x3 kernel of adjacent diffraction patterns at a time. Using only nine CBEDs per probe position in a 4D-STEM dataset implies that the dataset can practically be of arbitrary size and the reconstruction can be performed live during the experiment with appropriate accelerator hardware, such as a modern, single GPU. The method allows the retrieval of exit waves with a resolution theoretically only limited by the Nyquist frequency of the detector and thus enables super-resolution imaging at sufficiently high doses. Besides, we also show excellent low-dose imaging capability and easily interpretable contrast from the proposed reconstruction method. In this paper, the character and capability of the proposed method are discussed in detail and demonstrated on both simulated and experimental data.

Comparisons are also made with other possible live processing methods. To the best of our knowledge, the only reconstruction methods that go beyond utilizing the traditional annular detector setting and enabling live imaging are single sideband ptychography (SSB)[80] and integrated differential phase contrast (iDPC) or iCOM [36], which is why we focus our analysis on comparing the results of our proposed method to these two methods.

4.2 Theoretical Framework

The interaction of fast electrons with thin specimens can be conveniently described with the phase object approximation (POA). As an electron passes through an electrostatic potential landscape, its wavelength λ is temporarily altered, resulting in a phase shift of the passing electron wave [43]. With the application of the POA, the specimen is assumed to be thin enough so that the propagation of the wave as it goes through the material can be neglected, and the acquired phase shift $\phi(\vec{r})$ is directly proportional to the projected potential of the sample, as described by equation 2.1, and the resulting exit wave $\psi_e(\vec{r})$ by equation 2.5. We further define a transmission function $T(\vec{r})$ that relates the incident wave $\psi_i(\vec{r})$ to $\psi_e(\vec{r})$:

$$\begin{aligned} \psi_e(\vec{r}) &= \psi_i(\vec{r})T(\vec{r}) \\ T(\vec{r}) &= \exp\left(i\phi(\vec{r})\right) \end{aligned} \tag{4.1}$$

According to equation 4.1 a direct solution of the transmission function is only possible if both incident and exit wave are known, while in practice neither of them are known a priori. The incident wave $\psi_i(\vec{r})$, can be fairly well approximated by the inverse Fourier transform of the product of the aperture function $A(\vec{k})$ and a complex aberration function $\exp\left(i\chi(\vec{k})\right)$ [91].

$$\psi_i(\vec{r}) = \text{IFT}\left[A(\vec{k})\exp\left(i\chi(\vec{k})\right)\right] \tag{4.2}$$

where $A(\vec{k})$ and $\chi(\vec{k})$ is the aperture image that can be found on the diffraction plane ($\vec{k}$) and the phase profile of the aberration, respectively, and $\vec{r}$ describes a 2D space at the object plane. The function $\chi(\vec{k})$, considering only the spherical aberration C_s and defocus Δf, is given by:

$$\chi(\vec{k}) = \pi\lambda k^2\left(0.5C_s\lambda^2k^2 - \Delta f\right) \tag{4.3}$$

Assuming that at least the low order aberration parameters of $\chi(\vec{k})$ are known, equations 4.2 and 4.3 can be used to estimate $\psi_i(\vec{k})$. The other piece of missing information is then the exit wave $\psi_e(\vec{r})$. From a 4D-STEM experiment, only a measurement of the intensity of the wave $|\psi_e(\vec{k})|^2$ can be made (figure 4.1(2)), and thus a method to retrieve the exit waveform on the sample plane based on the information accessible from the experiment is required to solve equation 4.1 to get the object transmission function $T(\vec{r})$.

Complex object wave retrieval based on known intensity observations is a common inverse problem, but in the case of a 4D-STEM experiment, it is much more complicated. Shot noise of the measured intensity makes even the estimation of $|\psi_e(\vec{k})|$ a challenging task on its own, and the actual interaction between the fast electron and material could severely change the outcome predicted by the phase object approximation.

The idea in this study is to leverage the multislice formalism, incorporating the calculation of electrostatic atomic potentials [92] and the frozen phonon approximation [93], as a forward model to generate a large synthetic dataset. This dataset can then be used to train a convolutional neural network (CNN) to retrieve an estimate of ψ_e for any given experiment within the boundaries of the validity of the used forward model and within the parameter space of the training data, which will be discussed in section 4.3 and is given concisely in table 4.1.

4.3 Convoluted Neural Network

Workflow

The general concept of the proposed reconstruction method is schematically illustrated in figure 4.1. The workflow to retrieve the transmission function $T(\vec{r})$ (figure 4.1(1)) of a 4D-STEM dataset (figure 4.1(2)) can be divided into two main steps: firstly, a neural network, trained to solve the inverse problem as outlined in section 4.3, retrieves the phase ϕ_e and amplitude $|\psi_e|$ of the exit wave, based on the intensity measurements of a 3x3 kernel of adjacent diffraction patterns, as depicted in figure 4.1(3-5). Secondly, a patch of the object is reconstructed from the previously retrieved exit wave according to equation 4.1, as shown in figure 4.1(6).

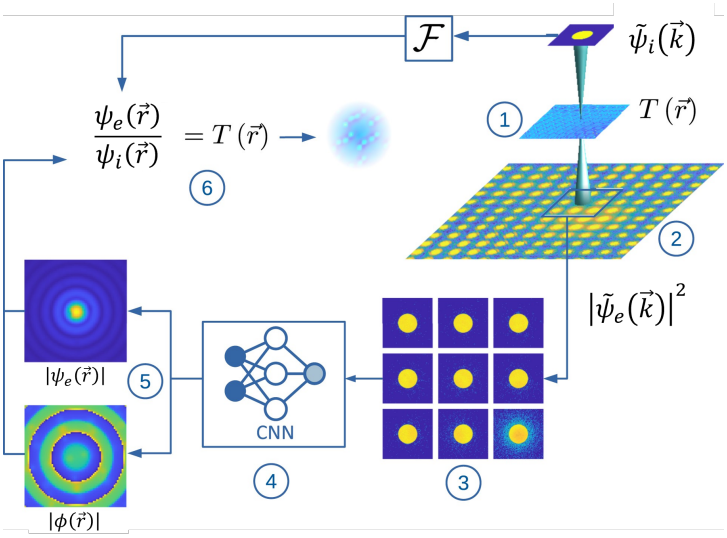


Figure 4.1: General workflow: A patch of the phase object (1) of a 4D-STEM dataset (2) is reconstructed by extracting a 3x3 kernel of adjacent CBEDs (3), using a CNN (4) to reconstruct the amplitude ($|\psi_e(\vec{r})|$) and phase ($\phi_e(\vec{r})$) of the exit wave of the central CBED (5) and using the phase object approximation to reconstruct the object patch (6) from the reconstructed exit wave and an estimated probe function $\psi_i(\vec{r})$. Patches are then stitched together by complex addition to yield a reconstruction of the full phase object.

The mismatch between the pixel size of the reconstructed object and the scan step needs to be taken into account before the complex addition can be applied. This is due to the fact that the object patches retrieved from different probe positions are always shifted by integer multiples of the step size in both directions, which is not guaranteed to be compatible with the pixel size of the object that is determined by the maximal scattering angle and pixel density of the detector. Therefore, a phase ramp is included in the approximated $\psi_i(\vec{k})$ and the predicted $\psi_e(\vec{k})$ to make the necessary sub-pixel shift of their real space counterpart $\psi_i(\vec{r})$ and $\psi_e(\vec{r})$, so that they can be found at the right position. This ramp $\varphi_{shift}(\vec{k})$ is applied in the $\vec{k}$-space that causes a shift in the $\vec{r}$ space, and is shown to be a function of the magnitude of the shift (Δx and Δy) and the dimension of the array ($N_x \times N_y$) that describes an object patch:

$$\varphi_{shift}(\vec{k}) = \exp\left[2\pi i\left(\frac{\Delta x}{N_x} + \frac{\Delta y}{N_y}\right)\vec{k}\right] \tag{4.4}$$

The real space wave functions considering the phase ramp are:

$$\begin{aligned}\psi_{i,n}(\vec{r}) &= \mathrm{IFT}\left\{\psi_i(\vec{k})\varphi_{shift,n}(\vec{k})\right\}\\ \psi_{e,n}(\vec{r}) &= \mathrm{IFT}\left\{\psi_e(\vec{k})\varphi_{shift,n}(\vec{k})\right\}\end{aligned} \tag{4.5}$$

In equation 4.5, the magnitude of displacements Δx and Δy are implied by n, which specifies the n^{th} probe position.

The phase object approximation assumes that the retrieved object patch should have a homogeneous amplitude of 1 with a phase distribution reflecting the projected potential of the imaged material. However, since the sample region traversed by the most intense part of the electron probe makes the most influence on the waveform of the transmitted electron wave, the retrieved object patches are given a weighting function ω_n according to the n^{th} incident probe intensity distribution, and the accordingly weighted object patch $\psi_{patch,n}$ is expressed as:

$$T_{patch,n}(\vec{r}) = \frac{\psi_{out,n}(\vec{r})/\psi_{in,n}(\vec{r})}{|\psi_{out,n}(\vec{r})/\psi_{in,n}(\vec{r})|} \times \omega_n(\vec{r}) \tag{4.6}$$

with the weighting function ω_n as:

$$\omega_n(\vec{r}) = \begin{cases} \frac{|\psi_{in,n}(\vec{r})|^2}{\sum_{\vec{r}} |\psi_{in,n}(\vec{r})|^2} & \text{if} \quad |\psi_{in,n}(\vec{r})|^2 > \frac{1}{10} max(|\psi_{in,n}(\vec{r})|^2) \\ 0 & \text{if} \quad otherwise \end{cases} \tag{4.7}$$

This procedure is repeated for all real space coordinates in the 4D-STEM dataset and the object patches are combined into the final phase object T by complex addition over n probe positions.

$$T(\vec{r}) = \sum_n T_{patch,n}(\vec{r}) \tag{4.8}$$

The object patch estimations coming from the CNN are not perfect but carry some errors. However, the full object is the combined result of the predictions made at multiple probe positions (equation 4.8). Even if one particular prediction is very inaccurate, its impact on the final result is limited as long as it is outweighed by the contributions from neighboring probe positions, which is the case when a significant probe overlap is established by a dense scanning raster.

Training data generation

We created a large synthetic dataset, using atomic structures extracted from the materials project [94] database. Based on the unit cell definitions we created bulk specimens in one of the low-index zone axis orientations given in table 4.1. Each sample consists of a 3x3 kernel of simulated CBED patterns as features and the corresponding exit wave in r-space as the label. The simulations were performed using the multislice formalism and the frozen phonon approximation. In the implementation given by [41, 42], following the derivation of [93] the CBED intensity can be expressed as the sum of coherent and incoherent intensities.

$$\left\langle \left|\psi_e(\vec{k})\right|^2 \right\rangle = \left|\left\langle \psi_e(\vec{k}) \right\rangle\right|^2 + \left\langle \left|\delta(\vec{k},t)\right|^2 \right\rangle \tag{4.9}$$

where $\psi_e(\vec{k})$ is the exit wave, δ is a phonon configuration(t)-dependent difference and $\langle\rangle$ denotes the average over t. This formalism gives us access to the average, coherent, complex wave function. We use this wave function of the central CBED of the 3x3 kernel as labels (i.e. the ground truth training target) and the CBED intensities of all patterns in the kernel, as given in equation 4.9 as features (i.e. the CNN input). Only low order aberration parameters Δf and C_s of $\chi(\vec{k})$ (equation 4.3) are considered as they typically are the only remaining influence with a properly chosen aperture size. We assume that including these effects phenomenologically with a constant, small C_s and corresponding Scherzer defocus is sufficient. Temporal and spatial incoherence are also not taken into account. This reduces the parameter space considerably but also implies that the method is (for now) limited to aberration-corrected well-adjusted microscopes. Further, the dataset includes the CBED size in Å^{-1}, the aperture size and the acceleration voltage, which allows the computation of the probe function $\psi_i(\vec{k})$ within the data pre-processing pipeline during the training using equations 4.2 and 4.3. Also, the effect of finite electron dose is applied as a data augmentation step during the training, assuming only Poisson noise. To accommodate the possibility that there may be no specimen interacting with the beam, another augmentation step replaces the 3×3 CBEDs with the aperture image $A(\vec{k})$ with a 3% chance. An example of the resulting training sample inputs and labels is illustrated in figure 4.2 in the "input" and "ground truth" panels respectively. All parameters describing the dataset are summarized in table 4.1. Visualizations of the parameter distributions are shown in the supplementary information (figure S.1). The data generation code was published open source under `https://github.com/ThFriedrich/ap_data_generation`, as well as the training dataset used [95].

Description	Value
Acceleration voltage	∈ {30, 40, 50, 60, 80, 100, 120, 140, 160, 180, 200, 300}kV
Step size ($\vec{r}$)	(0.0167...0.33) Å
Convergence angle	5...30 mrad
Spherical aberration	0.001 mm
Defocus	Scherzer defocus
# Frozen phonons	30
Atom rmsd	0.08 Å
Orientation	∈ {[1 1 0], [0 1 1], [1 0 1], [0 0 1], [1 0 0], [0 1 0], [1 1 1]}
Thickness	<30 Å
Dose	3...3e9 e/CBED
# Structures	126,335
# Samples	742,688

Table 4.1: Simulation parameters for the training dataset.

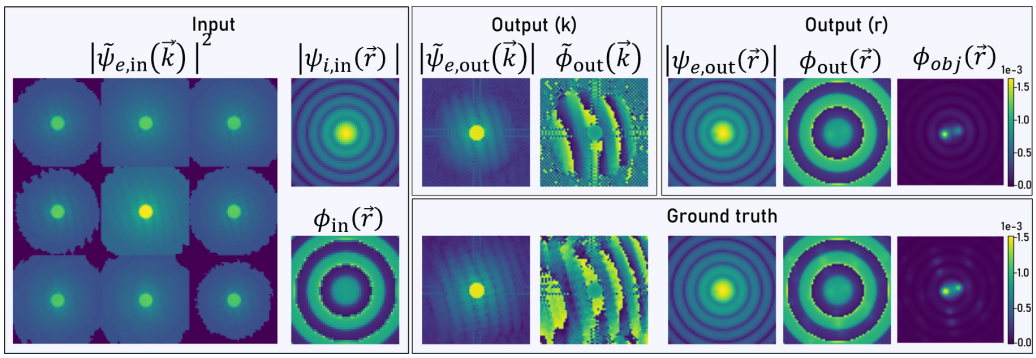


Figure 4.2: Example of an exit wave retrieval taken from the validation dataset, illustrating the inputs and outputs of the CNN, as well as the Fourier transforms of the (real space) exit waves. Intensities and amplitudes are depicted in log scale.

Neural network implementation

CNN architecture

The complexity of the inverse problem practically dictates the use of deep neural networks in this study. The U-Net architecture [96] is one of the most popular choices for deep learning applications, since it allows to expand the number of parameters considerably while maintaining strong backpropagation. Each depth level in the "U-shape" reduces the filter map size which accommodates the retrieval of features of different sizes. This makes the U-NET an extremely versatile and easy-to-train CNN and hence suitable as a starting point for this project as well [97]. However, since the training target is naturally complex-valued, a generic pixel-to-pixel mapping, as commonly employed for CNNs in image processing, cannot be used. Two popular ways of dealing with this exist. Firstly, both, phase and amplitude retrieval problems can essentially be treated separately by defining two outputs and optimizing for dedicated loss functions on the phase and amplitude components of the wave function. This treats the complex wave as two real-valued images, which in practice has the advantage that common, highly optimized AI tools can be readily employed. Another approach that naturally lends itself to this kind of problem is using complex-valued neural networks [98]; an approach that has found relatively few applications so far. However, the theoretical framework for complex CNNs is established [98] and implementations have been showing promising results [99, 100]. We implemented the U-NET architecture for both CNN types to test the main ideas. The complex networks delivered reconstruction results and accuracies of predicted phases, which are, for all intents and purposes, equivalent to the real-valued CNN, while decreasing the speed performance considerably. Since live imaging is an envisioned application, inference speed is a critical concern and the faster real-valued CNN was chosen in the study accordingly.

The structure of the U-NET is modified to account for physical considerations. The aim of the neural network is to model the electron-specimen interaction. Adding skip connections from the input probe function to the output exit waves (essentially enabling

global residual learning) isolates the specimen interaction contributions from the probe function contributions to the exit wave. The CNN, therefore, does not need to learn to actually model the electron probe. The skip connections have the additional benefit of providing a common template during inference. The training is done on isolated scan points of only 9 CBEDs per specimen, while during inference the probe function should be consistent for the entire dataset, which is a requirement that cannot be captured by any metric during the training. Providing an estimated probe function greatly promotes this consistency during inference. Global residual learning also enhances noise robustness because the probe function serves as the template, which is hardly altered if the input is merely noise.

On the input, the dynamic range of the CBEDs is being scaled by taking them to the power of 0.1 in a preprocessing step, which puts more relative weight on the dark field scattered electrons to support exit wave reconstructions beyond the convergence angle. Subsequently, each pattern is scaled, according to equation 4.10 depending on its distance from the central beam position, where ζ_{xy} and ζ_d correspond to the CBED weights adjacent to the central CBED along x and y and on the diagonal, respectively. This is a straightforward way to include the step size Δs into the workflow.

$$\zeta_{xy} = \frac{1}{\Delta s * 50} \qquad \zeta_d = \frac{1}{\sqrt{2 * \Delta s^2} * 50} \tag{4.10}$$

The constant factor of 50 accounts for the range of step sizes in the training datasets such that all ζ are between 0 and 1. The effect of this weighting can be seen in the "Input" panel of figure 4.2, evident by the higher mean intensity of the central CBED.

Other notable differences as compared to the original U-NET implementation[96] are different map sizes, the use of Swish-Activation functions [101] and the use of strided convolutional layers instead of max-pooling layers to avoid any loss of information when feature map sizes are reduced. Trainable scaling factors on the two output layers for the phase and amplitude were added to scale between the batch-normalized feature maps (with standard deviations of 1 and means of zero) in the CNN and the relatively small target value distributions, which correspond to the difference between exit waves and probe functions. The variables were initialized accordingly with small values of 1e-4 and optimized during training. The amplitude output layer includes a regularization that penalizes integrated exit wave intensities > 1. The phase output layer penalizes values larger than π and smaller than $-\pi$. These penalties are added to the loss during training. The resulting overall architecture is depicted in figure 4.3. The tensorflow implementations of the models and individual layers are available on GitHub at `https://github.com/ThFriedrich/airpi`. During the training the CNN can process >750,000 sample points in about 6 minutes on a single Nvidia RTX 3080 GPU, indicating that the model itself may be well suited for live processing at rates >2 kHz, if the pre- and post-processing pipelines are well-optimized.

Loss function

To facilitate the learning of a general representation for both the phase and amplitude we designed a multi-objective loss function as a linear combination of $\mathcal{L}_2(\vec{k}/\vec{r})$-losses on the

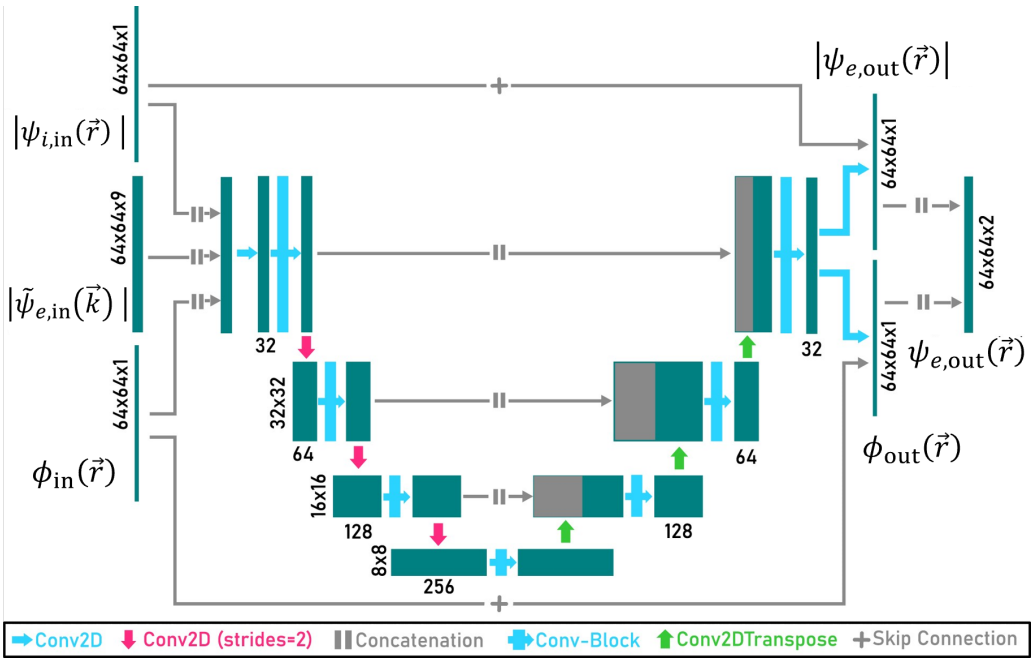


Figure 4.3: The CNN architecture used in this study is a modified U-NET with separate, real-valued phase and amplitude outputs. The inputs are the estimated probe function $|\psi_{i,\text{in}}(\vec{r})|$, the recorded CBED patterns $|\tilde{\psi}_{e,\text{in}}(\vec{k})|$, and the approximated aberration phase profile $\phi_i(\vec{r})$. The predicted output are the amplitude of the central CBED $|\psi_{e,\text{out}}(\vec{r})|$, the phase profile of the central CBED $\phi_{\text{out}}(\vec{r})$ The model leverages global residual learning through added skip connections of the probe function to the output. Each "convolution layer" is composed of a 2D convolution layer, batch normalization and a swish activation function. Each "convolution block" consists of 3 consecutive convolution layers.

phase and the amplitude in Fourier space and real space. Enforcing the correspondence between $\vec{r}$ and $\vec{k}$-space during the training encourages the CNN to abide by physical constraints. It was also observed that the decomposition of the phase into its sin and cos components facilitates better convergence, compared to just optimizing for the phase directly. This is presumably related to the decompositions being smooth so the CNN does not have to account for phase wrapping effects. Since the probe function is an input to the CNN the object can be solved directly and used in the loss function too. The optimization of the $\mathcal{L}_2$ error of the phase of the object $T(\vec{r})$ directly promotes an agreement of the object with the transmission function, which is practically the most meaningful metric. However, a good quantitative match may be impossible to achieve in certain scenarios (e.g. very low dose). To promote at least a visual match, the object phase is also optimized for its cross-correlation xc as given in equation 4.14. Furthermore, it was empirically determined that a higher weight on the phase in $\vec{k}$-space leads to faster convergence and overall better results. Putting all terms into a sum, for an exit wave on the sample plane $\psi_e(\vec{r})$ with phase $\phi(\vec{r})$, and on the diffraction plane $\tilde{\psi}_e(\vec{k})$ with phase $\tilde{\phi}(\vec{k})$, and also an object phase $\phi_{obj}(\vec{r})$ the loss function is given by:

$$\begin{aligned}\mathcal{L} = & \\ & \mathcal{L}_2(|\tilde{\psi}_e(\vec{k})|^2) + (\mathcal{L}_2(\sin\tilde{\phi}(\vec{k})) + \mathcal{L}_2(\cos\tilde{\phi}(\vec{k}))) * 3 \\ & + \mathcal{L}_2(|\psi_e(\vec{r})|^2) + \mathcal{L}_2(\sin\phi(\vec{r})) + \mathcal{L}_2(\cos\phi(\vec{r})) \\ & + \mathcal{L}_2(\phi_{obj}(\vec{r})) + \mathcal{L}_{xc}(\phi_{obj}(\vec{r}))\end{aligned} \tag{4.11}$$

with:

$$\mathcal{L}_2(x) = \left(x_{true} - x_{predicted}\right)^2 \tag{4.12}$$

$$\mathcal{L}_{xc}(x) = \left(1 - xc(x_{true}, x_{predicted}\right)/2 \tag{4.13}$$

where:

$$xc(x,y) = \frac{\sum\limits_{i,j}[(y(i,j) - \overline{y})(x(i,j) - \overline{x})]}{\sqrt{\sum\limits_{i,j}[y(i,j) - \overline{y}]^2}\sqrt{\sum\limits_{i,j}[x(i,j) - \overline{x}]^2}} \tag{4.14}$$

where x and y correspond to pixel values at the locations (i, j) and $\overline{x}$ and $\overline{y}$ are the mean values respectively.

Training

The training was performed using the Adam optimizer with a learning rate of $5e-4$, a batch size of 256 and a momentum setting of 0.9. The learning rate was decreased by a factor of 0.5 when the validation loss did not decrease for 3 epochs. Convergence was reached after ≈50 epochs. After convergence, the training was resumed for another 10 epochs with a weighting factor of 10 applied to the $\mathcal{L}_2(\phi_{obj}(\vec{r}))$ term of equation 4.11, which leads to a further small decrease on the object error. This step does not alter the reconstruction results considerably but improves the quantitative match between reconstructed objects and transmission functions somewhat.

4.4 Experiments and Simulations

The demonstration of the reconstruction methods is performed on both experimental and simulated datasets. For the experiment, probe-corrected Thermo Fisher Titan (X-Ant-TEM) and Themis (Advan-TEM) are used, the former is equipped with a MerlinEM direct electron detector (based on Medipix3) and a custom-made Timepix3 detector for the latter. For the experimental dataset of Au crystal and $SrTiO_3$ focused ion beam lamella, which can be found in the online repository [86], the acceleration voltage is set

at 300 kV, the convergence angle of the electron beam 20 mrad, and the scanning step size 0.2 Å and 0.185 Å, respectively. The USY zeolite dataset, which can be found in [102], is collected at 200 keV, with 12 mrad convergence angle and 0.15 Å step size. For the simulations, the twisted bilayer graphene dataset is generated with acceleration voltage 200 kV, convergence angle 25 mrad, and scan step size 0.2 Å, for the twisted bilayer MoS_2 dataset acceleration voltage 300 kV, convergence angle 20 mrad, and scan step size 0.1 Å, and for the MgO dataset acceleration voltage 300 kV, convergence angle 20 mrad, and scan step size 0.05 Å. All of the simulated datasets are generated with the MULTEM software [41].

4.5 Results and Discussion

Super-resolution

The reconstruction of the proposed method is based on solving equation 4.1 for the object using the incident and the exit wave functions, and therefore the resolution of the method is not explicitly limited by neither the optical conditions of the imaging system nor the sampling density of the electron probe. By reciprocity, the object plane is sampled with a maximum resolution determined by the maximum scattering angle the detector can reach, or the highest angle at which the exit wave can be accurately retrieved, which can potentially result in higher resolution than permitted by the former two limitations.

This super-resolution granted by the knowledge of the exit wave at higher scattering angles is demonstrated by the reconstruction of a simulated dataset of a twisted bilayer graphene sample at infinite dose [103]. The result from the CNN reconstruction is shown in figure 4.4a, and compared with a SSB reconstruction in figure 4.4b.

To analyze the spatial frequency achieved by each method, the Fourier transformed (FT) images are presented as well (figure 4.4d, e). The circle in the FT SSB image indicates twice the range of the convergence angle (α), which is the upper limit of the spatial frequency of this reconstruction method[104], and therefore all the frequency components beyond are eliminated. The reconstruction of CNN successfully retrieves components beyond this limitation, reflected in the ability to distinguish atoms with very short spacing in between, as can be seen in the atom pairs profile (figure 4.4c)

The improved resolution capabilities of the method are also demonstrated on an experimental USY-zeolite dataset [102]. To increase the accuracy of the neural network prediction, the CBED at each probe position is replaced by a summation of CBEDs within a 5×5 box, while the reconstruction is done with a step size twice as large as the original data. This treatment of data is, of course, not equivalent to sampling with a larger step size with more dose per CBED, but considering the very small step size used in recording the raw data set (0.05 Å) and the much larger probe size (1.1 Å, by equation 4.16), more benefits due to the reduction of shot noise can be gained compared to the loss in fidelity due to forcing a incoherent sum of CBEDs. This repetition in data usage increases the effective dose in the dataset, as individual CBED now contains $\frac{25}{4}$-times more electrons and greatly increases the accuracy of the neural network prediction. The actual dose that inflicts damage while interacting with the material, on the

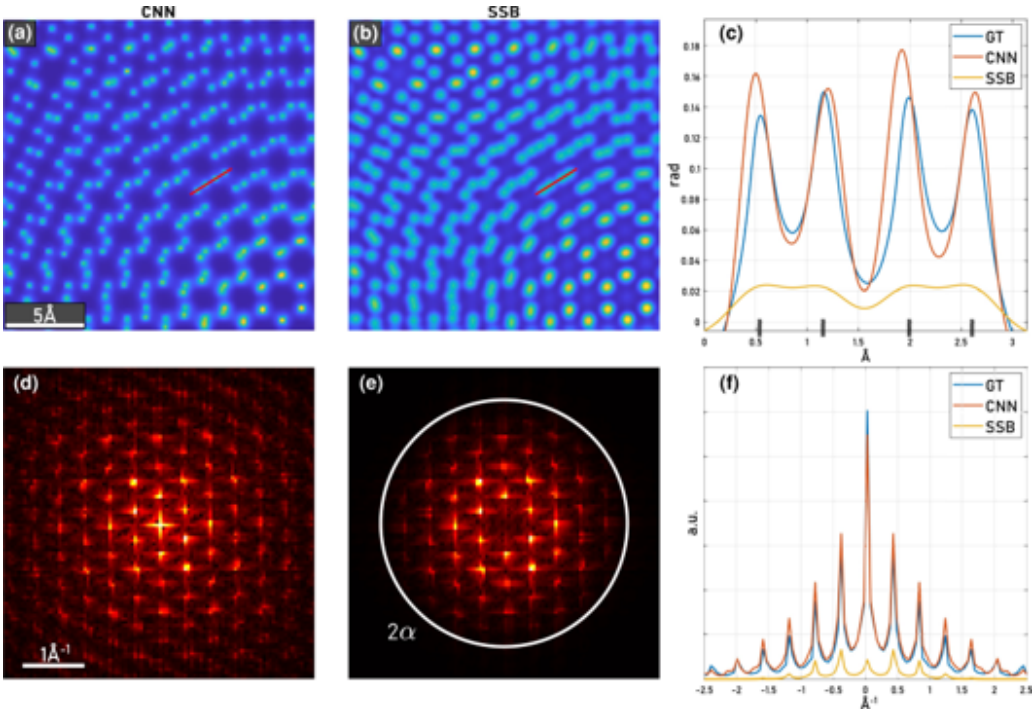


Figure 4.4: Demonstration of super-resolution capabilities on simulated datasets with infinite dose. Compared are CNN reconstructions and standard SSB ptychography (b). Their corresponding Fourier-transformed (FT) intensities show the maximal spatial frequency achieved by each method (d, e). For the FT image of the SSB result, a circle indicating twice the convergence angle is added, which corresponds to the maximal spatial frequency of the method. (c) shows the intensity along the line profile drawn in each image. Markers on the x-axis indicate atom positions. (f) depicts the integrated intensity of the FT images along the y-axis.

other hand, remains the same. For comparison, SSB is performed on the original dataset and a dataset with the same data repetition strategy applied. In figure 4.5, it is shown that the last three reconstructions successfully build a clear image of the zeolite crystal structure with atomic level resolution. The CNN reconstruction based on the original dataset does not showcase similar quality, since the dose for individual CBED is too low to make a meaningful prediction of the exit wave, but after data repetition is applied, the neural network gives results that capture details of the material. SSB, on the other hand, does neither benefit nor suffer from the repetition, at least not at a noticeable level. From the Fourier transform of the three images, one can estimate the resolution limits of the methods by comparing the most distant frequency component. The neural network reconstruction shows a maximal frequency component at 12.8 $Å^{-1}$, which according to the Raleigh criterion:

$$d = \frac{0.61\lambda}{\sin \alpha} \tag{4.15}$$

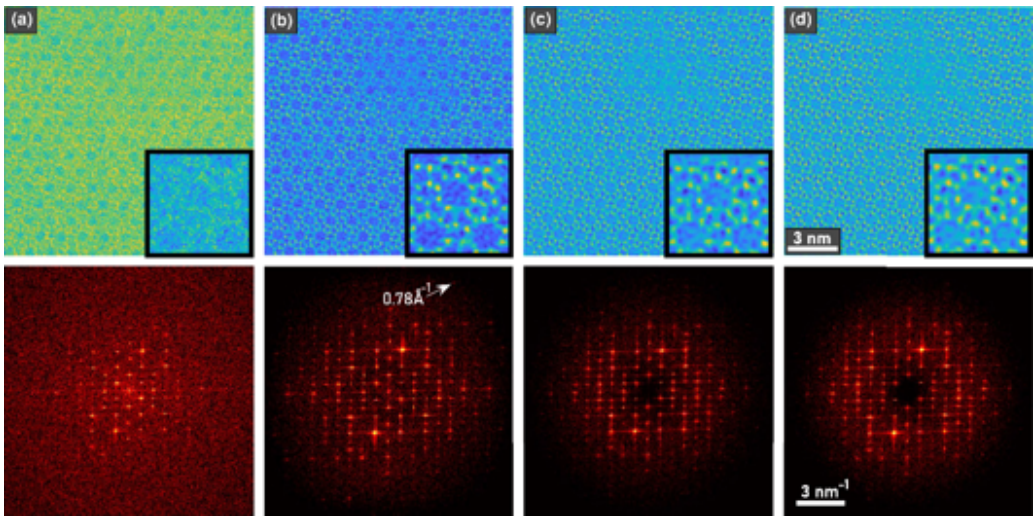


Figure 4.5: Reconstruction results of three different approaches. (a, b) Reconstruction performed on datasets without and with data repetition, respectively. (c, d) SSB reconstruction done on datasets without and with data repetition, respectively. The Fourier transforms of the reconstructed images are shown below. Notice that in (a) vertical streaks of very strong intensity can be found, which originate from an unknown defect of the detector, also reported by [82].

is equivalent to the resolving power of an un-aberrated perfect optical system of convergence angle 12 mrad, which is the same as the aperture size used in the experiment, at the electron wavelength (λ) of 0.02 Å. As most of the microscopes, even ones equipped with a probe corrector, cannot achieve the resolving power given by the Raleigh criterion, the purposed method shows the ability to overcome the effect of remaining aberration, shot noise, and other imperfection in the system to reach for higher resolution.

Step size

Since the proposed reconstruction method is based on retrieving individual object patches, which are commonly sampled finer than the step size of the scanning electron probe, a rather coarse scanning grid can produce high-quality images, as long as a good prediction of the exit wave, and hence the object patch, can be made. This character of the proposed reconstruction method is demonstrated on a simulated MgO particle with different step sizes of 0.1, 0.4, 0.8 and 1.6 Å. The "ratio" values shown in figure 4.6 refer to the ratio of the diameter of the incident probe function (1.2 Å) and the step size, where the probe size for a given convergence angle α in reciprocal Angstrom is defined by the first root of the Bessel function of the first kind and first order:

$$d = 2 \times \frac{3.8317}{\alpha\pi} \tag{4.16}$$

Figures in the left column are generated with infinite doses, and therefore the neural

network has very detailed knowledge of the amplitude of the exit wave to make accurate predictions. In this case, the difference between the results of an overlap ratio of 12 and an overlap ratio of 3 is barely noticeable. By further reducing the scan density, probe positions reach a distance where the weighting function forbids any overlap, as shown in equation 4.7. Despite the weighting function cutoff, which creates blank spaces between the object patches (see figure 4.6c,d), the actual probe positions in the simulation overlap with each other just enough to make exit wave predictions possible to maintain the crystal structure to a certain level in the reconstructed image. As the step size reaches 1.6 Å and the ratio drops below 1, the cropping resulting from the weighting function leaves large blanks (see figure 4.6a,b), and the retrieved object patches deteriorate severely, no longer reflecting any crystal periodicity. This failure shows that the neural network follows certain physical and mathematical constraints, such as necessary probe overlap for accurate exit wave retrieval, and that it would fail rather than make false predictions that continue to resemble atoms or the crystal. This failure can be identified by the user, not only based on the deviation of the resulting image from the expected appearance of the object but also by the wide blanks left between the object patches, indicating insufficient probe overlap.

The images in the right column of figure 4.6 were generated with the same dose per area. As mentioned in the previous section, the accuracy of the retrieved object patch is not directly related to the total dose in the dataset, but rather to the dose per CBED. By this consideration, it follows that larger step sizes work better for the neural network since this would mean fewer probe positions in the same area and a higher dose at every individual CBED. On the other hand, a certain level of probe overlap is also required for accurate predictions. Therefore, not only the total dose per area but also the scanning strategy is an important consideration for the proposed method. A balanced scan density will generate better results as compared to a very fine scan grid, even if the total dose per area would be the same. This behavior is illustrated in figure 4.6. The noise level is lowered significantly as the step/probe-width ratio drops from 12 to 3 in figures 4.6-h and 4.6-f.

The noise created by inaccurate prediction also creates different features as the step size changes. As the training is exclusively done on crystalline materials in zone axis orientations, the predicted object patches may show atomic-scale features, even if the input is merely noise. In other words, the frequency transfer function of each object patch is highest at the spatial frequency that would compose an image of an atom. This somewhat dangerous behavior of the neural network is compensated by the stitching of the patches, since atomic-scale features of the noise would not remain sharp as multiple object patches contribute to the same area, and would thus contribute to a cloudy, low-intensity background, as seen in figure 4.6-h. However, when the degree of overlap is reduced, such that the phase value is completely determined by one object patch, the risk to observe a false, atom-like feature, such as the ones seen in figures 4.6-f and 4.6-d, greatly increases. Awareness of this effect is therefore important when using the method and large step-size scanning patterns should be treated with extra caution. The other effect of noise is that the phase value drops in cases of very few electrons per CBED. This is essentially related to the failure of CNN at making accurate phase object prediction, but it also shows that as long as a sufficient dose is present in the 3×3 CBED input set, the phase value retrieved is not strongly related to step size.

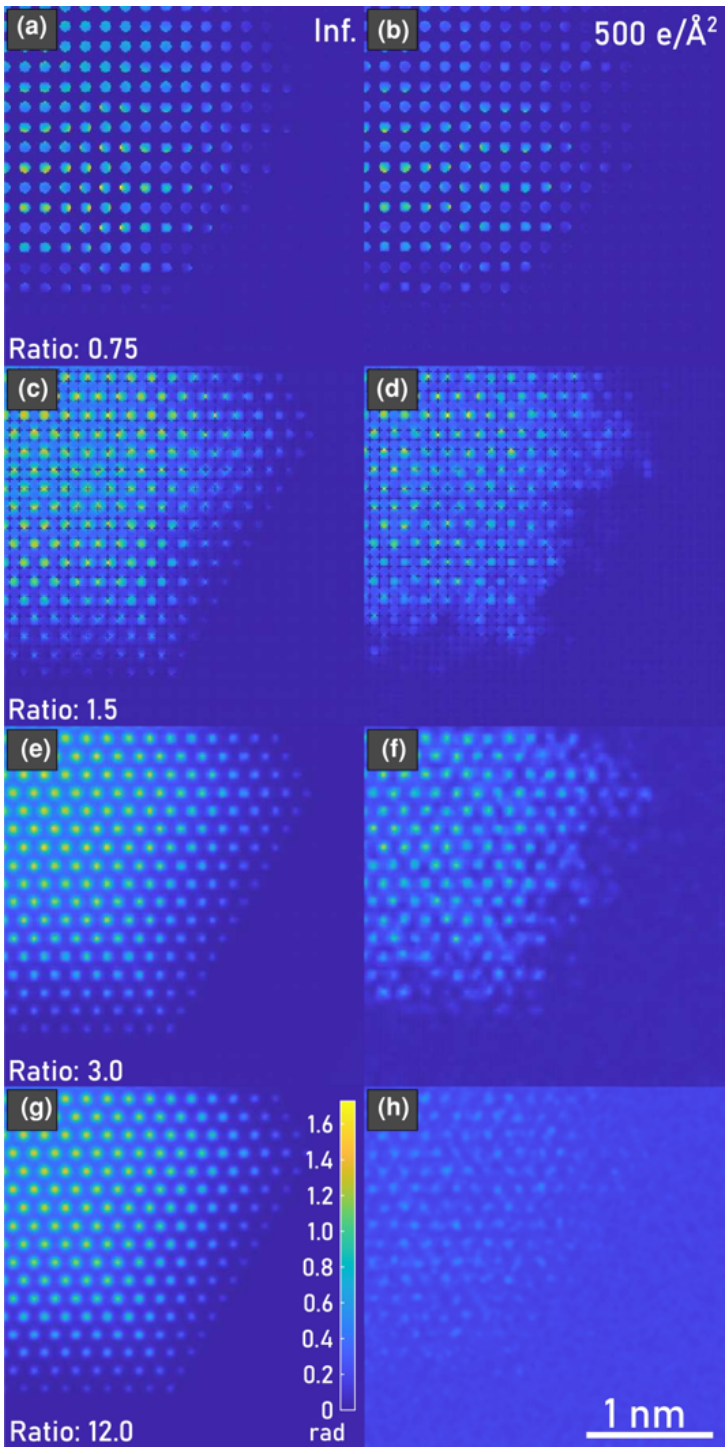


Figure 4.6: Reconstruction results of simulated MgO particle. In the left column (a, c, e, g) the images are generated with infinite dose, and in the right (b, d, f, h) the dose is set to be 500 electron per Å^2. The images in each row are built with the same step size, as well as the step/probe-width ratio. The colorbar in the bottom left panel applies to every image in the figure.

Contrast analysis

As outlined in section 4.2, the phase of the object is proportional to the electrostatic potential. By this relation, atomic species should, at least within the boundaries of the phase object approximation, be distinguishable. To verify whether this requirement holds true for the CNN reconstructions, 4D-STEM datasets of isolated, single atoms for each species in the periodic table up to Z=103 are simulated individually with a step size of 0.2 Å and a simulation box size of 3×3 Å. The retrieved phase objects of these datasets are compared to the ground truth transmission function, which is based on the parameterization by [105]. The comparison is carried out by taking averaged phase values from pixels within various ranges. In figure 4.7, from top to bottom, the curves show phase values at the peak only (figure 4.7-a), averaged phase over 3×3 pixels around the atomic position (figure 4.7-b), and the average over 5×5 pixels (4.7-c). Both, the curves of the ground truth transmission function and the CNN predictions, generally increase against atomic number, with the exception of certain dips at larger averaging ranges. This effect stems from different electron orbital distributions in the radial direction, and thus only the phase value averages at larger ranges are sensitive to this difference. The CNN predictions obviously also preserve these sub-atomic level details to some extent, as the shape of the curves bears a strong resemblance to the ground truth curves. Although the reconstructed phase values and the transmission functions do not match exactly, the predictions are accurate enough, such that the phase values of the reconstructed objects are indeed useful as an indicator for different atomic species, potentially even allowing semi-quantitatively predicting the exact atomic species.

For thick samples, the method is not expected to yield results in quantitative agreement with projected potentials, because even if the neural network would retrieve the correct exit wave, the reconstruction algorithm is still based on the POA and inherits its limitations. The analysis of thicker samples presented here is therefore done in a more qualitative/empirical manner and comparisons are made against ADF imaging and SSB. ADF images are well known for their strong contrast related to the scattering power of the imaged object and thus are suitable for examining the thickness variation of the sample [69] and local elemental compositions [23]. The contrast of SSB reconstructions is not as strong as ADF [106], yet the method is often used to study crystals containing elements of a wide range of atomic numbers due to its ability to image heavy and light atomic columns at the same time with distinguishable contrast [107]. Albeit a quantitative match can hardly be expected, it is important to verify whether reconstructed phase images still reflect the relative projected potentials of thicker samples to avoid misinterpretations. To that end, an experimental dataset of a $SrTiO_3$ FIB-lamella was analyzed. The reconstruction results are presented and compared to ADF and SSB images in figure 4.8.

For ADF imaging, the contrast difference between the Sr and O columns is too large, making it difficult to locate the O columns without the help of profiling. On the other hand, the SSB reconstruction does successfully image both atomic column types. While the peak intensities of the columns are ambiguous, they can still be distinguished by their corresponding size. The O columns are sharper than the Sr ones, indicating that an integrated signal from the area of each column could still be used as a reference of the local projected potential. The CNN reconstruction exhibits the advantage of both: while the light atom columns are observable, both the intensity and size differences are large enough to distinguish their types. This confirms that the Z-contrast sensitivity is

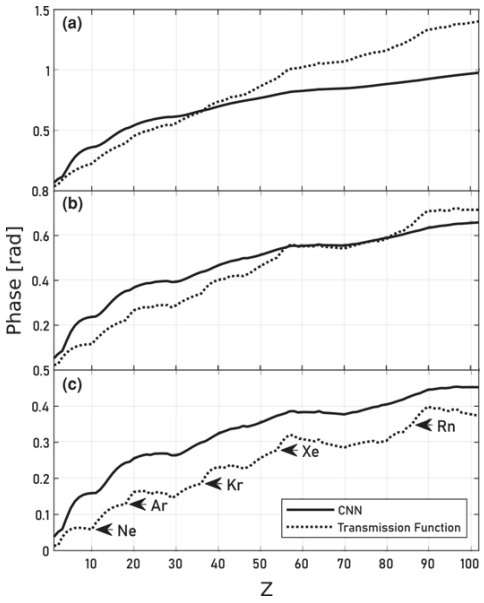


Figure 4.7: Phase response of the CNN compared to ground truth transmission functions for simulations of single atoms throughout the periodic table at infinite dose for (a) peak intensity, (b) mean over 3x3 pixels around the atomic position, and (c) mean over 5x5 pixels around the atomic position.

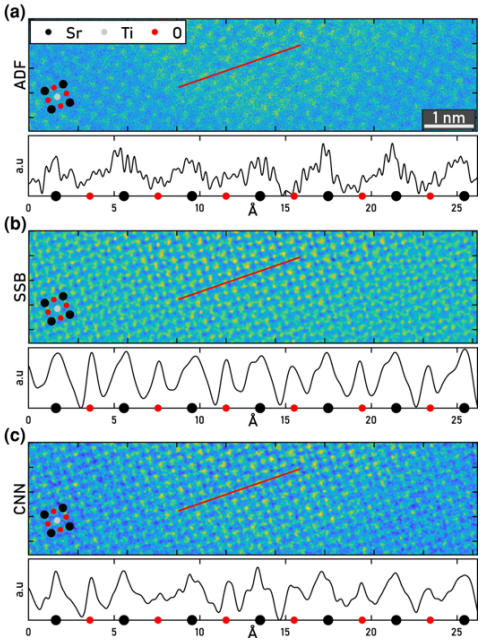


Figure 4.8: Phase response of the CNN compared to ground truth transmission functions for simulations of single atoms throughout the periodic table at infinite dose for (a) peak intensity, (b) mean over 3x3 pixels around the atomic position, and (c) mean over 5x5 pixels around the atomic position.

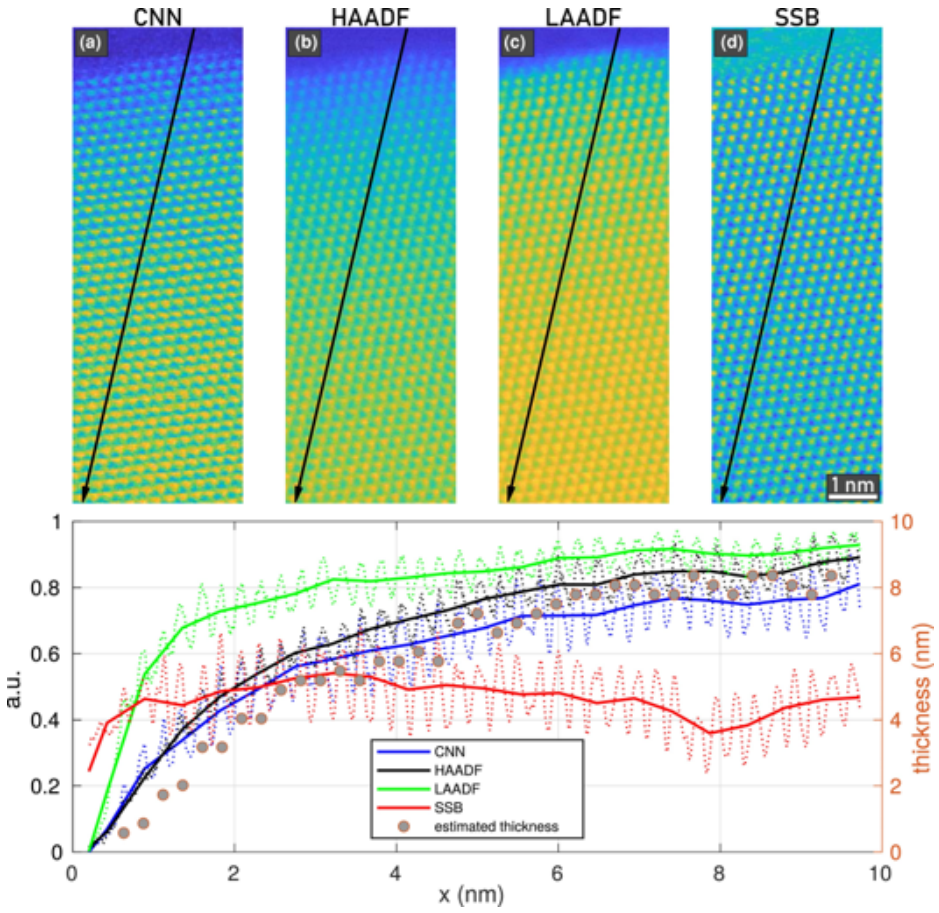


Figure 4.9: Reconstructed images of an edge of an Au crystal using f.l.t.r: CNN reconstruction, HAADF, LAADF and SSB. Line profiles across the nanorod illustrate the thickness dependence of the corresponding signals.

preserved for thicker specimens.

To further investigate the thickness dependence of the retrieved phase signal, an experimental dataset of the tip of an Au nanorod was used. As shown in figure 4.9, the intensity recorded by the virtual low angle ADF (LAADF) detector (20 mrad to 30 mrad) and HAADF detector (45 mrad and beyond) increases from the top to the bottom of the image. Based on the statistical analysis of the HAADF signal[69] to retrieve atom counts in each column, the thickest part in the image is about 9 nm. A line profile is then drawn for each imaging method presented in the figure. For the two ADF imaging methods at different collection angles, the profiles show a monotonic increase at a different pace against thickness, while the SSB profile only shows locations but the intensity is not correlated to the thickness of the atomic column. Compared to these profiles, the CNN reconstruction appears to be qualitatively most similar to the one of HAADF and correlates with the estimated thickness accordingly. It should be noted that the maximum thickness of ≈9 nm is well outside the parameter range of the training data (<3 nm). Also, the $SrTiO_3$-sample, being a FIB-lamella, should be well beyond 3 nm thickness.

This means that these examples also demonstrate the extrapolation capabilities of the CNN, which albeit being quantitatively arguably inaccurate, may still provide very useful reconstructions for imaging purposes. The strong resemblance with HAADF images at larger thicknesses may in fact be a very desirable characteristic, as it aligns well with many microscopists' experience and intuition.

By comparing the reconstructed images from the Au crystal and $SrTiO_3$, one would notice that SSB recovers contrast of higher spatial frequencies, such as the intensity and shape of atomic columns, but it does not recover long-range features induced by e.g. thickness variation. This is due to the band-pass nature of the method [87, 65]. Long-range features are built with low-frequency components, and thus for reconstruction methods that filter out, or cannot utilize signals that fall in the low-frequency end, these features are lost. For the CNN reconstruction, the object patches are also localized and no information beyond one probe position away is shared among the prediction of the exit wave. Therefore, the existence of long-range contrast variation can only be attributed to the good prediction accuracy of the CNN.

Noise robustness

The performance of the method under different dose conditions is demonstrated and analyzed on a simulated dataset of a twisted MoS_2 bilayer. The dose used for the reconstruction ranges from 500 to $10^5 e/\text{Å}^2$, and the dataset is processed by the proposed method, SSB, and iDPC. The methods reconstructions are illustrated in figure 4.10 and compared against the ground truth transmission function in terms of their normalized cross-correlations (equation 4.14).

It is evident that the method is strong in all three dose conditions, as can be seen visually in figure 4.10, and is confirmed by a cross-correlation higher than that of the corresponding SSB and iDPC reconstructions. The higher cross-correlation is not only due to a lower noise level but also the generally better-matched atom shape and phase value with respect to the ground truth, as indicated in the line profiles drawn under each dose condition (figures 4.10j-l). The sharper atom shape is due to the super-resolution that is achieved by accurate retrieval of DF waveform. This is confirmed by the better distinction of close atom pairs, as well as by the higher frequency component found in the Fourier transform of the image (figure 4.10). In figure 4.10-l, the atom profile of the CNN reconstruction at high dose almost strikes a perfect match with the phase of the transmission function. Note that the lines are not normalized or shifted, indicating that a very accurate wave retrieval is achieved by the CNN. Additionally, the contrast of the CNN results allows Mo and S_2 to be distinguished, which is more difficult with other methods, as their difference is much smaller. At a lower dose, the line profiles suggest better low-dose robustness may be found in iDPC, as the two peaks are preserved in the reconstructed image. However, the signal almost completely falls into the noise level, as confirmed by the Fourier transform, and thus this seemingly better low-dose performance could very likely be a coincidental noise distribution. The given example highlights the potential of the proposed method for low-dose imaging. As pointed out in sections 4.5 and 4.5 and illustrated in figures 4.5 and 4.6 respectively, the noise robustness may depend substantially on the step size and the effective dose per CBED. To gain an advantage over other methods in this regard this context needs to be taken into account

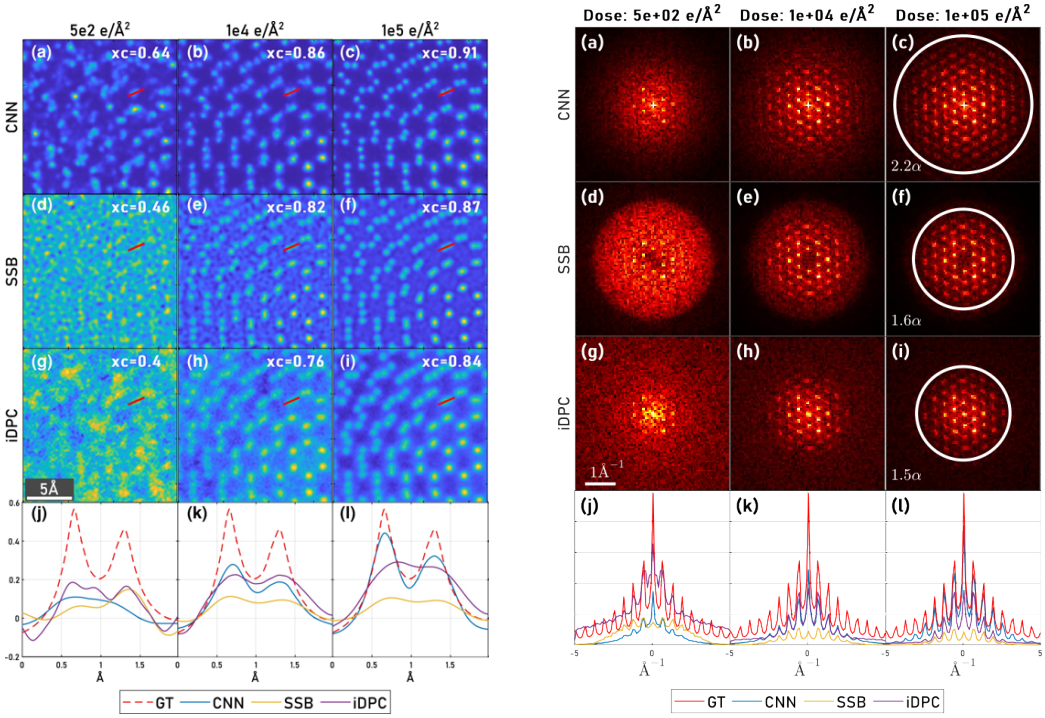


Figure 4.10: The dose dependency of the various reconstruction methods. The panels on the left show reconstruction results of a simulated twisted bilayer MoS_2 with the proposed methods (CNN), SSB, iDPC, and an intensity profile across two atoms indicated by the red line, from top to bottom respectively. The line profiles averaged in the perpendicular direction over 1.6 Å are drawn across a Mo-S_2 pair, and shown with the ground truth. cross-correlation values xc are given with respect to the ground truth phase object in the panel of each result image. IDPC values were normalized by the maximum value of the transmission function. The panels on the right show the Fourier transforms of the reconstructed object images. Profiles on the bottom show the integrated intensity of the Fourier transforms in the y direction.

and scanning strategies adapted accordingly.

4.6 Conclusion

This paper presents a new computational imaging method, leveraging a CNN to retrieve complex exit wave functions from CBEDs and an algorithm to reconstruct the phase object from the predictions of the neural network. Since the exit waves are retrieved for each real-space coordinate in a 4D-STEM dataset, based only on a small kernel of adjacent diffraction patterns, the method can be employed in a sequential manner, thus enabling live imaging during an experiment. The machine learning system is based on a well-established model but streamlined to the task at hand and adapted to account for physical constraints and considerations. The model was trained on a large synthetic

dataset of multislice simulations. Large and high-order aberrations, as well as CBED distortions, like non-centricity, geometric distortions and hot/dead pixels, are not considered in the training data. Therefore, experimental data may require a pre-processing step. The range of practical conditions for which the method works reliably is thus arguably limited accordingly to aberration-corrected, well-adjusted instruments at this stage. The trained model, code and training data are publicly available as summarized in the support information 4.7. In the discussion, multiple unique characteristics and advantages of the method are demonstrated. The CNN-based reconstruction was shown to enable higher resolutions than any other live-imaging-capable method considered, on simulated, as well as on experimental data, provided that a sufficiently high dose-per-CBED is maintained. In correspondence to this consideration, the effect of the step size is analyzed. While a better estimation of the exit wave is obtained with the electron dose-per-area distributed across fewer probe positions, some probe overlap is necessary to ensure the accuracy of the exit wave retrieval. Hence, the method is most suitably applied at a balanced scan density. If these considerations are taken into account the reconstruction method can be very dose efficient.

The Z-contrast was analyzed on single-atom simulations across the periodic table. The phase signal of the reconstructions could indeed be linked qualitatively to atomic properties and a semi-quantitative analysis of thin specimen within the limits of the POA, was shown to be possible. We confirmed the contrast sensitivity to atomic species and sample thickness on experimental datasets of a $SrTiO_3$ FIB-lamella and an Au nanorod respectively. The observed monotonic increase of the phase signal with thickness and nearly monotonic increase with the atomic number indicates that quantitative analyses based on the reconstruction results may be feasible.

Generally, we believe the proposed method presents an attractive imaging modality for its super-resolution capability, high noise robustness, and the feasibility of qualitative or even quantitative contrast analysis. While further studies would be necessary to obtain a more detailed view of the model performance over the entire parameter space (and beyond), we could already show that the method is robust for a wide range of practically meaningful applications, even exhibiting reasonably good extrapolation behavior well beyond the maximum sample thickness of the training data. The fact that none of the examples shown in this study exist in those exact configurations in the training data, further indicates that the system generalizes well within the parameter interpolation range as well.

4.7 Supporting Information

Code and Data

Trained Neural Network and Reconstruction implementations:

authors: Thomas Friedrich and Chu-Ping Yu
title: airpi
license: GNU GPL3
url: https://github.com/ThFriedrich/airpi

Training Datasets:

authors: Thomas Friedrich, Chu-Ping Yu, Jo Verbeeck and Sandra Van Aert
title: Phase Object Reconstruction for 4D-STEM using Deep Learning, (4D-STEM Training Data)
doi: 10.5281/zenodo.6971200
license: Creative Commons Attribution 4.0 International Public License
url: https://zenodo.org/record/6971200#.YwflDmxBxhE

Data generation code:

authors: Thomas Friedrich and Chu-Ping Yu
title: ap_data_generation
license: GNU GPL3
url: https://github.com/ThFriedrich/ap_data_generation

STO dataset:

authors: Chu-Ping Yu, Thomas Friedrich, Daen Jannis, Xie Xiaobin, Sandra Van Aert and Jo Verbeeck
title: Real Time Integration Center of Mass (riCOM) Reconstruction for 4D-STEM
doi: 10.5281/zenodo.6971200
license: Creative Commons Attribution 4.0 International Public License
url: https://doi.org/10.5281/zenodo.5572123

Zeolite dataset:

authors: Daen Jannis, Christoph Hofer, Chuang Gao, Xiaobin Xie, Armand Béché, Timothy J. Pennycook and Jo Verbeeck
title: Event driven 4D-STEM acquisition with a Timepix3 detector: microsecond dwelltime and faster scans for high precision and low dose applications
doi: 10.5281/zenodo.6971200
license: Creative Commons Attribution 4.0 International Public License
url: https://doi.org/10.5281/zenodo.5068510

Example Datasets:

authors: Thomas Friedrich, Chu-Ping Yu, Jo Verbeeck and Sandra van Aert
title: Phase Object Reconstruction for 4D-STEM using Deep Learning, (4D-STEM Example Data)
doi: 10.5281/zenodo.7034879
license: Creative Commons Attribution 4.0 International Public License
url: https://doi.org/10.5281/zenodo.7034879

Additional data and figures

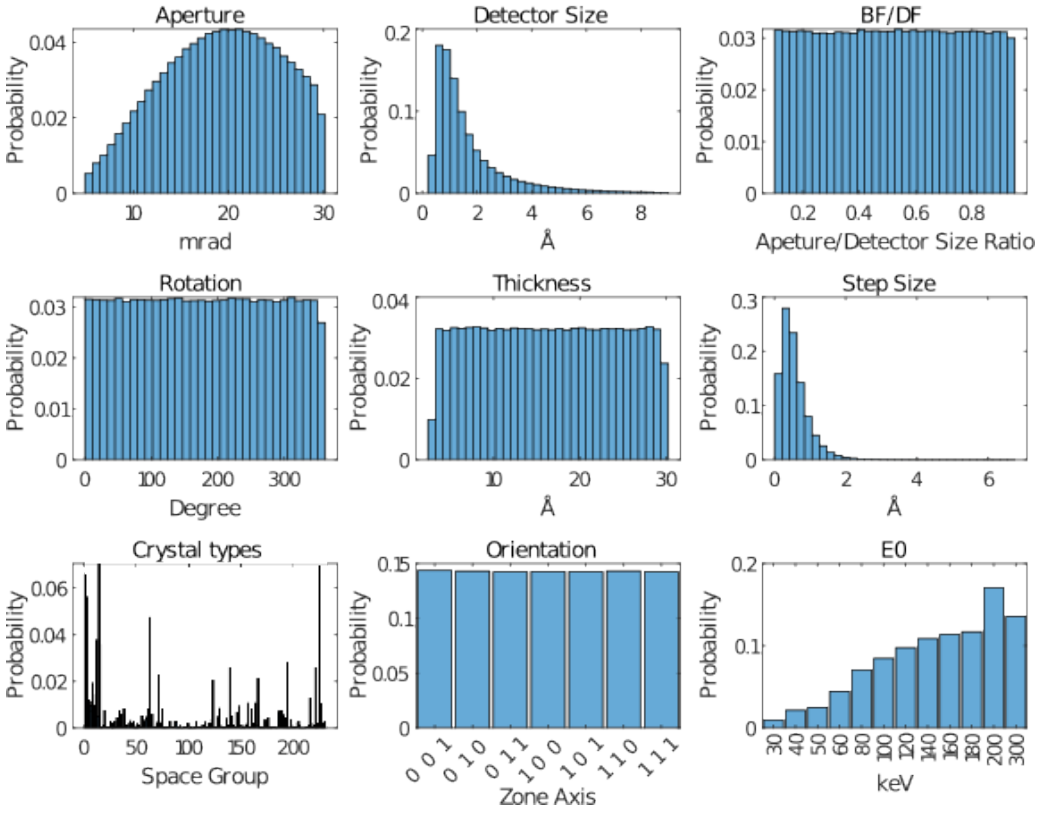


Figure 4.11: Parameter distributions of most important simulation parameters of the training dataset, consisting of 742,688 samples.

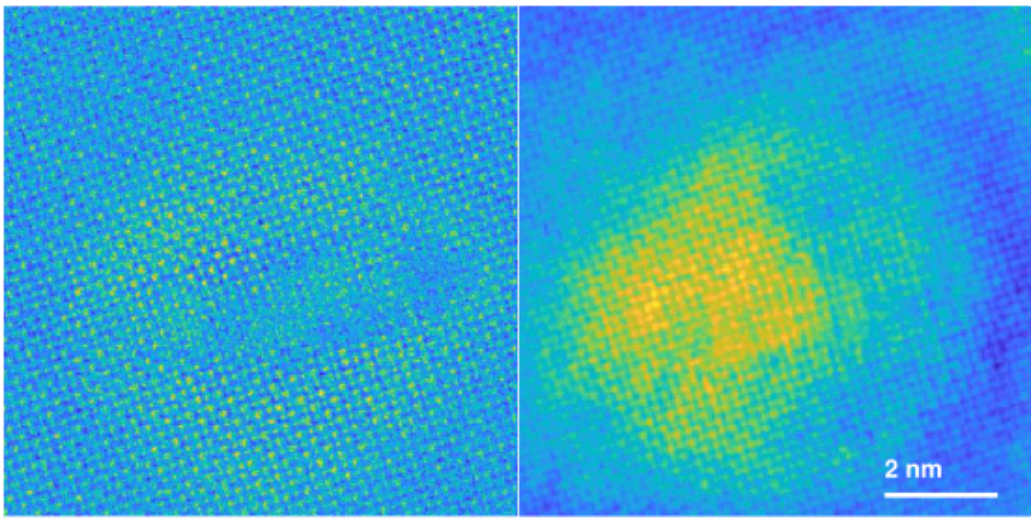


Figure 4.12: Comparison of the Neural network reconstruction (left) and iDPC (right) of a FIB lamella, including a hole in the center. This illustrates a reasonable tolerance of the proposed method towards thickness variations and bending.

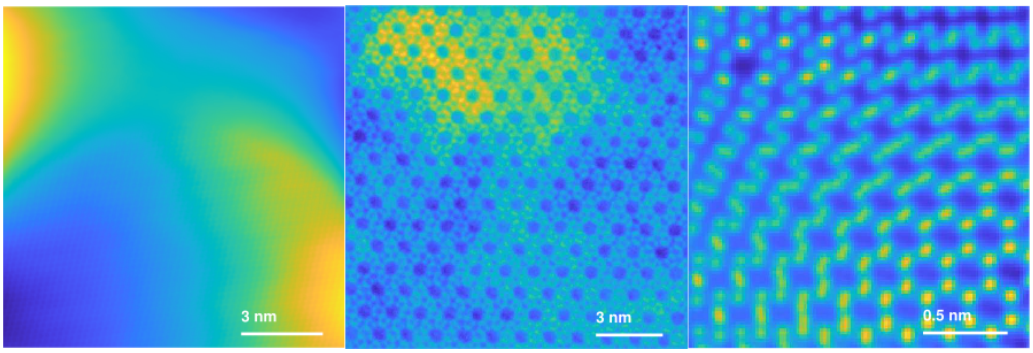


Figure 4.13: Additional iDPC reconstructions for comparison of samples presented in the paper, f.l.t.r: experimental Gold nanorod, experimental Zeolite, simulated twisted bilayer graphene with infinite dose.

Chapter 5

Quantum Wavefront Shaping with a 48-element Programmable Phase Plate for Electrons

This chapter is based on:
***Yu, C.P.**, Ibañez, F.V., Béché, A. and Verbeeck, J., 2023. Quantum Wavefront Shaping with a 48-element Programmable Phase Plate for Electrons. arXiv preprint arXiv:2308.16304.*

The author has major contribution to the conceptualization of the paper, characterizing the named device, and simulating/testing the named applications.

5.1 Introduction

Wavefront shaping, or the spatial and time-dependent control over the phase in coherent waves, has revolutionized many diverse scientific fields ranging from radio and light astronomy [108, 109], radar [110], acoustics [111, 112, 113], seismology [114], telecommunication [115, 116] and many more [117]. It requires a device that can apply a position-dependent phase change and can be augmented by adding a control loop to obtain adaptive optimization of the wave with respect to some goal function. In optics, this can be realized with so-called spatial light modulators, which can be based on moving arrays of mirrors or by liquid crystal-based setups that can change refractive index when applying an electric field [118, 119].

Matter waves, as introduced by de Broglie [120], are also amenable to this same concept. Indeed, the working principle of an electron microscope is entirely based on describing the free electrons as coherent quantum waves with wavelengths of the order of picometers. The capability of manipulating these electron waves is an indispensable part of a transmission electron microscope (TEM). The most relevant addition to phase manipulating devices in recent decades is, without a doubt, the spherical aberration corrector [121, 122, 123, 124], which flattens the phase front of the electron wave induced by (high-order) geometric aberrations of the microscope lenses and allows the forming of a sharper and more intense probe in scanning probe applications. Removing unwanted phase aberrations has significantly increased the resolution and current density

of the scanning transmission electron microscope (STEM) with many benefits in, e.g., spectroscopic applications.

Besides canceling geometric aberrations, the ability to arbitrarily shape the electron wavefront is gradually gaining attention with the hope of improving contrast or selectivity in electron microscopy setups. There has been a renewed surge of such phase modulators and their applications in the past few years. In soft material imaging, different phase plates such as Zernike [46, 125], Boersch[45, 126], Zach [127, 128], or Volta [129, 130, 131] have been implemented in the TEM to imprint a constant phase shift to a (central) part of the electron wave, to increase the contrast when imaging weak phase objects. Some other designs with relatively higher complexity may modify both the amplitude or phase configuration of the electron wave to create an electron probe of specific shape [132, 133], to increase contrast [35, 134], or to extract specific information from the electron-sample interaction [135, 136], to name a few. Some of these complex modulators even exhibit control over the parameters or magnitude of the modulation. The electrostatic phase plate reported by Verbeeck et al. [137] has demonstrated changes in interference between 4 partial waves by altering their mutual phase relation. Barwick and Batelaan [138] showed that a pulsed laser beam could induce a phase shift in the electron beam and that the contrast of the formed image can be optimized by tuning these laser pulses. Following this discovery, different realizations of using the ponderomotive force to change the phase of an electron beam appeared [139, 140, 141, 142]. The electrostatic phase plate reported by Tavabi et al. [143] has demonstrated a tuneable azimuthal phase by setting up specific electric field boundary conditions, which was interpreted as adding orbital angular momentum to the electron beam.

Here we report on an adaptive electrostatic phase plate based on the proof of principle demonstration by Verbeeck et al. [137], but with significantly increased complexity, performance, and practical usefulness. The phase plate consists of 48 openings, or pixels, transparent to an incoming coherent electron wave. The vertical walls of the pixels are made into electrodes so that an electric potential can be established inside, changing the wavelength of that part of the transmitted wave. Since independent voltage sources control each of the 48 pixels, the phase of the entire transmitting coherent electron wave can be programmed at will. This design and the electrostatic nature grant the phase plate several advantages, such as short response time, the ability to realize complex and arbitrary phase configurations, low power dissipation, compactness, low weight, and high stability and repeatability.

The experimental part of the paper provides a concise summary of the reported phase plate. The design of the phase plate is described first, as well as the components and mechanism to create a phase shift on an electron wavelet. The manufacturing design choices are briefly discussed in the scope of the challenges faced. The device's optical performance is then evaluated regarding its phase sensitivity and response time.

We discuss the applications of the phase plate in the scope of electron microscopy. Using the unique properties of a fast, hysteresis-free programmable phase plate, we demonstrate how novel imaging setups can expand or improve imaging modalities in TEM. We provide simulated examples and early experimental attempts towards electron wave modulation, complex sampling schemes, adaptive optics, and phase-coded ptychography to hint at what phase plates could bring to the electron microscopy community.

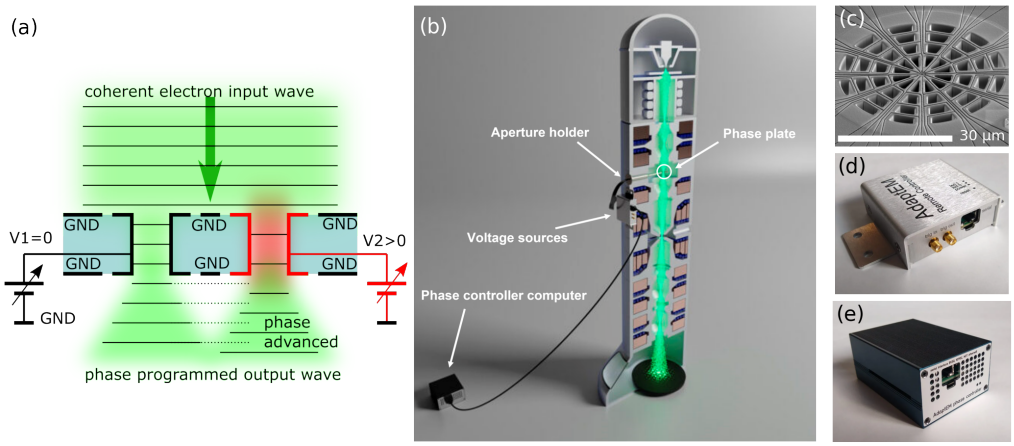


Figure 5.1: (a) Sketch of the working principle of the phase plate, where only 2 pixels are drawn. (b) 3D render of the setup and the main components, including (c) the phase plate, (d) the voltage sources, and (e) the phase controller computer.

5.2 Experimental Considerations

Description of the electrostatic phase plate

The basic working principle of the phase plate is sketched in 5.1-a. A coherent incoming electron wave is made to interact with an insulating membrane that has several holes. The top and bottom surfaces of the membrane are covered with a ground shield, while the inside of the holes is coated with a conductive layer that can be put to a controlled electrostatic potential (V_1 and V_2 in the simplified sketch). The potential surrounding the holes creates a potential landscape for the fast electrons that accelerates the electron upon entering and decelerates upon leaving this area. This will cause a phase change between the partial waves leaving these holes where one could imagine them as coherent Huygens sources that will constitute a now phase-programmed wave upon propagation in free space. The phase shift ϕ obtained is given by the electrostatic Aharanov-Bohm shift:

$$\phi = \frac{\pi e}{\lambda E_0} \int_\Gamma V(\vec{r}) dl \tag{5.1}$$

For an electron wave with wavelength λ and energy E_0 and crossing a region of space with an electrostatic potential $V(\vec{r})$ along a trajectory Γ. In the case of a weak perturbation, the electron's trajectory is not altered by this field, and the phase shift becomes directly related to the projected electrostatic potential. The goal of a pixelated phase plate is to create a potential profile that, in projection, leads to a constant phase shift proportional to the voltage applied to each pixel element. This occurs if the projected potential changes as little as possible over the surface of each hole, which can be obtained by choosing a high aspect ratio (height/diameter> 1).

From a practical perspective, the AdaptEM WaveCrafter phase plate [144] comprises three main elements shown in 5.1b-e: a dedicated condenser aperture holder containing the active phase plate chip, a 48-channel programmable voltage source and a remote computer for control and user interface, respectively. The phase plate used in this work is composed of 48 independent active elements, or pixels, arranged in 4 concentric rings and 12 petals (see 5.1b). Each element consists of a layered structure similar to the one described by Matsumoto and Tonomura for a single phase shifting element [145]. An aspect ratio of approximately 2 was chosen to avoid lensing effect, and a total diameter of the active area of 50 μm assures that a modern electron microscope can coherently illuminate the whole device.

One considerable advantage of this phase plate design lies in the relatively low voltage (in the mV range) required to induce a phase shift of 2π. This avoids high electric field breakdown issues in the nanoscale features of the chip and has the benefit that readily available voltage sources, which are simultaneously precise, stable, low power, fast and reliable, can be used.

Characterization

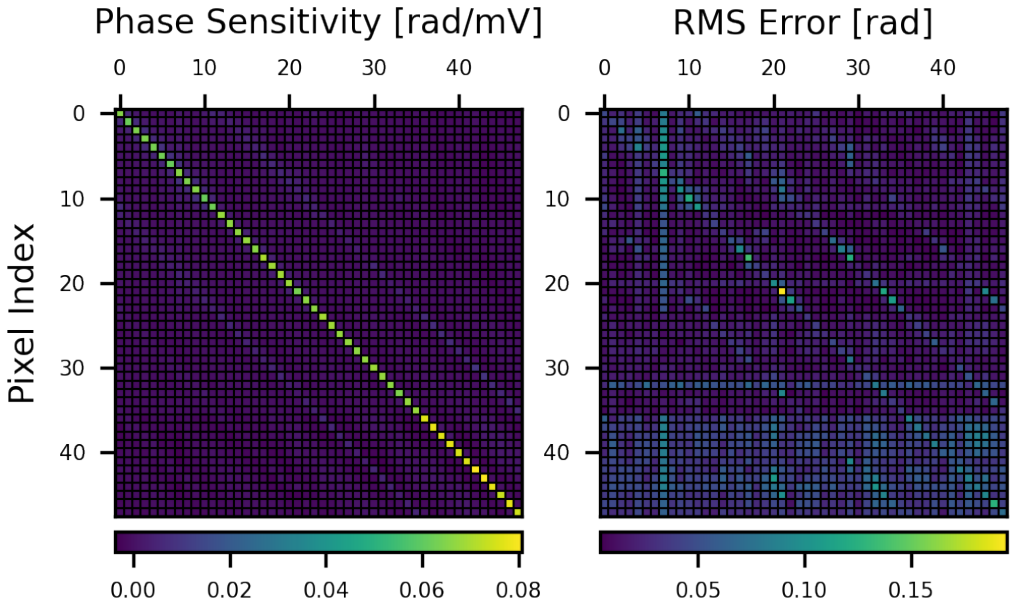


Figure 5.2: Phase sensitivity matrix and the corresponding root mean squared error of the linear fitting.

To experimentally examine the projected potential profile, phase reconstruction based on the Gerchberg-Saxton (GS) algorithm [58] was performed on a set of images of the phase plate, where each pixel is excited with increasing electrostatic potential (48 pixels, 11 voltage levels, 528 images in total). For the characterization, the phase plate is inserted in the sample plane of an FEI Tecnai Osiris S/TEM operating at 200 keV and illuminated with a parallel electron beam. The images are taken from the back focal plane of the objective

lens (diffraction mode), while the objective lens is largely defocused so that the detector can capture the near-field diffraction pattern of the phase plate. This experiment is aimed to characterize the projected potential on the phase plate when varying the phase inside each pixel in a range between 0 and 2π. A rough estimation of the voltage corresponding to a 2π phase shift was first found by assigning a gradually increasing voltage to half of the pixels randomly and repeatedly. Theoretically, a 2π phase shift should not result in any difference in the near-field pattern formed by the phase plate, and therefore the corresponding voltage can be found by slowly raising the voltage applied on random pixels until no major modifications to the pattern occurred. Once this voltage $V_{2\pi}$ was found, a series of images with 11 different potentials equally spread between 0 and $V_{2\pi}$ was taken for each pixel.

The defocused condition was specifically chosen so that outgoing waves from the electrodes interfered strongly with each other, and the phase difference between separate neighboring wavelets is significantly encoded in the recorded intensity images (see supplementary). This choice of detection plane was preferred over recording at an in-focus condition that interferes all of the wavelets together for several reasons. First of all, at the right focus, the transmitted electrons are concentrated in a very small region (less than 1 % of the size of the recorded defocused images), and creating a high enough camera length to sufficiently sample such patterns on a pixelated camera for phase retrieval is not trivial. On top of that, the inversion invariant nature of the wave intensity in the reciprocal space would also challenge obtaining a unique reconstruction and greatly hinder the retrieval algorithm's convergence.

The result of the reconstruction is summarized in 5.2. The phase response of all pixels, as they were individually excited, is fitted using a linear function, representing the phase sensitivity of that pixel to the applied voltage. A phase sensitivity matrix can be constructed showing the phase sensitivity of pixel i upon exciting pixel j. The phase sensitivity matrix in 5.2 shows a strong response on the diagonal, meaning that the excited pixel is the only one showing a significant linear phase shift against the applied voltage. An average phase sensitivity of 0.075 rad/mV is found, which translates to a theoretical phase resolution of approximately $3{\cdot}10^{-3}$ π according to the smallest step size provided by an ideal 16-bit DAC (maximum 2.5 V, smallest step 2.5×2^{-16} V). The error matrix, also shown in 5.2, indicates response deviation from the expected linear behavior, mainly resulting from imperfections in the phase retrieval process, such as the finite pixel size and non-ideal detector response. These can cause a difference between the recorded intensity and the actual waveform. The error is calculated by the root mean square error of the fitted result, which is found, at maximum, to be 3% of 2π (0.19 rad), while on average less than 0.5% of 2π (0.027 rad).

Besides the expected response of the phase plate, it is equally important to characterize any non-ideal behavior. The inhomogeneity describes the phase deviation within the pixel area from the ideal constant, homogeneous expectation. We evaluate the standard deviation of the reconstructed phase within each activated pixel and find it to be $< 1.7\%$ of 2π. The cross-talk refers to the phase response within a pixel region caused by the voltage applied to another pixel. We estimate this as the maximum linear response of a non-excited pixel as a function of any other excited pixel. The off-diagonal lines found exactly 12 pixels away from the main diagonal in both matrices in 5.2 indicate that the strongest cross-talk is, unsurprisingly, found between neighboring pixels due to how the pixels are ordered in the matrix (see supplementary). The cross-talk is measured

to be < 0.012 rad/mV, which amounts to 15% of the response of the excited pixel. In summary, the inhomogeneity only creates phase error much less than $\frac{2\pi}{10}$, which is generally accepted as very good in light optics [146, 147], while the cross-talk is clearly the biggest contributor to a non-ideal response. This behavior could be significantly improved in the next design iteration, where an additional top-ground layer could shield the effects from neighboring pixels and the conductive tracks leading to those pixels.

Characterizing the temporal response of the phase plate is also important for applications that rely on rapid switching between different electron probe shapes or phase configurations. Since the phase shift results from the projected potential in the electrodes, the response of the phase plate can be characterized by the time required to build up the potential. With the criterion of phase error < $\frac{2\pi}{10}$, the response time is measured to be less than 1.3 μs for reaching from 10 % to 90 % of $V_{2\pi}$ and is entirely dominated by the electronics performances.

5.3 Application Examples

Designer electron waveforms

To demonstrate the capability and visualize the effects of a freely programmable phase plate, we recorded the far-field diffraction patterns of various phase-modulated electron waves in a TEM (5.3). These patterns form rather complex configurations compared to ones formed by commonly-used round apertures, even when all phase plate elements are at ground potential. This is due to the amplitude modulation created by the set of holes, which produces highly delocalized tails. Previous theoretical research points out that the proportion of the electrons in these tails is directly related to the fill factor (% of the electron wave not blocked by the material of the phase plate) of the probe forming aperture [148]. Although improvement has been made on the fill factor (current design approximates 30%, while the proof of concept 2x2 version from 2018 [137] had only 17%), a large proportion of the electrons can still be expected in the tails.

Comparison between experimental intensity profiles and simulations reveals a close agreement. From 5.3, rows (b-e) show a phase shift of π applied to half of the total pixels with different patterns; therefore, the original single intense spot in the diffraction pattern is split into multiple parts due to destructive interference. Double-spots (b), quadruple-spots (c), and even a duodecuple-spot (d) consisting of six $0 - \pi$ pairs are shown. By taking into account the radial distribution of the rings, a checkerboard-like pattern (e) can be created. These patterns cover a few of the 48-dimensional Hadamard basis [149], which defines an orthogonal basis consisting entirely of pixels with either 0 or π phase. Lastly, (f) shows the result of a vortex setup with an orbital angular moment equal to 1 [150]. This is done by creating a phase ramp from 0 to 2π in the azimuthal direction. The vortex can be verified by the signature singularity point at the center of the resulting probe approximating one member of the Laguerre-Gaussian orthogonal basis set [151].

The phase plate can also create a phase profile imitating geometric optical elements and aberrations. Typically they can be modeled by a phase shift that follows a Zernike polynomial in the angle with respect to the optical axis [152]. How faithfully the phase

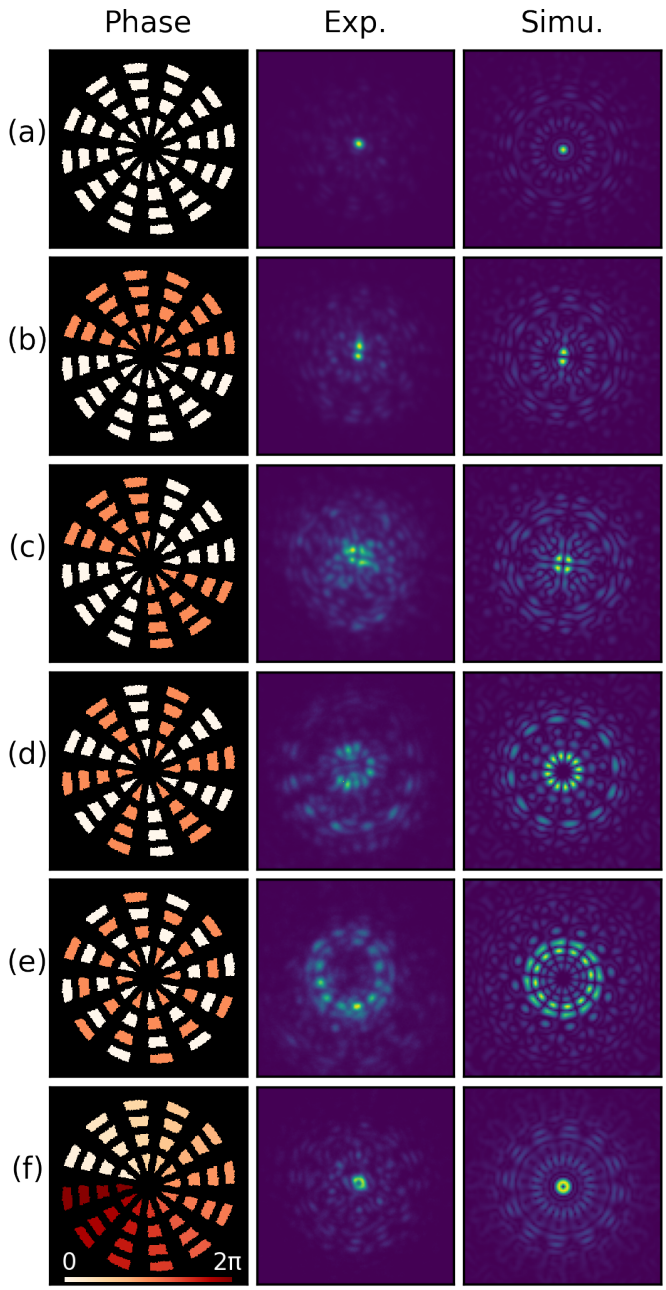


Figure 5.3: Realization of various electron quantum states. The three columns of figures, from left to right, are the phase configurations set on the phase plate, the resulting experimental probe images, and their simulated counterparts. Note the excellent agreement between expected and obtained results showing successful arbitrary wavefront shaping.

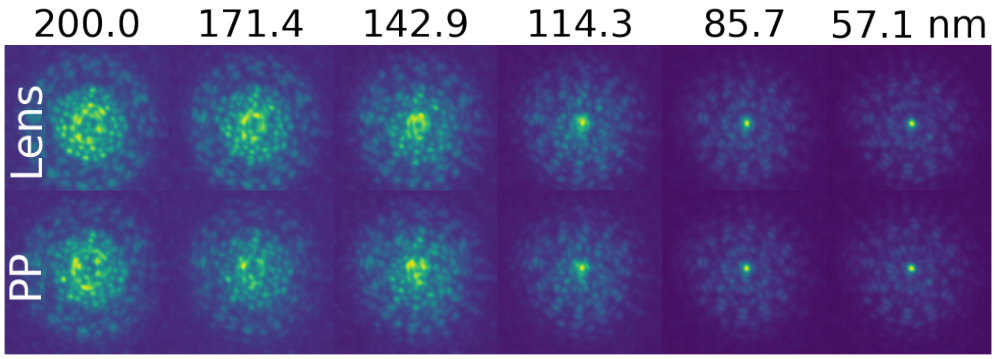


Figure 5.4: Defocused probes formed by defocusing the microscope lenses (top row) and the phase plate (bottom row) at 300 keV acceleration voltage and an opening angle of 1 mrad. Note the close similarity of both, showing that the phase plate can mimic the action of a round lens to up to 200 nm defocus.

plate can recreate such polynomials at different angles has been discussed theoretically in detail by Vega Ibáñez et al. [148] and relates to parameters such as the order of aberration, the fill factor, number of pixels, and pixel shape. Here, a defocus effect (second-order in angle) is introduced by either the conventional electromagnetic objective lens of the microscope or by the phase plate to demonstrate this concept. The resulting probe shapes are shown in 5.4, respectively. The two rows show good resemblance with each other up to 200 nm defocus. Further defocusing causes a steep phase ramp within the area of the individual pixels, which can not be faithfully reproduced anymore by the phase plate. For this reason, the phase plate can obviously not replace an actual (round) lens of any significant strength.

Object sampling with different wavefunctions

Electron microscopy is a process of sampling an unknown material with an electron wave. Once the incident wave interacts with the examined object, the information from the object is imprinted on the wave by creating changes in amplitude, phase, and the creation of inelastic scattering signals. When the measurement result of the interaction between the object and a beam with a given electron waveform provides insufficient information about the sample, different waves can be used to interrogate the object. For example, in-line holography [153, 154, 155] is done by recording the intensity of a beam while varying the phase by changing the defocus of the objective lens. STEM essentially describes a process to accumulate information about the material by a dense sampling while spatially scanning a localized electron beam. In both cases, multiple measurements while changing the incoming electron wave enriches the acquired information and eliminates confusion that can sometimes not be resolved with a measurement process that only uses a beam with a single static waveform.

Such multi-waveform sampling schemes rely entirely on the ability to alter the wavefunction of the electron beam state. Even though some form of modulation of the wavefunction is present in any electron microscope (e.g., defocus, beam tilt, beam shift,

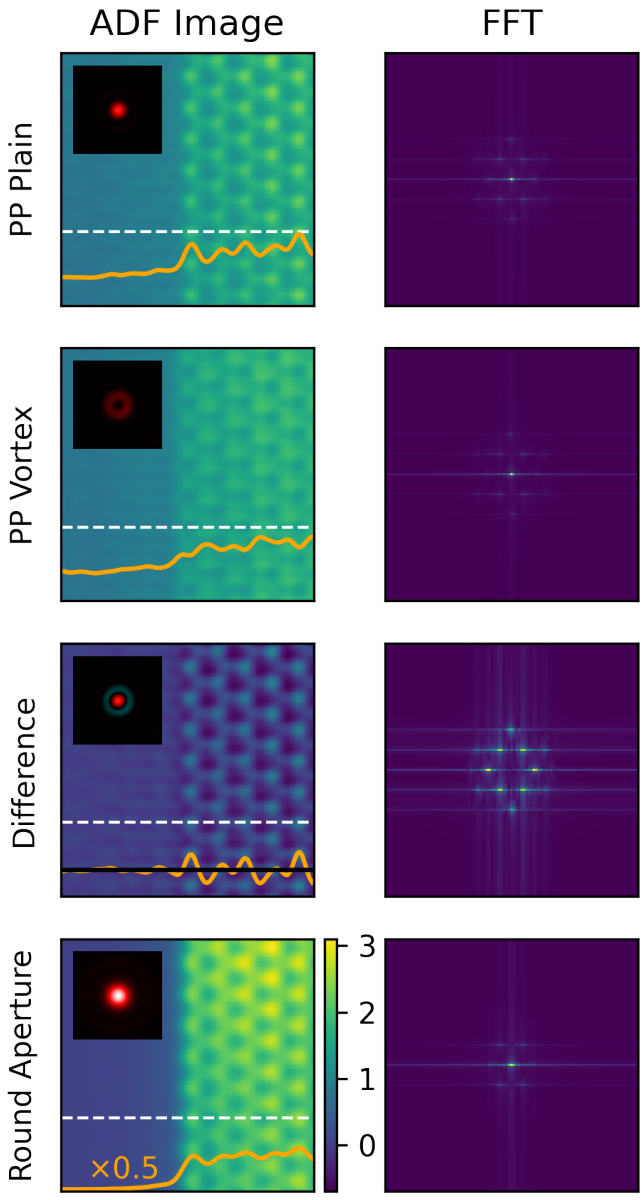


Figure 5.5: Simulated ADF images of various probe shapes (see the insets) and their Fourier transforms. The line profiles (orange lines) are taken at the position of the white dashed line in each image. Zero in the difference image is indicated with a black line, while in other images zero is set at the bottom edge of the figures. The intensity profile in the round aperture image is halved for better presentation.

or aberration correctors), they often rely on electromagnetic elements, which can suffer from slow settling times and hysteresis effects. For example, in the acquisition of through focal series images, an update rate in the order of seconds to minutes is typically applied to induce small focal changes in the objective lens[156, 157].

The phase plate presented here can update to an entirely new pattern in a few μs without hysteresis so that complex sampling schemes can be realized efficiently. For instance, the phase plate can cycle over a few different wavefront settings for each probe position in a STEM recording. Compared to through focal TEM acquisition, where the focus is changed between recording image frames, we could now update multiple focus levels for each probe position in a STEM scan, providing, e.g., increased depth of field. This dramatically reduces the difficulty of realigning each image, especially in cases of severe sample drift, and also avoids inconsistencies caused by contamination building up on the sample over time.

Changing defocus is just one of the possible wavefunctions to sample an object of interest. Being a non-orthogonal change to the beam wavefunction, it could be argued that this is not even an optimal choice of basis. The adaptability and rapid response of the phase plate can be extended to a wide variety of orthogonal basis sets that can be specifically chosen to efficiently encode selected knowledge about the sampled object into the probing electron waves.

This concept is widely used in light microscopy and serves as an important cornerstone for techniques such as stimulated emission depletion (STED) microscopy [158, 159, 160] and switching laser mode microscopy (SLAM) [161, 162]. Two or more waveforms sequentially illuminate the sample, and the sharp feature created by the difference between the illuminating waves can be exploited to increase the resolution of the final image.

The same concept can now be applied to electron microscopy with a two-fold beneficial effect. Indeed, changing between a probe state with and without orbital angular momentum will slightly improve image resolution due to differential imaging with both probes (super-resolution). But more importantly, this method also cancels the long probe tails arising from the amplitude modulation of the pixel shapes, as these tails are nearly identical for both probe wavefunctions. This is a far more critical effect as it dramatically increases the practical resolution that can be obtained even when the fill factor of the phase plate is not ideal and shows a way to significantly outperform the results presented earlier for the single waveform aberration correction prospects of programmable phase plates for electrons [148]. The result of this differential scheme is demonstrated with high-angle annular dark field (HAADF)-STEM simulation (5.5). Electron probes, as results of the far-field diffraction of three illuminating wave functions, created by a phase plate with zero phase, vortex phase, and a conventional round aperture, are used to scan a single-layer hexagonal boron nitride that shares half of the simulation box with vacuum, with 200 keV electron beam energy, a spherical aberration C_3 of 1.2 mm and operating at Scherzer defocus, in agreement with a typical uncorrected TEM instrument. The convergence angles of the electron probes are set to 9.5 and 11 mrad for the round aperture and the phase plate, respectively. We select a larger opening semi-angle for the phase plate since its capability to correct aberrations yields an optimal imaging condition at 11 mrad. The subtraction of the vortex image from the plain phase plate is then presented as the difference image.

The simulated images are then juxtaposed to illustrate the effect of the tails, and an

intensity profile (orange line) is drawn across each image (at the position of the white dashed lines). Both images from the phase plate have non-zero intensity in the vacuum area (the left half of the simulation box) due to the tails' interaction with the crystal. The profile from the image formed with the round aperture shows much faster decay as the intensity distribution of an aberrated Airy probe is more concentrated. The difference image demonstrates good cancellation of this false background, and the intensity profile quickly converges to zero, with small fluctuations due to slight differences between the tail configuration of the two probes. This result shows that the phase plate can indeed provide an excellent tail effect cancellation when alternating between a flat phase and vortex phase probe. The resolution is significantly improved over the non-corrected round aperture at the expense of some signal loss related to the fill factor and the loss of low-frequency sample information. This demonstrates the potential for aberration correction with a device that is significantly smaller (< 5 mm), lighter, faster (μs), more energy efficient (< 5 W), and requires far less stringent control over the precision of the voltage/current sources as compared to current multi-pole correctors.

Adaptive optics

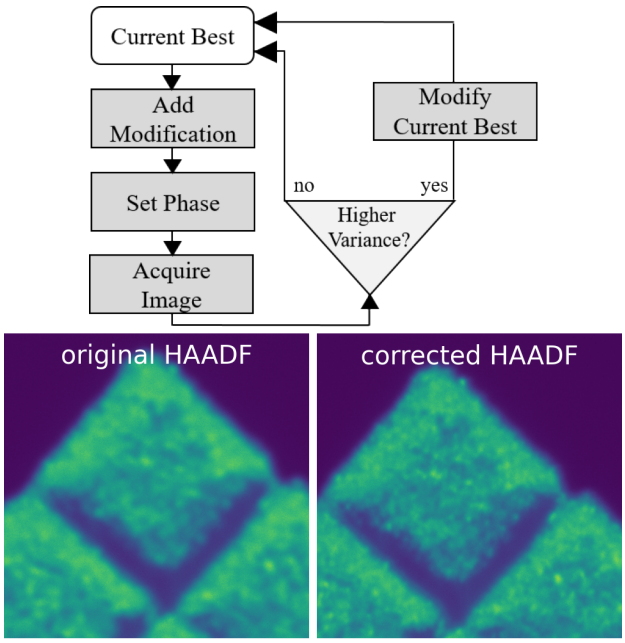


Figure 5.6: Schematic of the adaptive probe correction with phase plate. The HAADF images before and after the correction are shown below.

Using the fast and hysteresis-free phase programming offered by the electrostatic phase plate opens the attractive possibility of adaptive optics. As a proof of concept, such a setup is realized (fig. 5.6). An algorithm repeatedly reshapes the electron probe with the phase plate in order to reach a higher variance in the high-angle annular dark field (HAADF) image, which is taken as a figure of merit that links with 'image sharpness' [163].

The algorithm sequentially adds phase modifications from a list of discretized low-order Zernike polynomials to the latest best-performing phase configuration. Zernike polynomials are chosen since they exhibit close similarity to common aberration in the electron microscope and form a complete, orthogonal basis. A HAADF image is consequently recorded with every new probe. If the variance is higher in the new image, the current best is replaced with this new variation. Once all the configurations are tested, their magnitudes in terms of phase value are reduced by half for a further refinement step. The process is demonstrated by inserting the phase plate in the C2 aperture of a probe and image-corrected FEI-Titan operating at 300 keV in micro-probe mode with a convergence angle of 1 mrad (to minimize the effect of aberrations and partial coherence effects). The HAADF image is taken from a gold cross-grating test sample with a deliberately introduced initial defocus of approximately 1 μm. The result of the correction is shown in figure 5.6. The process converges after 32 iterations with a sharper resulting image, even though 1 μm defocus cannot be entirely compensated by the phase plate due to the steep phase profile. The result shows the feasibility of counteracting the lens defocus automatically. The process takes approximately 1 minute, but this time is currently dominated by sub-optimal software handshaking between scan engine control, image readout, and phase plate control, and can be dramatically improved in the future. As an estimate, with the assumption that an update can be made by evaluating a minimum area of 100x100 pixels at 1 μs dwell time (a reasonable dwell time to produce HAADF images with an acceptable noise level), the update rate for the correction scheme would be 1 khz. This frequency is easily within reach of the phase plate, which currently offers a maximum update rate of 100 kHz, limited by the electronics. This would result in an adaptively optimized image within 10 ms which would be a small fraction of the time to acquire, e.g., a full 1024x1024 frame. Of course, this time depends on the beam current, as enough image quality is required to make good decisions on the next step. Further work is needed to evaluate the best goal function and most optimum control loop, but the proof of concept demonstrates the scheme's feasibility.

This process could bring significant benefits for the automation of microscopy experiments. Automatic data acquisition and feature identification are widely used for life science research and quality control in the semiconductor industry. With them, the analysis of large amounts of samples can be done without operator intervention, and the demonstrated probe correction scheme can be utilized for maintaining the quality of the optical system over a much longer operation time.

This iterative optimization process can also be extended to any technique in electron microscopy where a specific quantifiable property is related to the shape or phase of the electron probe. For example, in electron energy loss spectroscopy, the intensity of a specific plasmon peak can be tracked while reshaping the electron probe until the optimal probe shape selectively highlights the corresponding plasmon mode [164].

Phase programmed phase reconstruction

Besides shaping a focused electron beam, the phase modulation capability of the phase plate operating under parallel-beam conditions can bring new opportunities for other microscopy applications. For example, coherent diffraction imaging and ptychography can benefit from using the phase plate as a "modulator" or a "diffuser" to break symmetry in the illuminating beam and thus increase the robustness and convergence rate of the

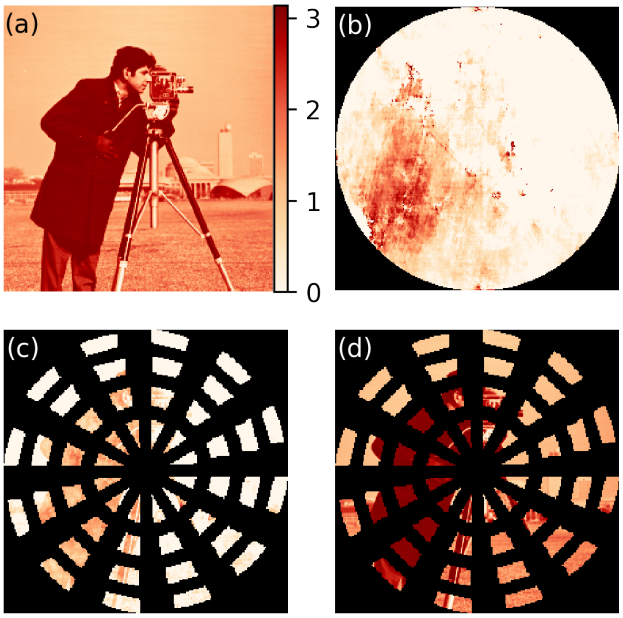


Figure 5.7: Simulated phase reconstruction from recorded diffraction patterns with various illuminating beams. (a) The ground truth phase image of the object. (b-d) Reconstruction results from illuminating beams formed by a conventional round aperture, flat phase plate, and phase plate with random phase configuration, respectively. The dark region indicates the opaque part of the aperture. Note the significant improvement in phase reconstruction quality when the incoming beam is phase randomized. As the object is only illuminated once, reconstruction is only possible in those areas where the amplitude is not zero.

reconstruction. The benefit of a modulator has been widely reported and studied in the field of light microscopy[165] and electron microscopy [166, 167, 168]. Among the reported realizations of ptychography in electron microscopy, with or without a modulator, the reconstruction of the complex object relies on repeated sampling at different locations of the object, with the criterion that the illuminating beam partially overlaps with the sampling at a nearby position. This overlap creates the so-called "information redundancy" [78, 169], which eliminates the twin-image artifact [59] that originates from the central symmetry of the illuminating beam. On the other hand, such symmetry can be easily broken by a random phase configuration introduced by the phase plate instead of the displacement of the beam or the sample.

We hereby demonstrate this concept by performing phase reconstruction on simulated diffraction patterns from a pure phase object target (5.7a). The diffraction patterns are generated by different illuminating waves, formed with a round aperture, the phase plate set to zero, and one randomly generated phase configuration. The phase reconstruction is again based on the GS algorithm, and the resolved objects are obtained after 50 iterations. The results are shown in 5.7(b-d). Neither a round aperture nor a zero phase plate could generate a convincing reconstruction result, as the geometry of both apertures is centrosymmetric. However, introducing a random phase configuration increases the reconstruction quality significantly, despite the sample being only illuminated at one

beam position.

The amplitude modulation of the phase plate inevitably results in missing information about the reconstructed object, which could be filled by moving the beam or the sample to illuminate the whole region of interest at least once. It should be noted here that the phase plate is placed in front of the sample, and all electrons interacting with it are recorded. This means that the limited fill factor does not reduce the electron dose efficiency nor increase beam damage on the sample.

5.4 Conclusion

We report the successful realization of arbitrary wavefront shaping of electrons with a novel 48-pixel programmable electrostatic phase plate. The phase plate is capable of introducing a phase shift of more than 60π, as well as fine-tuning the phase value with step size as small as $3{\cdot}10^{-3}$ π for 300 keV coherent electron beams. Cross-talk between pixels was shown to be $< 15\%$ and can be improved further with better shielding electrode geometries. This brings modern adaptive light optics concepts into the domain of electron beam instruments. The rapid response of the device allows up to 100 kHz update rates making it possible to do on-the-fly auto-tuning of differential contrast schemes without a noticeable recording time penalty for the user. The examples demonstrate the potential for a rich field of emerging applications offered by the phase degree of freedom. Immediate use cases focus on electron microscopy, but other electron beam instruments, such as, e.g. e-beam lithography or semiconductor inspection tools, could also profit significantly from this realization. With an even broader perspective, we demonstrate here the arbitrary preparation of coherent quantum states that might be exploited in novel quantum information/computing schemes over a much wider range of electron energies than the ones demonstrated here.

Chapter 6

Atom Differentiation with Elastically Scattered Electrons

6.1 Introduction

As briefly introduced in chapter 1, the composing elements of a material can be determined from signals stemming from inelastic scattering events due to the electron-sample interaction, such as energy loss and emitted X-ray. This is because the energies of the orbitals are quantized, and their energy levels are determined by the positive charge in the nuclei. Therefore, energy exchange corresponding to the excitation/de-excitation of the orbital electrons is characteristic of the element type.

On the other hand, the scattering events also contain information about the type of interacting atoms. From equation 1.1, a clear connection between the phase profile of the scattered electron wave and the sample potential can be found, and since the potential profile associated with each element is different, so should the waveform of the exit wave when interacting with different atoms. As the exit wave propagates, the sample-induced phase shift causes changes in both the amplitude and the phase of the wave, therefore allowing the effect of the electron-sample interaction to be recorded and measured. Such measurement provides valuable information to distinguish different types of atoms, which has already been reported [23, 170] and used as a primary technique to distinguish chemical differences, such as visualizing core-shell structures [171] and distinguish dopants [172] etc.

This chapter investigates the ability to differentiate elements using only the signal of elastically scattered electrons. An electron wave that experiences elastic scattering acquires a phase shift corresponding to the projected potential of the scatterer. Since each type of atom has a different projected potential, the induced phase shift, if it can be measured somehow, should reveal the atom's identity.

We begin with the physical expression of the measurement of a quantum state, such as an electron wave, from which the optimal measurement and incident waveform for atom-type differentiation can be derived. Although the phase shift due to the potential of an atom is invariant to the incident waveform, it is worth exploring whether waveform design helps imprint more phase information onto the amplitude of the transmitted

wave. With the conclusion drawn from the theoretical derivation, several designs of probing waveforms that can potentially be realized, for example, with a similar device as discussed in chapter 5, are shown. These designs are found to be consistent with other reported proposals of incident wave designs that boost contrasts of weak scatterers. The performance of these waveforms in creating differences in the transmitted waves, lowering the probability of error in determining atom species, and increasing contrast response to various types of scatterers is discussed.

6.2 Physical Formulation

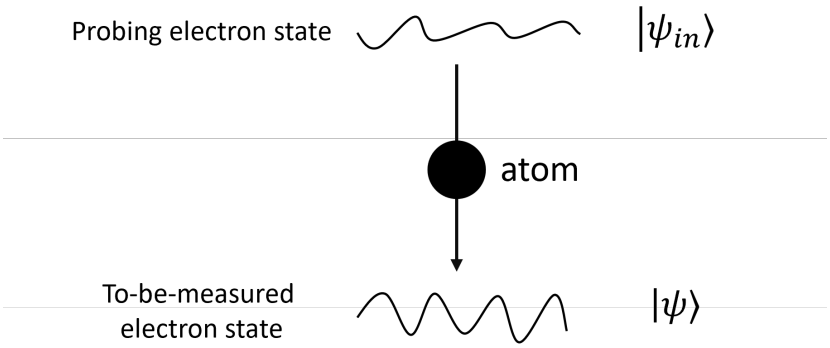


Figure 6.1: Determining the atom species by probing the atom with an electron at the prepared state $|\psi_{in}\rangle$. After the interaction with the atom and lenses in the imaging system of the TEM, the transmitted electron wave is found in state $|\psi\rangle$, from which the correct atom type may be determined with specific measurements.

In transmission electron microscopy (TEM), the wave-forming components of the microscope, including apertures, lenses, correctors, etc., prepare electron waves of certain waveforms in a specific quantum state to probe the sample. The transmitted electron wave that has interacted with the sample propagates with the information of the probed object and can be eventually recorded by the detector. In other words, the investigated object is never directly measured, but the electron wave that has interacted with it.

This scheme of measurement can also be used for the purpose of differentiating atoms of different types. An electron wave prepared by the optical components of the TEM is guided to interact with the unknown atoms so that information about the atom species is written to the waveform. This can cause variation in the intensity distribution of the transmitted wave, and based on its recording, the correct type of atom may be deduced if a connection between the wave intensity distribution and the atom type can be made. From the theoretical representation of an experimental scheme, shown in figure 6.1, both the measurement of the transmitted wave and the shape of the incident wave could be potentially optimized to maximize the accuracy when determining different atom species.

With that in mind, this section of the chapter is dedicated to the physical formulation of the measuring processes of such an experiment. The measurement of the wave intensity is formulated as an operation on the quantum state describing the electron wave function,

and it is shown as equivalent to the measurement of the projection of the electron state on a certain eigenstate of the operator. The most efficient measurement to distinguish wave functions can then be found as a result of the formulation, and the optimal design of the probing electron state for atom differentiation can also be derived based on the observation of how electron waves are distinguished.

Measuring an electron wave

The measurement of a quantum state associated with a specific operator can be expressed by its eigenstates $\left\{\left|\alpha^{(i)}\right\rangle\right\}$ and eigenvalues $a^{(i)}$:

$$\hat{A} = \sum_i a^{(i)} \left|\alpha^{(i)}\right\rangle \left\langle\alpha^{(i)}\right| \tag{6.1}$$

Measuring a state $|\psi\rangle$ with $\hat{A}$ throws the original state into the eigenstates of the operator:

$$\hat{A}\,|\psi\rangle = \sum_i a^{(i)} \left\langle\alpha^{(i)}\middle|\psi\right\rangle \left|\alpha^{(i)}\right\rangle \tag{6.2}$$

where $\left\langle\alpha^{(i)}\middle|\psi\right\rangle$ represents the inner product between the states $\left|\alpha^{(i)}\right\rangle$ and $|\psi\rangle$, as shown in figure 6.2a. The expectation of the measurement is:

$$\left\langle\psi\middle|\hat{A}\middle|\psi\right\rangle = \sum_i a^{(i)} \left\langle\psi\middle|\alpha^{(i)}\right\rangle \left\langle\alpha^{(i)}\middle|\psi\right\rangle \tag{6.3}$$

and the probability to acquire the eigenvalue $a^{(i)}$, or that the probability of the measured state collapsing into the eigenstate $\left|\alpha^{(i)}\right\rangle$, is:

$$\left\langle\psi\middle|\alpha^{(i)}\right\rangle \left\langle\alpha^{(i)}\middle|\psi\right\rangle = \left|\left\langle\alpha^{(i)}\middle|\psi\right\rangle\right|^2 \tag{6.4}$$

which is often referred to as the Born rule. A classic example of measuring a quantum state is the measurement of the polarization of a photon. With the state $|\psi\rangle$ representing the photon, the measurement can be expressed as an operator $\hat{A}_p$:

$$\hat{A}_p = H\,|H\rangle\,\langle H| + V\,|V\rangle\,\langle V| \tag{6.5}$$

where $|H\rangle$ and $|V\rangle$ are the orthogonal polarized states of photon. If the experiment is designed so that only the detection of the horizontally polarized state would lead to a signal output, we can make the eigenvalue associated with the measurement as such: $H = 1,\ V = 0$, and the probability to acquire the eigenvalue 1 can then be found according to equation 6.4 as $|\langle H\,|\,\psi\rangle|^2$.

The detection of an electron can also be formulated in such a manner. This process happens as the incident electron interacts with the sensing layer of the detector and dumps its energy to create photons or electron-hole pairs. As a consequence, a signal indicating the detection of the electron is triggered, and the position of the electron can also be found where the exchange happens. Such measurement can be written in the form of equation 6.2, with $|\psi\rangle$ as the wave function of the electron before the measurement takes place and the triggered signal as the eigenvalue associated with the measurement. After the measurement, the position of the electron can be determined by the creation of a signal, and therefore the eigenstates into which $|\psi\rangle$ falls are a set of localized delta functions whose position in the detector plane is unambiguous:

$$\left|\alpha^{(i)}\right\rangle = \delta(\vec{k} - \vec{k}_i) \tag{6.6}$$

where the vector $\vec{k}$ describes the detector plane, and $\vec{k}_i$ is the position of the delta function's peak corresponding to the eigenstate $\left|\alpha^{(i)}\right\rangle$.

In summary, the recording of the electron wave intensity with a detector can be expressed as an operator $\hat{A}_D$:

$$\begin{aligned} \hat{A}_D &= \sum_i a^{(i)} \left|\alpha^{(i)}\right\rangle \left\langle\alpha^{(i)}\right| \\ a^{(i)} &= \begin{cases} 1 & \text{if } \ i \in \mathrm{D} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{6.7}$$

where D describes a subspace on the detector plane that is covered by the detector, with the eigenvalue $a^{(i)}$ equal to 1 if the state $\left|\alpha^{(i)}\right\rangle$ belongs to D and 0 if not.

Following this narrative, the eigenstates of the measurement are fixed to the space where the detector is placed, but it is not to be forgotten that other operations can be done on the wave so that the desired component of the wave can be projected onto the bases defined by the detector. Such transformation may be achieved by an optical component in the TEM that transforms the electron wavefunction from state $|\psi\rangle$ to $|\psi'\rangle$. The corresponding operator $\hat{M}$ with the effect of the said transformation can be written as.

$$\hat{M}\,|\psi\rangle = |\psi'\rangle \tag{6.8}$$

The probability of acquiring $a^{(i)}$ by measuring the transformed wave is:

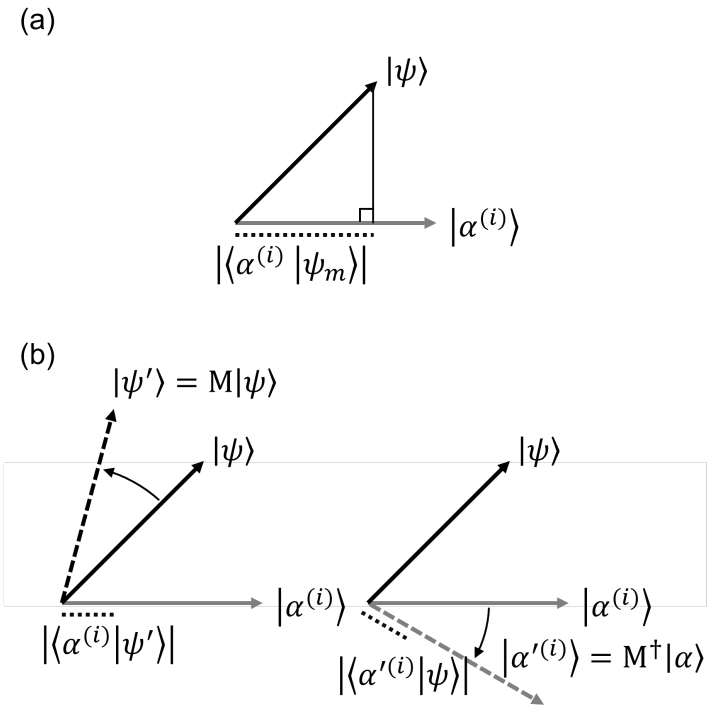


Figure 6.2: Vectorial representation of the measuring process. (a) shows the probability of $|\psi\rangle$ collapsing into $\left|\alpha^{(i)}\right\rangle$ as the square inner product between the two states. (b) demonstrates that the transformation operator on the state $|\psi\rangle$ can be applied instead to the eigenstate of the measuring operator $\left|\alpha^{(i)}\right\rangle$ without changing the probability of acquiring the eigenvalue $a^{(i)}$

$$\begin{aligned} & Pr(a^{(i)}, \psi') \\ &= \left|\left\langle \alpha^{(i)} \middle| \psi' \right\rangle\right|^2 \\ &= \left|\left\langle \alpha^{(i)} \middle| \hat{M} \middle| \psi \right\rangle\right|^2 \\ &= \left|\left\langle \alpha'^{(i)} \middle| \psi \right\rangle\right|^2 \end{aligned} \tag{6.9}$$

where:

$$\hat{M}^\dagger \left|\alpha^{(i)}\right\rangle = \left|\alpha'^{(i)}\right\rangle \tag{6.10}$$

Equations 6.9 and 6.10 show that the operator $\hat{M}$ that applies the transformation to the wave can be included in the set of bases $\{|\alpha\rangle\}$ of the measurement to create a new set of bases $\{|\alpha'\rangle\}$. This observation shows that an arbitrary component of the electron

wave $\alpha'^{(i)}$ can be measured if the corresponding operator $\hat{M}$ can be realized. If neither amplification, attenuation, nor cropping of the wave happens in the transformation, the transformed bases $\{\alpha'\}$ should be unitary, while the orientation can be freely defined, as shown in figure 6.2b.

The transformation described here can be found in many places of a TEM. For instance, if $|\psi\rangle$ describes the state of the electron on the sample plane after its interaction with the material, the diffraction formed by the electron can be found with a Fraunhofer diffraction operator, or the exit wave after propagation of a short distance can be found with a Fresnel diffraction operator. These transformations do not introduce any more knowledge about the sample into the wave as they happen after the electron-sample interaction; however, they serve crucial roles in TEM experiments since the concerned property of the material may be hidden in the phase of the exit wave and can only be measured and analyzed with such transformation.

Optimal measurement for waveform differentiation

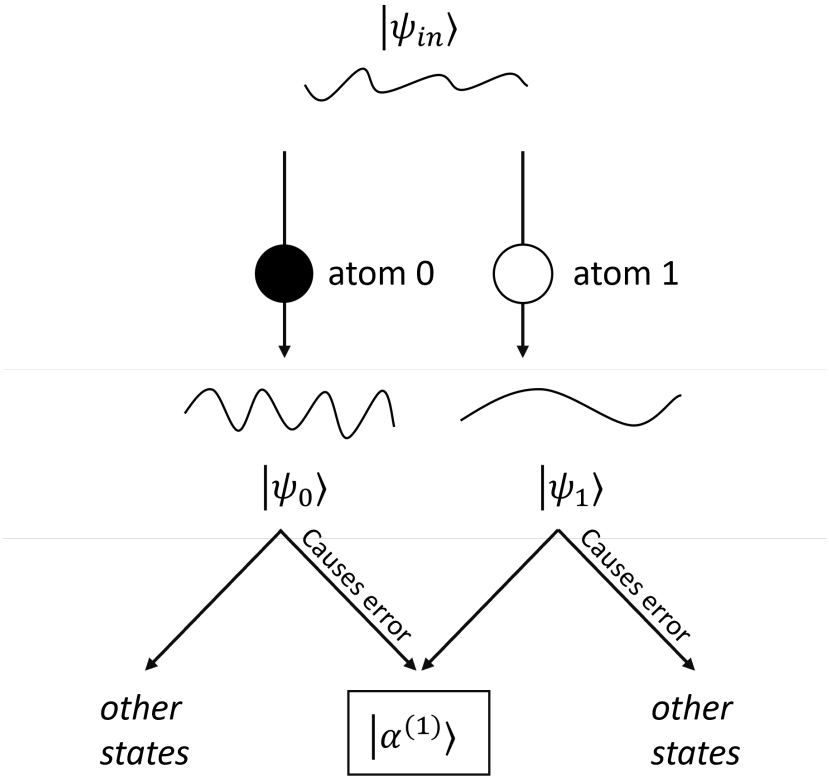


Figure 6.3: Determining whether the state $|\psi\rangle$ is $|\psi_1\rangle$ or $|\psi_0\rangle$ by checking if the state collapses into $|\alpha^{(1)}\rangle$ or not. Possible causes of error in the determination of the state are indicated.

As mentioned above, transformations of the electron wave can be performed to optimize the measuring of specific properties of the material. For example, the periodicity of

the sample can be easily found by observing the diffraction of the exit wave, and the positions of the atomic columns can also be known by defocusing the wave so that the phase information is reflected on the amplitude variation. For the specific purpose of atom differentiation, there may also be an optimal measuring scheme that maximizes the differences in the observation related to the atom species

Considering a case where the probed atom is one of the two different species with a fixed state describing the probing electron wavefunction, the resulting electron state after the interaction with the atom is described by $|\psi\rangle$. Since we are already sure that the probed atom is one of the two types, the exit wave is either $|\psi_0\rangle$ if the interacting atom is of type 0 or $|\psi_1\rangle$ for type 1. In this case, the optimal measurement to distinguish the type of the atom can be formulated as an operator that would result in the most drastic difference when applied to $|\psi_0\rangle$ or $|\psi_1\rangle$, for example, an operator that maximizes the probability that $|\psi_1\rangle$ would collapse into the eigenstate $\left|\alpha^{(1)}\right\rangle$ and $|\psi_0\rangle$ into other eigenstates can distinguish the two states by observing whether the electron after the measurement is in state $\left|\alpha^{(1)}\right\rangle$ or not.

However, a pair of states $|\psi_0\rangle$ and $|\psi_1\rangle$, as a result of two arbitrarily chosen atoms, is not guaranteed to be orthogonal to each other. As a consequence, the differentiation cannot be perfectly accurate as there would be a non-zero probability that $|\psi_0\rangle$ collapses into $\left|\alpha^{(1)}\right\rangle$ or $|\psi_1\rangle$ collapses into other eigenstates. When such a scenario happens, it would lead to an error in the differentiation of the wave and the atom species. The optimal measurement scheme is thus created by minimizing the probability of error (P_e), which in the case of distinguishing $|\psi_0\rangle$ and $|\psi_1\rangle$ can be expressed as:

$$P_e = \frac{1}{2}\left|\left\langle \alpha^{(1)} \middle| \psi_0 \right\rangle\right|^2 + \frac{1}{2}\left(1 - \left|\left\langle \alpha^{(1)} \middle| \psi_1 \right\rangle\right|^2\right) \tag{6.11}$$

The first term of equation 6.11 represents the probability of a state $|\psi_0\rangle$ to be thrown into the eigenstate $|\alpha^{(1)}\rangle$ after the measurement, and the second term is the probability of a state ψ_1 not to collapse into $|\alpha^{(1)}\rangle$, as shown in figure 6.3. Both terms are multiplied by a factor of 1/2, indicating that both occurrences have equal probability. To minimize the P_e, the $|\alpha^{(1)}\rangle$ should be chosen to minimize the square inner product with the state $|\psi_0\rangle$ and maximize the square inner product with $|\psi_1\rangle$. Since the square inner product between the two unitary states can be found as $\cos^2\theta$, with θ being the angle between both states (see figure 6.4), equation 6.11 can thus be reformulated with the assumption that all three states are normalized:

$$P_e = \frac{1}{2}\cos^2\varphi + \frac{1}{2} - \frac{1}{2}\cos^2(\varphi - \theta) \tag{6.12}$$

In this equation, φ is defined as the angle between $|\psi_0\rangle$ and $\left|\alpha^{(1)}\right\rangle$, and θ the angle between $|\psi_0\rangle$ and $|\psi_1\rangle$ (see figure 6.4). To minimize the P_e by choosing the optimal $\left|\alpha^{(1)}\right\rangle$, we take the derivative of equation 6.12 and look for the angle φ that results in 0:

$$\begin{aligned}
&\frac{d}{d\varphi}\left(\frac{1}{2}\cos^2\varphi + \frac{1}{2} - \frac{1}{2}\cos^2(\varphi-\theta)\right) \\
&= -\cos\varphi\sin\varphi + \cos(\varphi-\theta)\sin(\varphi-\theta) \\
&= \frac{1}{2}\left[-\sin(2\varphi) + \sin(2\varphi - 2\theta)\right]
\end{aligned} \tag{6.13}$$

The solutions to equation 6.13 are found to be:

$$\begin{aligned}
&\varphi_0 = \text{anything, while } \theta = k\pi \\
&\varphi_1 = \frac{\pi}{4} + \frac{\theta}{2} + k\pi
\end{aligned} \tag{6.14}$$

The first solution is a trivial case where the two quantum states are identical or pointing in the opposite direction in their parameter space, as shown in figure 6.4a. This results in $P_e = 1/2$ and makes φ not relevant, which is expected since the projection of such two states will always have equal length, regardless of the choice of $\left|\alpha^{(1)}\right\rangle$. The second solution φ_1 minimizes (or maximizes) the P_e for cases of the two different quantum states. As shown in 6.4b, when $\varphi = \varphi_1$, the projections of the two states reach the largest difference in length. An even value of k yields the minimum P_e, whereas an odd k maximizes the PE, which corresponds to the identical minimum P_e if $|\psi_0\rangle$ is made to collapse into the eigenstate $\left|\alpha^(1)\right\rangle$ in the first place. Since it makes no fundamental difference by switching the assignment, we choose $k = 0$ for the following discussion

Using the solution of φ_1, the P_e can be further derived as:

$$\begin{aligned}
&\frac{1}{2}\cos^2\varphi + \frac{1}{2} - \frac{1}{2}\cos^2(\varphi-\theta) \\
&= \frac{1}{2}\left(1 + \cos^2\varphi - \sin^2\varphi\right) \\
&= \cos^2\varphi \\
&= \frac{1}{2}\left(1 - |\sin\theta|\right)
\end{aligned} \tag{6.15}$$

It is worth mentioning here the Helstrom bound [173], which states the minimal P_e to distinguish any two quantum states is:

$$P_e = \frac{1}{2}\left(1 - \sqrt{1 - 4P_0P_1\left|\langle\psi_0\,|\,\psi_1\rangle\right|^2}\right) \tag{6.16}$$

where P_0 and P_1 represent the probabilities that the observed wave is ψ_0 or ψ_1, respectively. By making both 1/2, the Helstrom bound can be simplified as:

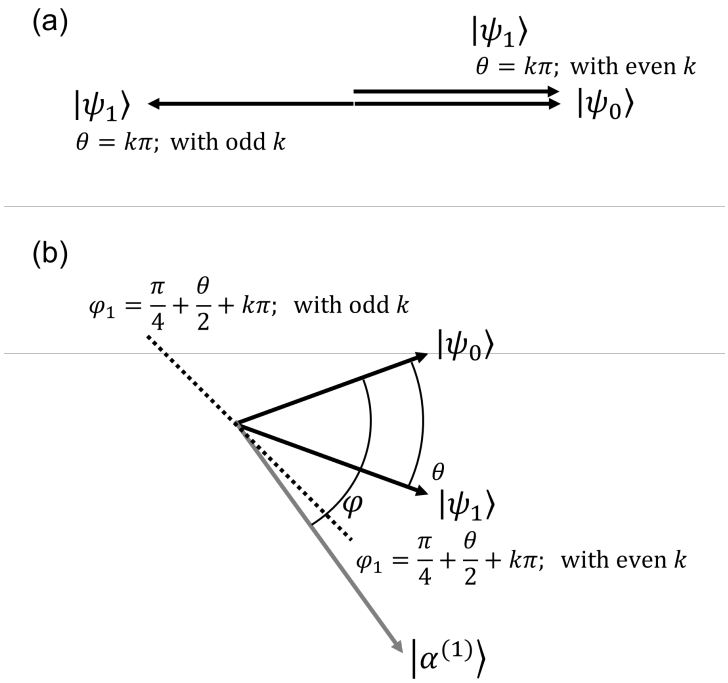


Figure 6.4: Distinguishing states $|\psi_0\rangle$ and $|\psi_1\rangle$ with a projection vector $|\alpha^{(1)}\rangle$. The angle between the two state vectors is θ, and φ the angle between $|\psi_0\rangle$ and $|\alpha^{(1)}\rangle$. (a) shows the special case where the two states can not be distinguished, regardless of the choice of the measuring states $|\alpha^{(1)}\rangle$. (b) shows the angles corresponding to the maximum and minimum of P_e.

$$\begin{aligned} P_e &= \frac{1}{2}\left(1 - \sqrt{1 - |\langle\psi_0 \,|\, \psi_1\rangle|^2}\right) \\ &= \frac{1}{2}\left(1 - \sqrt{1 - \cos^2\theta}\right) \\ &= \frac{1}{2}\left(1 - |\sin\theta|\right) \end{aligned} \tag{6.17}$$

which is the same as the result of our derivation in equation 6.15.

It is also worth mentioning one of the fundamental differences between the measuring process described here and the one occurring in a TEM experiment. The P_e derived here is based on the measurement of a quantum state that collapses into one of the eigenstates of the measurement operator, and thus the original quantum state is destroyed afterward. For atom differentiation, this description remains true for the measured state (the electron wave can only be detected once), but the probed atom may be assumed to stay in its original state, and therefore copies of the transmitted electron wave can be created. With that, the decision on the right atom species is no longer made on the result of a single measurement. Instead, the probability of receiving the scattered electron can

be used to characterize the scatterer by repeating the measuring process. In other words, the decision is no longer made on the resulting state of a single measurement but on the probability of it happening, learned after multiple trials. As described in equation 6.4, the expected amount of times the unknown state $|\psi\rangle$ collapsing into $\left|\alpha^{(1)}\right\rangle$ after N independent trials can be found as:

$$\text{Exp.} = N\left|\left\langle \alpha^{(1)} \middle| \psi \right\rangle\right|^2 \tag{6.18}$$

Assigning an eigenvalue of 1 to $\left|\alpha^{(1)}\right\rangle$ and 0 to all other eigenstates, the probability density function (PDF) of the measured results after N trials follows a binomial distribution with the probability of success p equal to $\left|\left\langle \alpha^{(1)} \middle| \psi \right\rangle\right|^2$:

$$Pr(x) = \frac{N!}{(N-x)!x!}p^x(1-p)^{N-x} \tag{6.19}$$

It shows that the PDF is not only a function of the measurement $\hat{A}$, but also the dose N, and thus a universal optimal measurement with minimized P_e at different N cannot be found. However, an important conclusion can still be drawn from equation 6.15: the final form of the formula to calculate the P_e writes $(1 - |\sin\theta|)/2$, showing that the error stems from the fact that the two to-be-distinguished states are not orthogonal to each other. For orthogonal states where θ is equal to $\pi/2$, the P_e is then effectively reduced to 0, even for the case where the measurement is done with only one electron.

Incident wave design

The conclusion from the last part shows that the P_e in distinguishing waveform and atom type is lowered if the quantum states representing the transmitted waves can be more orthogonal to each other or that the angle between the two states is closer to $\pi/2$. This information can then be used to design the optimal probing wave for the differentiation.

To disentangle the measurement of the transmitted wave and the gain in the incident wave design, we notice that the angle between the two states is invariant to any transformation that is equivalent to a unitary operator, which possesses the property that its inverse is also its adjoint, or $\hat{U}^{-1} = \hat{U}^{\dagger}$. This can be shown in the following:

$$\begin{aligned}\left\langle \psi_0' \middle| \psi_1' \right\rangle &= \langle \psi_0 | \, \mathrm{U}^{\dagger}\mathrm{U} \, | \psi_1 \rangle \\ &= \langle \psi_0 \, | \, \psi_1 \rangle\end{aligned} \tag{6.20}$$

which is found to be the same as if the operator was not used in the first place, therefore yielding the same P_e according to equation 6.15. This observation thus allows us to conveniently optimize the waveform in the sample space while the measurement is likely to be done in another, as long as the two space can be connected by the transformation of a unitary operator.

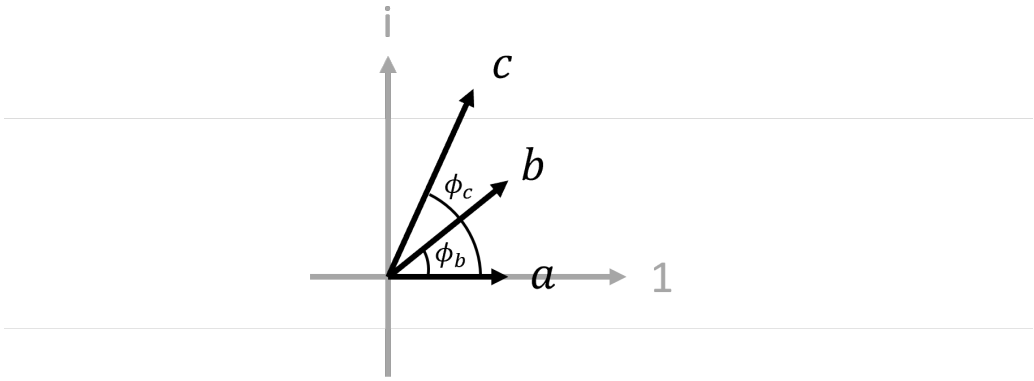


Figure 6.5: A simplified case with only three complex numbers.

Consider a case that two different atoms are illuminated in turn by an incident wave of a fixed waveform, by the phase object approximation (POA) the exit waves are found to be:

$$\begin{aligned} \psi_0(\vec{r}) &= \psi_{in}(\vec{r})e^{i\phi_0(\vec{r})} \\ \psi_1(\vec{r}) &= \psi_{in}(\vec{r})e^{i\phi_1(\vec{r})} \end{aligned} \tag{6.21}$$

where ψ_{in} is the incident wave and ψ_0, ψ_1 are exit waves influenced by two different types of atoms. According to the POA, the projected potential of an atom gives the electron wave a corresponding phase profile without redistributing the amplitude of the wave, and thus the difference between the incident and the exit wave is only a phase shift $e^{i\phi(\vec{r})}$ that describes the potential profile of the atom.

As mentioned in the last section, P_e can be minimized by maximizing the orthogonality between the two transmitted waves. The square inner product of the states of the two waves is thus found to be:

$$\begin{aligned} |\langle\psi_1 \mid \psi_0\rangle|^2 &= \left| \int \psi_{in}^*(\vec{r})\psi_{in}(\vec{r})e^{i\Delta\phi(\vec{r})}\, d\vec{r} \right|^2 \\ &= \left| \int \left|\psi_{in}(\vec{r})\right|^2 e^{i\Delta\phi(\vec{r})}\, d\vec{r} \right|^2 \end{aligned} . \tag{6.22}$$

where $\Delta\phi$ is the difference between the phase profiles of the two atoms, the term $\left|\psi_{in}(\vec{r})\right|^2$ describes the intensity distribution of the incident wave function, and the $*$ sign indicates the complex conjugate. Since the total intensity is fixed, the integral in equation 6.22 can be seen as a sum of a series of complex number $e^{i\Delta\phi}$ with a weight distribution $\left|\psi_{in}(\vec{r})\right|^2$.

For the minimization, we can approach the problem in its simplest form, where we need to reduce the square sum of three complex numbers $a = |a|$, $b = |b|e^{i\phi_b}$, and $c = |c|e^{i\phi_c}$

by changing their magnitude. Here, complex numbers a and c are assumed to have the largest phase difference. Since the total intensity of the electron wave is fixed, the equivalent part in the simplified case, which is the sum $|a| + |b| + |c|$, is kept equal to 1 (see figure 6.5). The square sum of the three numbers is then found to be:

$$\begin{aligned}
&|a+b+c|^2 \\
=&\Re\{a+b+c\}^2 + \Im\{a+b+c\}^2 \\
=&|a|^2 + |b|^2 + |c|^2 + 2|ab|\cos\phi_{ab} + 2|ac|\cos\phi_{ac} + 2|bc|\cos\phi_{bc} \\
=&(|a| + |b| + |c|)^2 - (2|ab|\left(1-\cos\phi_{ab}\right) + 2|ac|\left(1-\cos\phi_{ac}\right) + 2|bc|\left(1-\cos\phi_{bc}\right))
\end{aligned} \tag{6.23}$$

As $|a| + |b| + |c|$ amounts to 1, the square sum is minimized when $2|ab|\left(1-\cos\phi_{ab}\right) + 2|ac|\left(1-\cos\phi_{ac}\right) + 2|bc|\left(1-\cos\phi_{bc}\right)$ is maximized. Since $\phi_{ac} > \phi_{ab}$ and $\phi_{ac} > \phi_{bc}$, and in cases where $\left|\phi_{ac}\right| < \pi/2$, the sum can be minimized with $|ac|$ maximized and $|bc|$, $|ab|$ minimized, which naturally happens when $|b|$ is reduced to zero.

The square sum of the remaining two complex numbers with shared weights can be shown as:

$$\begin{aligned}
&|a+c|^2 \\
=&\left||a|e^{-i\frac{\phi_c}{2}} + |c|e^{i\frac{\phi_c}{2}}\right|^2 \\
=&(|a| + |c|)^2\cos^2\frac{\phi_c}{2} + (|a| - |c|)^2\sin^2\frac{\phi_c}{2}
\end{aligned} \tag{6.24}$$

The minimum thus can be found when $|a| = |c| = 1/2$.

This simplified case shows that if weights are distributed among complex numbers, the minimal sum of these numbers can be found by giving each of the two complex numbers with the largest phase difference half of the weight. With this conclusion, the minimum of the integral in equation 6.22 can be achieved if half of the intensity $|\psi_{in}|^2$ is found at the position where the phase difference is at its minimum $\Delta\phi_{\min}$ (most negative) and the other half at the maximum $\Delta\phi_{\max}$ (most positive).

Coming back to the case of probing different atoms with an electron wave, it is now clear that the square inner product in equation 6.22 can be minimized if the electron wave intensity is concentrated on the minimum and maximum points of $\Delta\phi(\vec{r})$, which is the difference between the phase profile of the two atoms. The two points are marked in figure 6.6, where the phase profiles and differences between the profiles of multiple atoms are plotted.

The square inner product $|\langle\psi_1 \mid \psi_0\rangle|^2$ with an incident wave that has half of the intensity at $\Delta\phi_{\min}$ and the other half at $\Delta\phi_{\max}$ is shown to be:

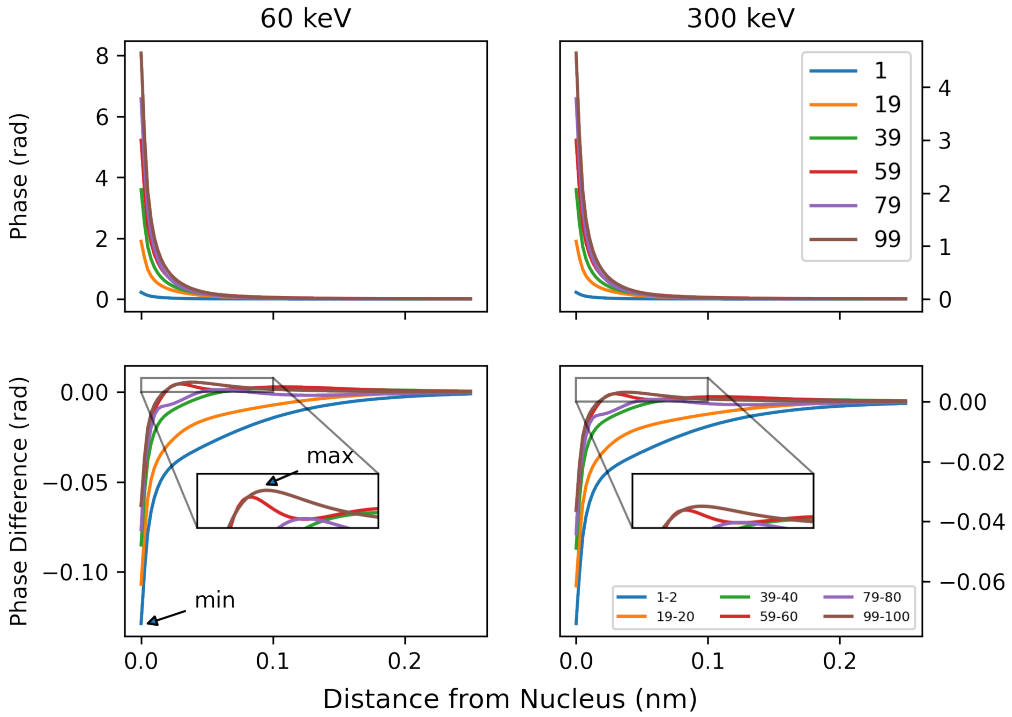


Figure 6.6: The equivalent phase profile of different atoms under 60 and 300 keV. The difference between the profiles of neighboring atoms (n and $n+1$) is also shown. The points of maximal difference with different signs are indicated by i and j. Notice that the equivalent phase decreases when a higher acceleration voltage is used.

$$
\begin{aligned}
|\langle\psi_1 \mid \psi_0\rangle|^2 &= \left| \int \left|\psi_{in}(\vec{r})\right|^2 e^{i\Delta\phi(\vec{r})}\, d\vec{r} \right|^2 \\
&= \left| \frac{|\psi_{in}|^2}{2} e^{i\Delta\phi_{\min}} + \frac{|\psi_{in}|^2}{2} e^{i\Delta\phi_{\max}} \right|^2 .
\end{aligned}
\tag{6.25}
$$

The resulting maximum angle between the states is then found to be:

$$
\theta = \cos^{-1}\left(\left| \frac{1}{2} e^{i\Delta\phi_{\min}} + \frac{1}{2} e^{i\Delta\phi_{\max}} \right| \right) \tag{6.26}
$$

To realize such a waveform that achieves the minimum P_e in differentiation, two criteria must be fulfilled: first, the electron distribution needs to be sharp so that the sampling can be precise and the wave intensity can be concentrated on spots instead of fill areas; second, the location of the distribution needs to be controllable, so that the positions corresponding to $\Delta\phi_{\min}$ and $\Delta\phi_{\max}$ can be accurately sampled. With the criteria as the guideline, the potential design of the electron wave is shown and evaluated in the following section.

6.3 Methods

For the simulation shown in this work, the atomic projected potential is created using the model described by Lobato et. al. [92], which is a part of the MULTEM simulation package [41, 42]. The sampling of the atomic potential is done with a pixel size of 2.5 pm, and in total 2000 points are used to sample each dimension. For the sampling of the convergent beam electron diffraction (CBED) patterns, a maximal scattering angle of approximately 400 mrad is allowed, with a pixel size of 0.4 mrad at 300 keV, and a maximal 973 mrad scattering angle with a pixel size of 1 mrad is used at 60 keV.

6.4 Discussion

Equations 6.25 and 6.26 point out the most efficient waveform to probe and distinguish atoms of different types. While one of the two illuminated points of the phase profile is always at the center of the nuclei, the position of the second point changes according to the atom species, which indicates that the optimal probing waveform cannot be employed without prior knowledge of the expected element type. However, it can be seen that the magnitude of $\Delta\phi_j$ is tiny and not much different from 0, and therefore placing the second half of the intensity in the vacuum far from the nucleus has a negligible effect on the performance of waveform differentiation, thus enabling a uniform design of the optimal probing electron waveform. Another advantage of having half of the wave intensity probing a distant vacuum instead of a part of the potential profile is the reduction of damage. Since damage is only inflicted when the atom and the orbiting electron are

affected by the passing-by fast electron, the partial wave that is sent in the vacuum would not cause any damage to the inspected sample and therefore effectively halves the total damage, or double the tolerable dose level.

This concept agrees with several wave modulation approaches and hardware reported in other works. A binary grating that modifies the amplitude of the wave [174] has been shown capable of splitting the electron probe into a central probe and many side probes. By utilizing one of the side probes to probe the vacuum, a reference phase representing the vacuum can be used to find the phase of the sample. Besides, a reference wave can also be created as some set of concentric rings. Rose has proposed using a concentric phase profile created by aberration and defocus to increase the phase contrast of the diffraction pattern [90], and a matching set of concentric detectors can efficiently pick up this contrast [175]. This idea was later materialized by a fabricated phase plate, which creates the desired phase grating in the shape of concentric rings, and the resulting images have shown improved contrast, especially for light elements [35]. The sensitivity of this imaging mode to the z position of the sample was then further discussed in [176].

Here we summarize the two types of grating and then test their performance for atom differentiation under various imaging conditions. The first type employs 1D gratings (vertical grating in the following discussion) that modify both the amplitude and the phase of the electron wave. Three designs of the incident wave are shown in figure 6.7. The first design is a combination of a homogeneous aperture function with a phase ramp:

$$\tilde{\psi}(\vec{q}) = A(\vec{q}) + A(\vec{q}) \exp\left(i2\pi q_x d\right) \tag{6.27}$$

The function $A(\vec{q})$ describes a round aperture that crops the electron wave, the vector $\vec{q}$ defines the condenser aperture plane where the wave is modified, and d is the slope of the phase ramp. The result of this design can be easily understood as a combination of one shifted probe for sampling the vacuum and another unshifted one to probe the atom, and the distance between the two is determined by the slope of the phase ramp. Despite being conceptually simple, the realization of such a design requires both the amplitude and the phase of the wave to be modified in a controllable manner. Achieving such a design is not trivial work since it necessitates specific materials to create the exact projected potential corresponding to the desired phase shift, and the absorption by the material also needs to be tailored to match the waveform.

The second modification employs a binary amplitude grating to create diffraction that looks like multiple probes on the sample plane:

$$\begin{aligned} \tilde{\psi}(\vec{q}) &= A(\vec{q}) \times G_a(q_x) \\ G_a(q_x) &= \begin{cases} 1, & \text{mod}\ (q_x, d) <= \frac{d}{2} \\ 0, & \text{mod}\ (q_x, d) > \frac{d}{2} \end{cases} \end{aligned} \tag{6.28}$$

where $G_a(q_x)$ is a binary function describing the amplitude grating. Additionally, a similar design realized with phase grating is also shown:

$$\tilde{\psi}(\vec{q}) = A(\vec{q}) \times G_a(q_x)$$
$$G_a(q_x) = \begin{cases} 1, & \text{mod } (q_x, d) <= \frac{d}{2} \\ i, & \text{mod } (q_x, d) > \frac{d}{2} \end{cases} \tag{6.29}$$

Figure 6.7 shows that the distances of the closest side probes are the same for all three setups, but for the two cases that utilize gratings instead of a phase-amplitude ramp, side probes on both sides of the central probe can be found due to the symmetry of the wave modulator. The number of the side probes, in principle, does not affect the performance of atom differentiation as long as they are sufficiently far to not interact with the atom, and they in total occupy half of the intensity of the wave.

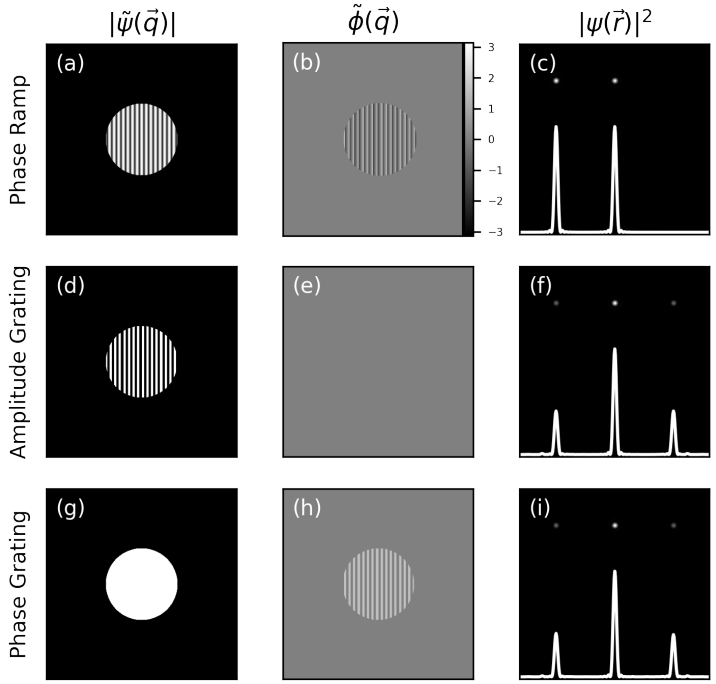


Figure 6.7: Three probe designs with (a,b,c) phase ramp, (d,e,f) amplitude grating, and (g,h,i) phase grating on the aperture plane. The amplitude and phase of the waves on the aperture plane are shown in the first and second columns, respectively. The resulting probe functions on the sample plane are shown in the third column, with a line profile across the center of the probes.

The second type of design can be seen as an extension of the first type, where the reference wave is not created as side probes but as rings surrounding the central probe by utilizing a radial grating, as shown in figure 6.8

To demonstrate the ability of the designed waveform to distinguish atoms, a series of simulated exit waves are generated by positioning the incident wave at the center of the illuminated atom. In figure 6.9, the orthogonality between exit waves resulting from

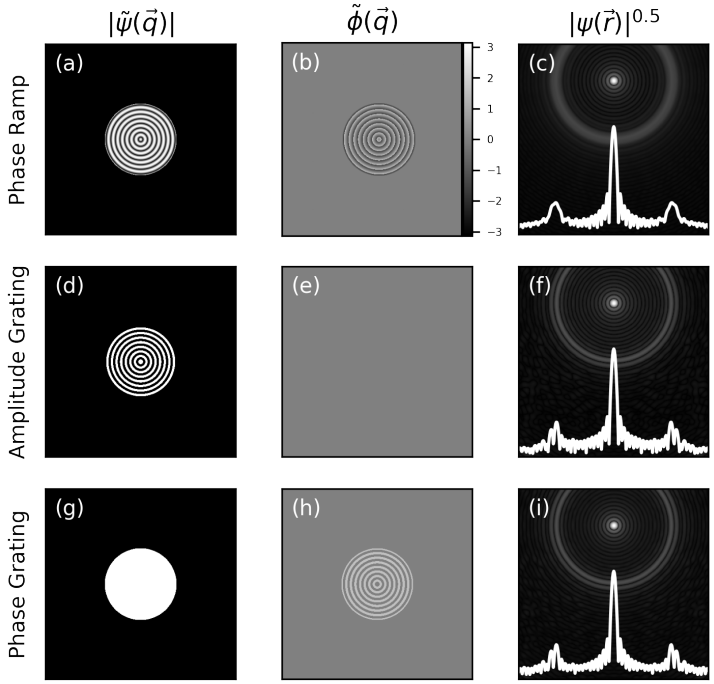


Figure 6.8: Designs with (a,b,c) phase ramp, (d,e,f) amplitude grating, and (g,h,i) phase grating on the aperture plane to create an extra ring on the sample plane. Notice that for (c,f,i) the square root of the amplitude of the probe is shown, instead of the intensity, due to the ring being much dimmer than the side probes.

atoms of neighboring atomic number, i.e. $Z = i$ and $Z = i+1$, is represented by the angle θ between their corresponding vectors. The ramp and gratings are designed to create side probes or rings 1 nm away from the central probe, and from figure 6.6 it can be seen that the projected potential is already effectively zero at a shorter distance. For the calculated angle between the exit waves, the choice of the wave modulator (amplitude grating, phase grating, or combined) does not make noticeable difference, and thus the three results from modified waves are represented together by the gray curves. The maximal achievable angles θ predicted by equation 6.26 are presented by the black curves, and the results of the unmodified probes, i.e. probes formed by hollow round aperture without any phase profile, are shown in dashed curves. The angles between the exit waves are calculated under four imaging conditions. With an acceleration voltage of 300 kV, convergence angles of 25, 100, and 150 mrad are chosen to show the effect of probe sharpness. Besides, one example of 60 kV with 150 mrad convergence angle is also demonstrated, exhibiting the effect of the kinetic energy of the beam.

Figure 6.9 shows that the type of the grating (vertical or circular) does not make a noticeable difference in the orthogonality of the exit wave. This is expected since with both designs the central probe interacts with the nucleus, and the other half of the intensity is distributed somewhere where the projected potential of the atom does not affect it, and by the way that θ is calculated, both waveforms should result in the same angle. By increasing the convergence angle, the angle between the exit waves approaches

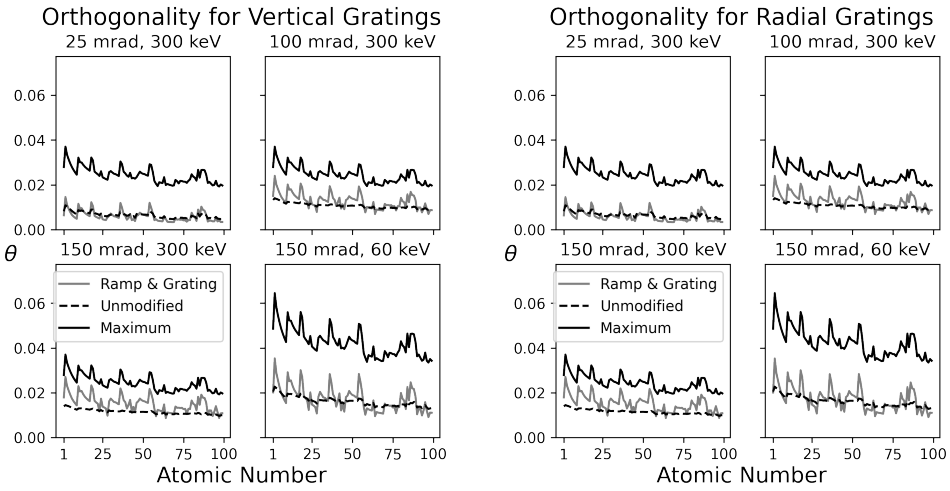


Figure 6.9: Orthogonality of the exit waves modified with both gratings and phase ramps. Three curves are shown in each panel: The black curve indicates a theoretical maximum angle, the gray curve draws the angle calculated from the exit waves of modified incident waves, and the dashed curve represents the result from the unmodified probe. Different conditions, including various convergence angles and energies, are tested.

the predicted maxima. This reflects the fact that with a small convergence angle, the formed electron probe is not sharp enough to sample only the desired position of the potential. At 60 keV, the theoretical maximal angle is greater than at 300 keV. This is also expected since the equivalent phase shift of the same potential is higher when the electron energy is lower. However, it is worth pointing out that the curve of the modified wave overlaps strongly with the unmodified probe, even at a 150 mrad convergence angle. This is due to the broadening of the electron distribution on the sample plane at a lower voltage. At such a condition, an even higher convergence angle is required to create a sharp electron probe and fulfill the first criterion.

To calculate the P_e for atom differentiation, the resulting CBED patterns by the transmitted waves through various types of elements are compared with each other. These patterns are assumed to each consist of 100 electrons and are recorded by a pixelated detector. As the number of electrons collected in the area of a small pixel of the detector can be described by Poisson distribution, the P_e to distinguish two waveforms based on the recorded CBED patterns can be calculated by utilizing the central limit theorem [67].

$$
\begin{aligned}
P_e &= \frac{1}{2}\left[C\left(\frac{-\mu_1}{\sigma_1}\right) + C\left(\frac{-\mu_0}{\sigma_0}\right)\right] \\
\mu_n &= \sum_i^N \left(\lambda_{n,1} \ln \frac{\lambda_{1,i}}{\lambda_{0,i}} - \lambda_{1,i} + \lambda_{0,i}\right) \\
\sigma_n^2 &= \sum_i^N \lambda_{n,1} \left(\ln \frac{\lambda_{1,i}}{\lambda_{0,i}}\right)^2
\end{aligned}
\tag{6.30}
$$

where $C(x)$ is the cumulative distribution function of the standard Gaussian distribution, and $\lambda_{n,i}$ is the expected intensity at pixel i if the interacting atom is of type n.

With that, the P_e in distinguishing atom species of neighboring atomic number based on the measurement of the transmitted wave can be calculated. The same set of imaging conditions considered in the discussion of exit wave orthogonality is used again, with each CBED presented with 100 electrons. The results are shown in figure 6.10.

The P_e follows a similar trend to that of the orthogonality: a larger convergence angle yields a lower P_e and thus a higher accuracy when distinguishing between the waves and corresponding atoms In addition, following the argument that higher dose usage is allowed by probing the vacuum, results with twice the dose are shown as dotted lines, which should cause the same level of damage compared with a regular unmodified electron probe.

Besides, apparent dips in the P_e can be found in the modified wave cases, even at the smallest convergence angle of 25 mrad. Interestingly, these dips become more obvious for the modified waves at a larger convergence angle, but the exact opposite behavior is exhibited by the unmodified waves. The cause of these dips is the orbital contraction that happens as the orbital is gradually filled. Whenever an electron is filled in a new orbital, the distribution of the negative charge will expand, and thus the positive charge at the nucleus is more exposed. This description matches the positions of the dips that are found at the atomic numbers corresponding to each noble element, which marks the onset of the filling of a new orbital. From the perspective of the projected potential, this behavior of the orbital electrons results in a jump of the phase at the center, but the shape of the local phase profile is not greatly changed. Therefore, when sampling with a single electron probe (unaberrated, convergence angle larger than 100 mrad) that is situated at the center of the atom, this difference made by the orbital contraction is not noticeable since the change only happens in the phase value. On the other hand, the reference wave records the phase level of the vacuum, and therefore contrast variation will be created when the two parts of the waves with different phase levels are joined on the diffraction plane. The extra contrast created by the phase reference thus lowers the probability of error in distinguishing waves and stresses the importance and benefit of this holography-like sampling scheme.

The most common imaging modes such as ADF or ABF rely on the integration of the electron counts on a specific region in the detector plane. The same or similar method can also be applied to the modulated waves so that the response of the probed atom or sample can be represented by a single value. For waves modified by the phase grating, the contrast is created by subtracting the intensity of the partial wave that was assigned with 0 phase from the one that was assigned with a phase of $\pi/2$ (see equation 6.29), as described in [35]. For waves modified by the amplitude grating, as well as the unmodified wave, the intensity is integrated from the DF region, which covers every part of the detector plane that is blocked by the aperture, including the grating of the wave modulator. The resulting contrast of the atoms created by these designed waves is shown in figure 6.11.

The electron count found in a defined region in the detector plane can be seen as a result of a binary trial, which defines success and failure based on whether the electron falls within the region. The measured intensity thus follows the binomial distribution, with the probability of success equal to the normalized intensity integrated from the area and the number of trials equal to the dose. The variance of a binomial distribution is:

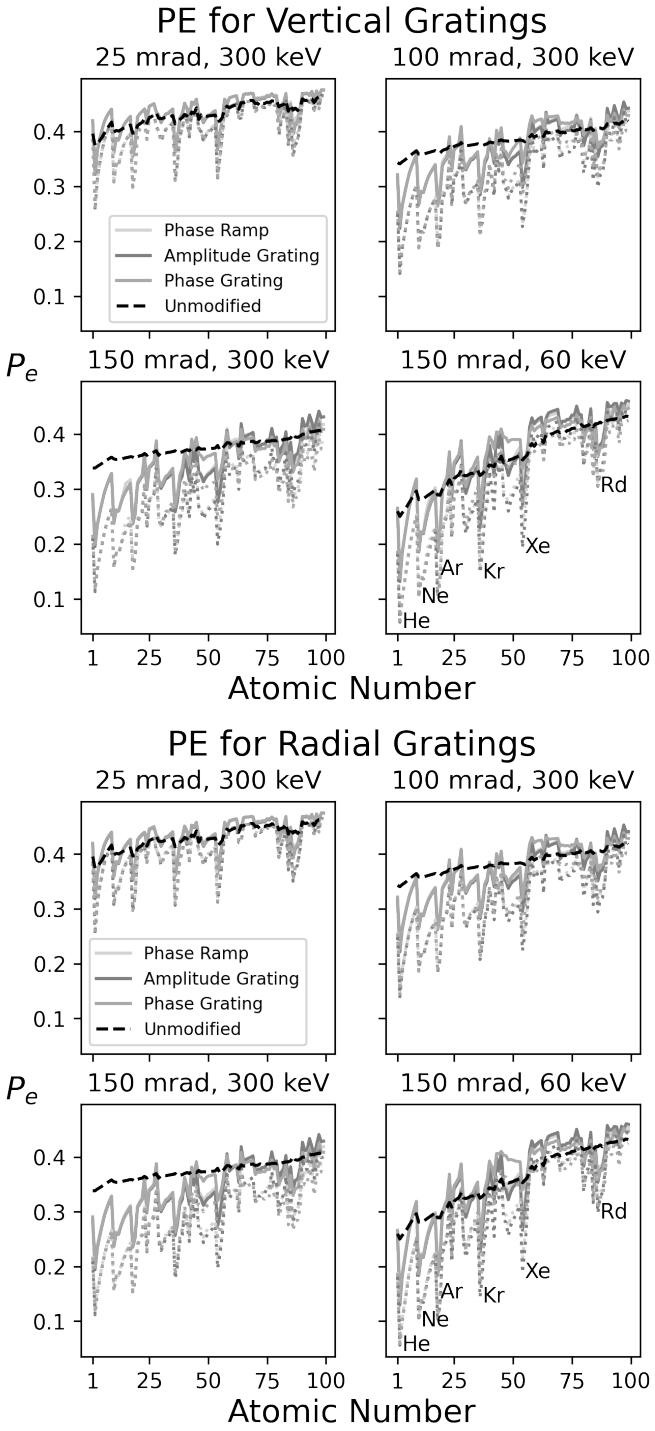


Figure 6.10: Probability of error of atom differentiation. The tested CBED patterns are generated from interaction with atoms that differ in atomic number by only 1. The dotted gray curves represent the P_e measured with twice the dose, while descriptions of curves of other colors are included in the legend.

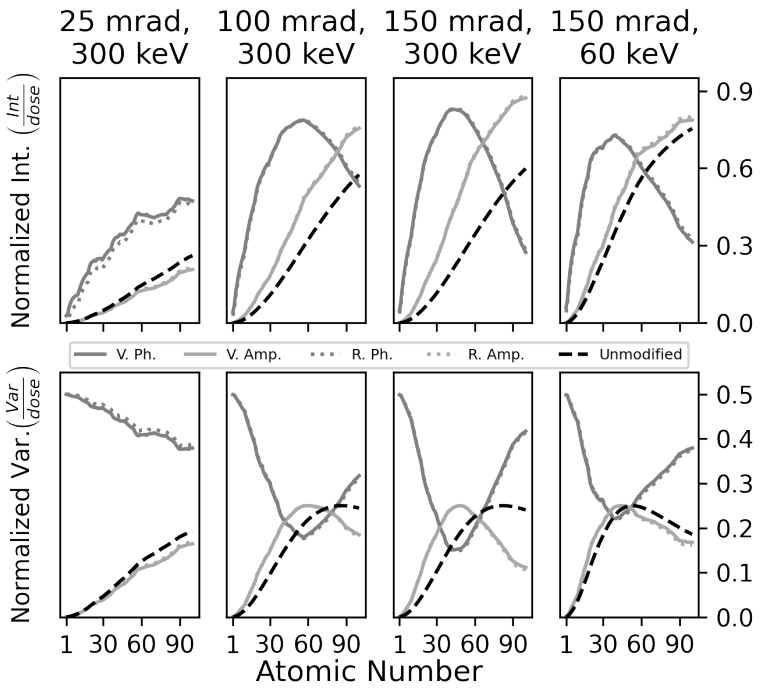


Figure 6.11: Contrast of different wave designs against the atomic numbers. Both the normalized intensity and the normalized variance are shown.

$$\text{var} = np(1-p) \tag{6.31}$$

where n is the number of trials and p is the probability of success. Since the variance is proportional to the dose, the normalized variance, which is only a function of the success rate, is then shown in the figure. For the contrast created by phase grating, which is taken as the difference between the intensities integrated from two regions, the intensity no longer follows the binomial distribution, and thus the variance cannot be calculated this way. With the assumption that the number of trials is large enough to approximate the individual intensity distribution of the two regions as Gaussians, the variance of the differential signal between the two can thus be found as the sum of the variances of the individual signals.

Figure 6.11 shows that waves modified by the phase grating possess high sensitivity to atomic number at the low Z regime, yet the variance is also very high. For cases with the larger convergence angle, the phase grating indeed shows reduced variance at the atomic number around 50; however, the sensitivity to different atom species is also much reduced, even exhibiting reversed contrast against increasing Z. This behavior indicates that the phase grating may not be an ideal way of wave modification if both dose reduction and improved contrast from weak scatterers are desired. For the electron probes formed with amplitude grating, an improved contrast sensitivity compared to ADF is shown at large convergence angles. Even though the contrast response may not be as strong as the one achieved with phase grating, the variance is much lower, and thus

a higher signal-to-noise ratio can be expected.

Since the main advantage of the designed waveform is the sensitivity to the projected potential value compared to the vacuum rather than only the local difference of it, a series of transmitted waves with different modulation are simulated, assuming interacting with a 1 nm square homogeneous phase object. The phase shift induced by the phase object ranges from 0 to 2π, and the resulting contrast under different imaging conditions is shown in figure 6.12

In this simulated experiment, no local phase variation can be found within the phase object. For sharp unmodified electron probes, this results in almost unnoticeable variation in the transmitted wave, and thus almost no signal in the dark field can be found, with the exception of some very weak contrast for 25 mrad probe at phase shift around π, which is a result of the long tail of the probe at smaller convergence angle. The result acquired with amplitude grating, on the contrary, shows contrast directly related to the interference between the central probe and the side probes, as described in equation 6.26, and thus the contrast peaks at phase shift equal to π. Contrast variation as a result of phase-grating modified incident wave shows a more complicated relationship with phase shift. The contrast here is a result of the intensity difference between the two regions that originally possess a phase of 0 or $\pi/2$. When the interacting object creates a phase shift between 0 and π, constructive interference is created in the region with $\pi/2$ original phase shift, and when the object creates a phase shift between π and 2π, constructive interference appears in the opposite region. As a consequence, a contrast reversal can be found at the phase shift of π, as seen in figure 6.12. Furthermore, it is also shown that vertical grating of both phase and amplitude shows stronger contrast than the radial ones. This is due to a higher residual amplitude found inside the ring, which consequently reduces the intensity distribution that samples the vacuum and thus the condition for the optimal wave design is not strictly met.

This hypothetical experiment shows that the modified waveforms of both phase and amplitude grating are very effective in imaging phase objects of a relatively flat phase profile. However, it should be noted here that to create side probes or rings at a meaningful distance, the density of the grating needs to be very high; otherwise, unavoidable interaction between the reference wave with the sample would reduce the benefit.

6.5 Conclusion

In this work, the maximal difference that can be created in the exit waves by different atom species is introduced together with the concept of orthogonality between the quantum states describing the electron waves. As suggested by equations 6.15 and 6.26, a waveform that interacts with only the most different part of the projected potential and the vacuum can be utilized to distinguish various types of atoms with the maximum efficiency. Several designs are thus tested and indeed show the said improvement. The P_e for exit wave or atom differentiation is lowered with the designed waveform. The effect of the modification is especially pronounced when the convergence angle is large, as the sampling of the projected potential can be done more precisely. The contrast response against the atomic number is also enhanced when the modified waves are used. Finally, the tests with flat phase objects show that for the modified waves, the intensity variation

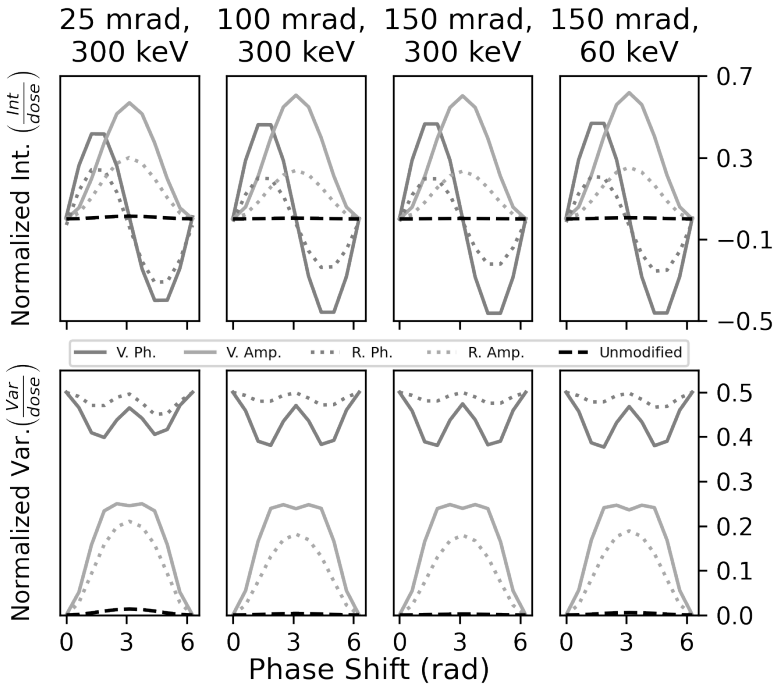


Figure 6.12: Contrast of different wave designs against homogeneous phase objects that create different levels of phase shift.

in the CBED does not rely on local phase variation thanks to the reference wave that probes the vacuum, which is an advantage that cannot be exploited by sharp electron probes formed by regular apertures.

For the sake of the discussion in this work, the imaging condition used to demonstrate the benefit of waveform design may be rather difficult to achieve for the technology nowadays, but it also provides valuable knowledge of do's and don'ts for future development of electron microscopy. For example, figures 6.10 and 6.11 imply that the pursuit of a higher convergence angle and sharper electron probe without the employment of wave modification will face its limit in the improvement of the contrast and sensitivity to atom species. In contrast, with the current probe correction technology or even optical systems without a probe corrector, the efficiency in imaging weak scatterers can already be improved with a designed waveform.

Chapter 7

General Conclusion and Outlook

This thesis is dedicated to the development of methods that utilize the result of elastic scattering of the fast electrons to inspect the sample. Two imaging methods are thus proposed, as well as potential novel techniques based on not only the elastic scattering events but also the concept of waveform design. These works are not established on new understandings of the scattering event, but the advancement of technology, including faster detectors, more powerful computers, and the creation of programmable wave modulators, makes new playing grounds on which once unrealistic ideas may now be realized.

The first chapter of the thesis writes basic knowledge about the electron beam, including its interaction with materials, its creation, and factors that affect its properties. With this understanding, the imaging mechanisms of a few common imaging modes of TEM are discussed, including the corresponding hardware that collects the necessary signals.

In the second chapter, several important methods involved in the making of this thesis are introduced. It begins with an introduction to the modeling of the elastic scattering events and the propagation of electron waves in the material and followed by several renowned imaging methods based on the wave nature of the transmitted electron and the phase shift created by the electron-sample interaction. In addition, statistical analyses of TEM data are discussed, with a few examples related to the investigation of sample thickness variation.

In the third chapter, the riCOM imaging method is discussed. The formula and the contrast transfer function are presented in detail, and the benefits of the imaging method are also showcased, including the sensitivity to light elements, the low-dose imaging capability, and the low computation demands, which leads to excellent processing speed. The method is implemented with a streamed data pipeline from the detector, and with that, real-time imaging on a microscope has been demonstrated. Live feedback that allows operations, such as sample search, focusing, and astigmatism correction to be performed at a lower dose is therefore achieved.

The fourth chapter introduces an approach to achieving real-time imaging with AI. A CNN is trained to perform waveform retrieval with data collected only from a few probe positions in a STEM experiment, and the predicted exit wave is then used to reconstruct the transmission function localized at the probe position, which is closely related to the sample potential. It is shown that this imaging method possesses the ability to achieve

super-resolution with regular or higher doses and also generates an interpretable output at a lower minimum dose limit compared to traditional imaging modes. Besides, the contrast of the reconstruction shows that both light and heavy elements can be imaged without sacrificing the sensitivity to atom column density, which means that different element types and column thicknesses can be intuitively distinguished with the resulting contrast.

The fifth chapter describes a device that introduces a programmable phase plate that can create phase shifts to different parts of the passing electron wave. The phase sensitivity to the applied voltage and its beam shaping ability has been shown, as well as several possible applications of the device in TEM. Despite that most of these applications are only realized on a proof-of-concept level or demonstrated with simulations, they serve as important examples of how TEM experiments can benefit from the ability to modulate waves at a high degree of freedom.

The sixth chapter discusses the possibility of differentiating elements with elastically scattered electrons and the optimal waveform design for this purpose. With the derivation based on distinguishing quantum states, the result shows that our search for optimal wave design arrives at a similar conclusion as a few other reported works. The ideal waveform is shown to be composed of two coherent parts: the central probe that is used to sample the imaged target and a partial wave with an equal intensity that samples the vacuum and provides a referencing phase level. It is shown with simulated examples that the improvement in the accuracy of the differentiation is significant when the central probe is sufficiently sharp to interact with only the very center of an atom. The current probe correction technology cannot create a probe with adequate sharpness fulfilling the criterion to optimize the differentiation of targets as small as an atom, yet it is shown that for the imaging of a broad and flat phase object, the reference wave can indeed enhance the contrast between two such phase objects of difference phase level.

The third and fourth chapters are dedicated to imaging methods that provide live feedback. A dose-efficient imaging method with live feedback is necessary to enable true low-dose electron microscopy as it reduces electron-inflicted damage during various operations on the microscope. For both proposed imaging methods, there is room for improvement. As mentioned in chapter 3, riCOM's algorithm does not take into account the radial distribution of the electron. Not utilizing every information from the scattering is obviously sub-optimal when it comes to achieving maximum efficiency. It also makes riCOM not yet capable of completely replacing other common imaging methods, such as HAADF, since it is not sensitive to features of comparable size of the probe. This problem can be partially compensated by simultaneous imaging with other modes, for example, ADF or second-moment imaging. Nevertheless, it is worth exploring the possibility of creating an algorithm that utilizes information including the COM shift, the radial distribution of the diffraction pattern, and possibly more while maintaining high efficiency and low computational demand. Like other applications utilizing machine learning, the CNN used in AIRPI can always be trained better with more data. For example, data generated with a larger variety of sample types and optical conditions can help make the reconstruction more robust and applicable to complex situations. Improvements in other aspects of the CNN design may also benefit the performance. For instance, AIRPI is currently trained with 2D images of the CBED patterns, which can make the reconstruction process straightforward since the retrieved waveform is built with 2D arrays given by the CNN. However, such data input of the CNN is also incapable of taking

full advantage of the data-driven detectors since frames need to be built from the events before the neural network can work on the data. Besides, the symmetry of the data is not utilized. For example, if data are rotated 90 degree for both the 3×3 CBEDs cube and the 2D arrays of each individual CBED, the retrieved waveform should be the same as the original with a 90-degree rotation. Such symmetry may be exploited with a different network design, and if so, the complexity of the CNN may be reduced, and the efficiency of the reconstruction can thus be increased. In addition to the improvement of the algorithm for better imaging results, it is also worth exploring how TEM applications can take advantage of the live updates and the low-dose imaging efficiency. For example, non-delayed feedback is helpful not only to a human operator but can be arguably even more beneficial for the automation of the TEM. An image analysis program can process the real-time reconstructed images acquired under a low-dose condition and give a prediction of the following step/operation for the acquisition of the desired data. This information can then be shared with, e.g., the scanning engine, the controller of the lenses, and even the probe corrector to achieve complete and adaptive TEM automation with maximum efficiency in terms of dose usage and time consumption.

For the phase device introduced in chapter five, to realize the expected applications and to achieve what cannot be done with current TEM, it requires further improvement in some aspects. Without the ability to constantly monitor the shape of the electron probe in a TEM, the desired probe shape can only be assumed with prior knowledge about the characteristics of the phase plate, which necessitates precise and stable phase modulation abilities that are not yet the case based on experience. To make such improvements, new ways to characterize and inspect possible defects of the device are crucial. With the current workflow, many aspects of the phase plate's performance can only be learned when the phase plate is inserted in a TEM, which is only possible once the manufacturing of the phase plate is nearly finished. This restricts corrections and improvements that may be done once a problem is noticed and also limits the understanding of the problem since the response of the phase plate can not be safely probed with the high-energy electron probe of a TEM. It seems that a smooth production of a stable phase plate is only possible if a comprehensive performance examination can be done at an elementary level, which may or may not rely on a TEM. This would enable inspection and amendment of the device at different stages of manufacturing, reducing non-ideal behaviors of the final product.

When a stable and highly adaptive waveform can be easily generated, improvements and implementation of the optimal wave design can be realized and tested on a TEM. Within the discussion of the sixth chapter, the advantage of using a reference wave boils down to increasing contrasts in an individual CBED collected from a single probe position. As demonstrated in the discussion, for imaging methods that directly utilize such contrasts, such as ADF, improvements in sensitivity to the induced phase shift can thus be expected. For other imaging modes that perform phase or waveform retrieval based on the recorded results of scattering events, enhanced contrast should in theory also improve the reconstruction outcomes. Further investigation into a reconstruction algorithm that utilizes the information provided by the reference wave, as well as the optimal imaging condition, not only has the potential to create a new dose-efficient imaging method but also possibly bring new insight into the design of the optimal waveform and sampling strategy. By combining the wave modulation and the formulation in quantum mechanics, the optimal measurement scheme for other material-related properties may be created. Some of the current TEM experiments include procedures to first acquire a

high-resolution image of the investigated sample, from which the desired information is then extracted. However, such an approach may require a higher dose than necessary due to the dense sampling for the intermediating step of image formation. Instead, if the investigated property of the material is not very localized, one or a series of broader illumination with designed waveforms may be more efficient in obtaining the necessary information describing a general property of the inspected material.

Most of the works included in this thesis come with a certain level of realization or on-machine implementation, with the only exception of chapter 6, where the realization of the concept has already been done and reported in other publications. Indeed, more than often a proposed method is only accepted once it is tested on actual data or run on an actual machine, and thus implementation, in that sense, is an inseparable part of creation. While the implementation of an idea can be full of technical issues and hardware limitations, often giving the impression that the project has derailed further and further away from the pursuit of science, the process, on the contrary, is surprisingly educational and often provides clarity of the very concept that is being implemented. During the course of my Ph.D. education, I feel very fortunate to find myself surrounded by people full of experience in materializing and implementing ideas. Besides hoping that the work in the thesis can help make a TEM a better tool for inspecting materials, I also wish to convey a positive attitude toward making tangible creations based on novel ideas and concepts.

Appendix A

List of Abbreviations

2D	Two Dimensional
4D STEM	Four Dimensional Scanning Transmission Electron Microscopy
ABF	Annular Bright Field
ADF	Annular Dark Field
BF	Bright Field
CBED	Convergent Beam Electron Diffraction
CCD	Charge Coupled Device
CDF	Cumulative Distribution Function
COM	Center of Mass
CTEM	Conventional Transmission Electron Microscopy
DED	Direct Electron Detector
DF	Dark Field
DPC	Differential Phase Contrast
EDX	Energy Dispersive X-ray Spectroscopy
EELS	Electron Energy Loss Spectroscopy
FEG	Field Emission Gun
FIB	Focused Ion Beam
GPU	Graphics Processing Unit
GS	Gechberg-Saxton
HAADF	High-Angle Annular Dark-Field
ICL	Integrated Classification Likelihood
iCOM	Integrated Center of Mass
iDPC	Integrated Differential Phase Contrast
LAADF	Low-Angle Annular Dark-Field
PE	Probability of Error
POA	Phase Object Approximation
PMT	Photomultiplier Tube
riCOM	Real-time Integrated Center of Mass
SCS	Scattering Cross-Section
SEM	Scanning Electron Microscope
SSB	Singel Sideband
STEM	Scanning Transmission Electron Microscopy
TEM	Transmission Electron Microscope
WPOA	Weak Phase Object Approximation

Appendix B

Author's Contribution

B.1 Peer-Reviewed Publications

[1] Chu-Ping Yu, Thomas Friedrich, Daen Jannis, Sandra Van Aert, and Johan Verbeeck. Real-time integration center of mass (ricom) reconstruction for 4D STEM. *Microscopy and Microanalysis*, 28(5):1526–1537, 2022. 15, 56

[2] Thomas Friedrich, Chu-Ping Yu, Johan Verbeeck, and Sandra Van Aert. Phase object reconstruction for 4D-STEM using deep learning. *Microscopy and Microanalysis*, 29(1):395–407, 2023. 15

[3] Thomas Friedrich, Chu-Ping Yu, Johan Verbeek, Timothy Pennycook, and Sandra Van Aert. Phase retrieval from 4-dimensional electron diffraction datasets. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3453–3457. IEEE, 2021.

[4] Annelies De Wael, Annick De Backer, Chu-Ping Yu, Duygu Gizem Sentürk, Ivan Lobato, Christel Faes, and Sandra Van Aert. Three approaches for representing the statistical uncertainty on atom-counting results in quantitative ADF STEM. *Microscopy and Microanalysis*, 29(1):374–383, 2023.

[5] Duygu Gizem Şentürk, Chu-Ping Yu, Annick De Backer, and Sandra Van Aert. Atom counting from a combination of two adf stem images. *Ultramicroscopy*, 255:113859, 2024.

[6] Nadine J Schrenker, Chu-Ping Yu, Thomas Friedrich, Nikolaos Livakas, Annick De Backer, Daen Jannis, Armand Béché, Maarten Roeffaers, Johan Hofkens, Liberto Manna, Johan Verbeeck, Sandra Van Aert, and Sara Bals. Octahedral tilt and quantitative low-dose imaging of formamidinium lead bromide nanocrystals via 4D-STEM.

[7] Chu-Ping Yu, Francisco Vega Ibáñez, Armand Béché, and Johan Verbeeck. Quantum wavefront shaping with a 48-element programmable phase plate for electrons.

[8] Qi Zhang, Xue-Feng Wang, Shu-Hong Shen, Qi Lu, Xiaozhi Liu, Haoyi Li, Jingying Zheng, Chu-Ping Yu, Xiaoyan Zhong, Lin Gu, Tian-Ling Ren, and Liying Jiao. Simultaneous synthesis and integration of two-dimensional electronic components. *Nature Electronics*, 2(4):164–170, 2019.

[9] Ke Deng, Mingzhe Yan, Chu-Ping Yu, Jiaheng Li, Xue Zhou, Kenan Zhang, Yuxin Zhao, Koji Miyamoto, Taichi Okuda, Wenhui Duan, Yang Wu, Xiaoyan Zhong, and Shuyun Zhou. Crossover from 2D metal to 3D dirac semimetal in metallic PtTe2 films with local Rashba effect. *Science Bulletin*, 64(15):1044–1048, 2019.

[10] Yang Zhang, Wenbin Wang, Wandong Xing, Shaobo Cheng, Shiqing Deng, Manuel Angst, Chu-Ping Yu, Fanli Lan, Zhiying Cheng, David Mandrus, Brian Sales, Jian Shen, Xiaoyan Zhong, Nyan-Hwa Tai, Rong Yu, and Jing Zhu. Effect of oxygen interstitial ordering on multiple order parameters in rare earth ferrite. *Physical review letters*, 123(24):247601, 2019.

[11] Ping-Luen Ho, Chu-Ping Yu, Qiqi Zhang, Kyung Song, James P. Buban, Si-Young Choi, Rafal E. Dunin-Borkowski, Joachim Mayer, Nyan-Hwa Tai, Jing Zhu, Lei Jin, and Xiaoyan Zhong. Effect of cation ratio and order on magnetic circular dichroism in the double perovskite Sr2Fe1+ xRe1-xO6. *Ultramicroscopy*, 193:137–142, 2018.

[12] Xinfeng Chen, Soma Higashikozono, Keita Ito, Lei Jin, Ping-Luen Ho, Chu-Ping Yu, Nyan-Hwa Tai, Joachim Mayer, Rafal E. Dunin-Borkowski, Takashi Suemasu, and Xiaoyan Zhong. Nanoscale measurement of giant saturation magnetization in α''-Fe16N2 by electron energy-loss magnetic chiral dichroism. *Ultramicroscopy*, 203:37–43, 2019.

[1, 2] The first two authors have contributed equally.
[6, 7] are in preparation.
[8, 9, 10, 11, 12] are not related to the work of this thesis.

B.2 Conference Appearances

ATOM2020
Contribution: Talk
At: Online
Topic: Optimized Probe Design for Atom Identification

ATOM2021
Contribution: Talk
At: Online
Topic: Image Reconstruction from 4D STEM Dataset Using Neural Network

MC2021
Contribution: Talk
At: Online
Topic: Phase Object Reconstruction from 4D STEM Dataset with Neural Network

MC2023
Contribution: Poster
At: Darmstadt, Germany
Topic: Adaptive Probe Correction with Electrostatic Phase Plate

PICO2022
Contribution: Poster
At: Vaalsbroek, Nederland
Topic: Real-Time Integration Center of Mass (RiCOM) Reconstruction for 4D-STEM

EAMC2023
Contribution: Talk
At: Online
Topic: Real Time 4D STEM Reconstruction - Two Recent Approaches

B.3 Other Publications

EP4075126A1
Type: Patent
Title: PHASE RETRIEVAL IN ELECTRON MICROSCOPY
Role: Inventor (40%)
URL: https://worldwide.espacenet.com

Can AI do your Ptychography?
Type: Magazine Article
Author: T. Friedrich, C.-P. Yu, J. Verbeeck, and S. Van Aert
Journal: Imaging & Microscopy

Real-Time Integration Center of Mass (riCOM) Reconstruction for 4D STEM
Type: Highlight
Author: C.-P. Yu, T. Friedrich, D. Jannis, S. Van Aert, and J. Verbeeck
Journal: Microscopy Today

Bibliography

[1] J. I. Goldstein, D. E. Newbury, J. R. Michael, N. W. Ritchie, J. H. J. Scott, and D. C. Joy, *Scanning electron microscopy and X-ray microanalysis*. springer, 2017. 2

[2] R. F. Egerton, *Electron energy-loss spectroscopy in the electron microscope*. Springer Science & Business Media, 2011. 2

[3] Y. Aharonov and D. Bohm, "Significance of electromagnetic potentials in the quantum theory," *Physical Review*, vol. 115, no. 3, p. 485, 1959. 2

[4] G. Catalan, J. Seidel, R. Ramesh, and J. F. Scott, "Domain wall nanoelectronics," *Reviews of Modern Physics*, vol. 84, no. 1, p. 119, 2012. 3

[5] M. Tanase and A. K. Petford-Long, "In situ TEM observation of magnetic materials," *Microscopy Research and Technique*, vol. 72, no. 3, pp. 187–196, 2009. 3

[6] K. Harada, "Lorentz microscopy observation of vortices in high-T c superconductors using a 1-mv field emission transmission electron microscope," *Microscopy*, vol. 62, no. suppl_1, pp. S3–S15, 2013. 3

[7] N. Nagaosa and Y. Tokura, "Topological properties and dynamics of magnetic skyrmions," *Nature nanotechnology*, vol. 8, no. 12, pp. 899–911, 2013. 3

[8] L. E. Franken, K. Grünewald, E. J. Boekema, and M. C. Stuart, "A technical introduction to transmission electron microscopy for soft-matter: Imaging, possibilities, choices, and technical developments," *Small*, vol. 16, no. 14, p. 1906198, 2020. 3

[9] 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. 6

[10] R. Egerton, P. Li, and M. Malac, "Radiation damage in the TEM and sem," *Micron*, vol. 35, no. 6, pp. 399–409, 2004. 7

[11] H. Sawada, T. Sasaki, F. Hosokawa, and K. Suenaga, "Atomic-resolution STEM imaging of graphene at low voltage of 30 kv with resolution enhancement by using large convergence angle," *Physical Review Letters*, vol. 114, no. 16, p. 166102, 2015. 7, 10

[12] L. Frank, J. Nebesářová, M. Vancová, A. Paták, and I. Müllerová, "Imaging of tissue sections with very slow electrons," *Ultramicroscopy*, vol. 148, pp. 146–150, 2015. 7

[13] M. J. Peet, R. Henderson, and C. J. Russo, "The energy dependence of contrast and damage in electron cryomicroscopy of biological molecules," *Ultramicroscopy*, vol. 203, pp. 125–131, 2019. 7

[14] L. Möller, G. Holland, and M. Laue, "Diagnostic electron microscopy of viruses with low-voltage electron microscopes," *Journal of Histochemistry & Cytochemistry*, vol. 68, no. 6, pp. 389–402, 2020. 7

[15] C. Ricolleau, J. Nelayah, T. Oikawa, Y. Kohno, N. Braidy, G. Wang, F. Hue, L. Florea, V. Pierron Bohnes, and D. Alloyeau, "Performances of an 80–200 kv microscope employing a cold-feg and an aberration-corrected objective lens," *Microscopy*, vol. 62, no. 2, pp. 283–293, 2013. 7

[16] B. Freitag, G. Knippels, S. Kujawa, M. Van der Stam, D. Hubert, P. Tiemeijer, C. Kisielowski, P. Denes, A. Minor, and U. Dahmen, "First performance measurements and application results of a new high brightness schottky field emitter for HR-S/TEM at 80-300kv acceleration voltage," *Microscopy and Microanalysis*, vol. 14, no. S2, pp. 1370–1371, 2008. 7

[17] J. Verbeeck, A. Béché, and W. Van den Broek, "A holographic method to measure the source size broadening in STEM," *Ultramicroscopy*, vol. 120, pp. 35–40, 2012. 8

[18] B. Freitag, S. Kujawa, P. Mul, J. Ringnalda, and P. Tiemeijer, "Breaking the spherical and chromatic aberration barrier in transmission electron microscopy," *Ultramicroscopy*, vol. 102, no. 3, pp. 209–214, 2005. 8

[19] M. Haider, P. Hartel, H. Müller, S. Uhlemann, and J. Zach, "Information transfer in a TEM corrected for spherical and chromatic aberration," *Microscopy and Microanalysis*, vol. 16, no. 4, pp. 393–408, 2010. 8

[20] P. Hartel, H. Rose, and C. Dinges, "Conditions and reasons for incoherent imaging in STEM," *Ultramicroscopy*, vol. 63, no. 2, pp. 93–114, 1996. 10

[21] S. J. Pennycook and P. D. Nellist, *Scanning transmission electron microscopy: imaging and analysis*. Springer Science & Business Media, 2011. 10, 24

[22] S. J. Pennycook, "Z-contrast STEM for materials science," *Ultramicroscopy*, vol. 30, no. 1-2, pp. 58–69, 1989. 10

[23] S. Pennycook and L. Boatner, "Chemically sensitive structure-imaging with a scanning transmission electron microscope," *Nature*, vol. 336, no. 6199, pp. 565–567, 1988. 10, 70, 93

[24] J. M. LeBeau, S. D. Findlay, L. J. Allen, and S. STEMmer, "Standardless atom counting in scanning transmission electron microscopy," *Nano letters*, vol. 10, no. 11, pp. 4405–4408, 2010. 10

[25] Z. Yu, D. A. Muller, and J. Silcox, "Study of strain fields at a-Si/c-Si interface," *Journal of Applied Physics*, vol. 95, no. 7, pp. 3362–3371, 2004. 10

[26] O. F. Ngome Okello, K.-Y. Doh, H. S. Kang, K. Song, Y.-T. Kim, K. H. Kim, D. Lee, and S.-Y. Choi, "Visualization of transition metal decoration on h-BN surface," *Nano Letters*, vol. 21, no. 24, pp. 10562–10569, 2021. 10

[27] T. J. Pennycook, A. R. Lupini, H. Yang, M. F. Murfitt, L. Jones, and P. D. Nellist, "Efficient phase contrast imaging in STEM using a pixelated detector. part 1: Experimental demonstration at atomic resolution," *Ultramicroscopy*, vol. 151, pp. 160–167, 2015. 10, 23

[28] J. Zuo, "Electron detection characteristics of a slow-scan CCD camera, imaging plates and film, and electron image restoration," *Microscopy research and technique*, vol. 49, no. 3, pp. 245–268, 2000. 11

[29] T. Mullarkey, M. Geever, J. J. Peters, I. Griffiths, P. D. Nellist, and L. Jones, "How fast is your detector? the effect of temporal response on image quality," *arXiv preprint arXiv:2212.04050*, 2022. 11, 12

[30] X. Sang and J. M. LeBeau, "Characterizing the response of a scintillator-based detector to single electrons," *Ultramicroscopy*, vol. 161, pp. 3–9, 2016. 11

[31] R. Ballabriga, J. Alozy, G. Blaj, M. Campbell, M. Fiederle, E. Frojdh, E. H. M. Heijne, X. Llopart, M. Pichotka, S. Procz, L. Tlustos, and W. Wong, "The medipix3rx: a high resolution, zero dead-time pixel detector readout chip allowing spectroscopic imaging," *Journal of Instrumentation*, vol. 8, no. 02, p. C02016, 2013. 12

[32] T. Poikela, J. Plosila, T. Westerlund, M. Campbell, M. D. Gaspari, X. Llopart, V. Gromov, R. Kluit, M. van Beuzekom, F. Zappon, V. Zivkovic, C. Brezina, K. Desch, Y. Fu, and A. Kruth, "Timepix3: a 65K channel hybrid pixel readout chip with simultaneous ToA/ToT and sparse readout," *Journal of instrumentation*, vol. 9, no. 05, p. C05013, 2014. 12

[33] M. Malínskỳ, G. van Hoften, O. Vyroubal, V. Doležal, M. Červinková, and L. Yu, "Ceta-f: Scintillator camera for entry level 100kv single particle analysis," *Microscopy and Microanalysis*, vol. 27, no. S1, pp. 1640–1641, 2021. 13

[34] Z. Hlavenková, D. Karia, M. Malínskỳ, D. Němeček, F. Grollios, V. Doležal, O. Sháněl, A. Kotecha, M. Červinková, L. Yu, and A. Mulder, "Thermo scientific™ Tundra Cryo-TEM: 100kv Cryo-TEM dedicated for single particle analysis," *Microscopy and Microanalysis*, vol. 27, no. S1, pp. 1330–1332, 2021. 13

[35] C. Ophus, J. Ciston, J. Pierce, T. R. Harvey, J. Chess, B. J. McMorran, C. Czarnik, H. H. Rose, and P. Ercius, "Efficient linear phase contrast in scanning transmission electron microscopy with matched illumination and detector interferometry," *Nature communications*, vol. 7, no. 1, p. 10719, 2016. 15, 80, 107, 111

[36] C.-P. Yu, T. Friedrich, D. Jannis, S. Van Aert, and J. Verbeeck, "Real-time integration center of mass (riCOM) reconstruction for 4D STEM," *Microscopy and Microanalysis*, vol. 28, no. 5, pp. 1526–1537, 2022. 15, 56

[37] T. Friedrich, C.-P. Yu, J. Verbeeck, and S. Van Aert, "Phase object reconstruction for 4D-STEM using deep learning," *Microscopy and Microanalysis*, vol. 29, no. 1, pp. 395–407, 2023. 15

[38] A. De Backer, G. T. Martinez, A. Rosenauer, and S. Van Aert, "Atom counting in HAADF STEM using a statistical model-based approach: Methodology, possibilities, and inherent limitations," *Ultramicroscopy*, vol. 134, pp. 23–33, 2013. 15, 25, 26

[39] K. Kuramochi, T. Yamazaki, Y. Kotaka, M. Ohtsuka, I. Hashimoto, and K. Watanabe, "Effect of chromatic aberration on atomic-resolved spherical aberration corrected STEM images," *Ultramicroscopy*, vol. 110, no. 1, pp. 36–42, 2009. 15

[40] J. Barthel, "Dr. Probe: A software for high-resolution STEM image simulation," *Ultramicroscopy*, vol. 193, pp. 1–11, 2018. 15

[41] I. Lobato and D. Van Dyck, "MULTEM: A new multislice program to perform accurate and fast electron diffraction and imaging simulations using Graphics Processing Units with CUDA," *Ultramicroscopy*, vol. 156, no. 2015, pp. 9–17, 2015. 15, 59, 65, 106

[42] I. Lobato, S. van Aert, and J. Verbeeck, "Progress and new advances in simulating electron microscopy datasets using MULTEM," *Ultramicroscopy*, vol. 168, no. 2016, pp. 17–27, 2016. 15, 59, 106

[43] E. J. Kirkland, *Advanced computing in electron microscopy*, vol. 12. Springer, 1998. 16, 17, 56

[44] D. Van Dyck, "Image calculations in high-resolution electron microscopy: problems, progress, and prospects," *Advances in electronics and electron physics*, vol. 65, pp. 295–355, 1985. 17

[45] E. Majorovits, B. Barton, K. Schultheiss, F. Perez-Willard, D. Gerthsen, and R. R. Schröder, "Optimizing phase contrast in transmission electron microscopy with an electrostatic (boersch) phase plate," *Ultramicroscopy*, vol. 107, no. 2-3, pp. 213–226, 2007. 18, 80

[46] R. Danev and K. Nagayama, "Transmission electron microscopy with zernike phase plate," *Ultramicroscopy*, vol. 88, no. 4, pp. 243–252, 2001. 18, 80

[47] N. Shibata, S. D. Findlay, T. Matsumoto, Y. Kohno, T. Seki, G. Sanchez-Santolino, and Y. Ikuhara, "Direct visualization of local electromagnetic field structures by scanning transmission electron microscopy," *Accounts of chemical research*, vol. 50, no. 7, pp. 1502–1512, 2017. 19

[48] M. Krajnak, D. McGrouther, D. Maneuski, V. O'Shea, and S. McVitie, "Pixelated detectors and improved efficiency for magnetic imaging in STEM differential phase contrast," *Ultramicroscopy*, vol. 165, pp. 42–50, 2016. 19

[49] K. Müller, F. F. Krause, A. Béché, M. Schowalter, V. Galioit, S. Löffler, J. Verbeeck, J. Zweck, P. Schattschneider, and A. Rosenauer, "Atomic electric fields revealed by a quantum mechanical approach to electron picodiffraction," *Nature communications*, vol. 5, no. 1, p. 5653, 2014. 19, 33

[50] 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. 19

[51] I. Lazić, E. G. Bosch, and S. Lazar, "Phase contrast STEM for thin samples: Integrated differential phase contrast," *Ultramicroscopy*, vol. 160, pp. 265–280, 2016. 20

[52] D. Zhang, Y. Zhu, L. Liu, X. Ying, C.-E. Hsiung, R. Sougrat, K. Li, and Y. Han, "Atomic-resolution transmission electron microscopy of electron beam–sensitive crystalline materials," *Science*, vol. 359, no. 6376, pp. 675–679, 2018. 21

[53] M. J. Zachman, Z. Yang, Y. Du, and M. Chi, "Robust atomic-resolution imaging of lithium in battery materials by center-of-mass scanning transmission electron microscopy," *ACS nano*, vol. 16, no. 1, pp. 1358–1367, 2022. 21

[54] H. Rose, "Nonstandard imaging methods in electron microscopy," *Ultramicroscopy*, vol. 2, pp. 251–267, 1976. 21

[55] P. Nellist and J. Rodenburg, "Electron ptychography. I. experimental demonstration beyond the conventional resolution limits," *Acta Crystallographica Section A: Foundations of Crystallography*, vol. 54, no. 1, pp. 49–60, 1998. 21

[56] 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. 21

[57] P. M. Pelz, H. G. Brown, S. Stonemeyer, S. D. Findlay, A. Zettl, P. Ercius, Y. Zhang, J. Ciston, M. Scott, and C. Ophus, "Phase-contrast imaging of multiply-scattering extended objects at atomic resolution by reconstruction of the scattering matrix," *Physical Review Research*, vol. 3, no. 2, p. 023159, 2021. 21

[58] S. W. O. Gerchberg R. W., "A practical algorithm for the determination of plane from image and diffraction pictures," *Optik*, vol. 35, no. 2, pp. 237–246, 1972. 21, 82

[59] M. Guizar-Sicairos and J. R. Fienup, "Understanding the twin-image problem in phase retrieval," *JOSA A*, vol. 29, no. 11, pp. 2367–2375, 2012. 22, 91

[60] 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. 22

[61] F. Hüe, J. Rodenburg, A. Maiden, and P. Midgley, "Extended ptychography in the transmission electron microscope: Possibilities and limitations," *Ultramicroscopy*, vol. 111, no. 8, pp. 1117–1123, 2011. 23

[62] Z. Chen, M. Odstrcil, Y. Jiang, Y. Han, M.-H. Chiu, L.-J. Li, and D. A. Muller, "Mixed-state electron ptychography enables sub-angstrom resolution imaging with picometer precision at low dose," *Nature communications*, vol. 11, no. 1, p. 2994, 2020. 23

[63] H. Sha, J. Cui, and R. Yu, "Deep sub-angstrom resolution imaging by electron ptychography with misorientation correction," *Science Advances*, vol. 8, no. 19, p. eabn2275, 2022. 23

[64] M. Odstrčil, A. Menzel, and M. Guizar-Sicairos, "Iterative least-squares solver for generalized maximum-likelihood ptychography," *Optics express*, vol. 26, no. 3, pp. 3108–3123, 2018. 23

[65] 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. 24, 47, 53, 73

[66] V. Prabhakara, D. Jannis, G. Guzzinati, A. Béché, H. Bender, and J. Verbeeck, "HAADF-STEM block-scanning strategy for local measurement of strain at the nanoscale," *Ultramicroscopy*, vol. 219, p. 113099, 2020. 24

[67] A. De Backer, J. Gonnissen, and S. Van Aert, "Optimal experimental design for nano-particle atom-counting from high-resolution STEM images," *Ultramicroscopy*, vol. 151, pp. 46–55, 2015. 24, 29, 31, 110

[68] K. MacArthur, T. Pennycook, E. Okunishi, A. D'Alfonso, N. Lugg, L. Allen, and P. Nellist, "Probe integrated scattering cross sections in the analysis of atomic resolution HAADF STEM images," *Ultramicroscopy*, vol. 133, pp. 109–119, 2013. 25

[69] A. De Backer, K. Van den Bos, W. Van den Broek, J. Sijbers, and S. Van Aert, "StatSTEM: An efficient approach for accurate and precise model-based quantification of atomic resolution electron microscopy images," *Ultramicroscopy*, vol. 171, pp. 104–116, 2016. 25, 70, 72

[70] M. Nord, P. E. Vullum, I. MacLaren, T. Tybell, and R. Holmestad, "Atomap: a new software tool for the automated analysis of atomic resolution images using two-dimensional gaussian fitting," *Advanced structural and chemical imaging*, vol. 3, pp. 1–12, 2017. 25

[71] H. Akaike, "A new look at the statistical model identification," *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974. 26

[72] P. Broersen and H. E. Wensink, "On finite sample theory for autoregressive model order selection," *IEEE Transactions on Signal Processing*, vol. 41, no. 1, p. 194, 1993. 26

[73] G. Schwarz, "Estimating the dimension of a model," *The annals of statistics*, pp. 461–464, 1978. 26

[74] C. Biernacki, G. Celeux, and G. Govaert, "Assessing a mixture model for clustering with the integrated completed likelihood," *IEEE transactions on pattern analysis and machine intelligence*, vol. 22, no. 7, pp. 719–725, 2000. 26

[75] E. Arslan Irmak, P. Liu, S. Bals, and S. Van Aert, "3D atomic structure of supported metallic nanoparticles estimated from 2D ADF STEM images: A combination of atom-counting and a local minima search algorithm," *Small Methods*, vol. 5, no. 12, p. 2101150, 2021. 27

[76] M. Monai, K. Jenkinson, A. E. M. Melcherts, J. N. Louwen, E. A. Irmak, S. V. Aert, T. Altantzis, C. Vogt, W. van der Stam, T. Duchoň, B. Šmíd, E. Groeneveld, P. Berben, S. Bals, and B. M. Weckhuysen, "Restructuring of titanium oxide overlayers over nickel nanoparticles during catalysis," *Science*, vol. 380, no. 6645, pp. 644–651, 2023. 28

[77] P. Geuens and D. Van Dyck, "The s-state model for electron channeling in high-resolution electron microscopy," *Advances in Imaging and Electron Physics*, vol. 136, pp. 111–226, 2005. 29

[78] L. Zhou, J. Song, J. S. Kim, X. Pei, C. Huang, M. Boyce, L. Mendonça, D. Clare, A. Siebert, C. S. Allen, E. Liberti, D. Stuart, X. Pan, P. D. Nellist, P. Zhang, A. I. Kirkland, and P. Wang, "Low-dose phase retrieval of biological specimens using cryo-electron ptychography," *Nature communications*, vol. 11, no. 1, p. 2773, 2020. 33, 91

[79] 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. 33

[80] A. Strauch, D. Weber, A. Clausen, A. Lesnichaia, A. Bangun, B. März, F. J. Lyu, Q. Chen, A. Rosenauer, R. Dunin-Borkowski, and K. Müller-Caspary, "Live processing of momentum-resolved STEM data for first moment imaging and ptychography," *Microscopy and Microanalysis*, pp. 1–15, 2021. 34, 56

[81] P. M. Pelz, I. Johnson, C. Ophus, P. Ercius, and M. C. Scott, "Real-time interactive 4D-STEM phase-contrast imaging from electron event representation data," *arXiv preprint arXiv:2104.06336*, 2021. 34

[82] D. Jannis, C. Hofer, C. Gao, X. Xie, A. Béché, T. J. Pennycook, and J. Verbeeck, "Event driven 4D STEM acquisition with a timepix3 detector: microsecond dwell time and faster scans for high precision and low dose applications," *Ultramicroscopy*, p. 113423, 2021. 34, 45, 67

[83] H. Guo, E. Franken, Y. Deng, S. Benlekbir, G. Singla Lezcano, B. Janssen, L. Yu, Z. A. Ripstein, Y. Z. Tan, and J. L. Rubinstein, "Electron-event representation data enable efficient cryoem file storage with full preservation of spatial and temporal resolution," *IUCrJ*, vol. 7, no. 5, 2020. 34

[84] "Quantum detectors." Accessed on July fifth, 2023. 39

[85] "Amsterdam scientific instrument." Accessed on July fifth, 2023. 39

[86] Y. Chu-Ping, F. Thomas, D. Jannis, X. Xiaobin, V. A. Sandra, and V. Johan, "Real Time Integration Center of Mass (riCOM) Reconstruction for 4D-STEM," Dec. 2021. 43, 64

[87] H. Yang, T. J. Pennycook, and P. D. Nellist, "Efficient phase contrast imaging in STEM using a pixelated detector. part II: Optimisation of imaging conditions," *Ultramicroscopy*, vol. 151, pp. 232–239, 2015. 47, 53, 73

[88] L. Liu, N. Wang, C. Zhu, X. Liu, Y. Zhu, P. Guo, L. Alfilfil, X. Dong, D. Zhang, and Y. Han, "Direct imaging of atomically dispersed molybdenum that enables location of aluminum in the framework of zeolite ZSM-5," *Angewandte Chemie International Edition*, vol. 59, no. 2, pp. 819–825, 2020. 47

[89] D. Liu, L. Liu, K. Wu, J. Zhou, Q. Cheng, J. Lv, T. Cao, D. Zhang, F. Lin, and Y. Han, "Possible misidentification of heteroatom species in scanning transmission electron microscopy imaging of zeolites," *The Journal of Physical Chemistry C*, vol. 125, no. 34, pp. 18952–18960, 2021. 47

[90] H. Rose, "Phase contrast in scanning transmission electron microscopy," *Optik*, vol. 39, no. 4, pp. 416–436, 1974. 53, 107

[91] J. M. Zuo and J. C. Spence, *Advanced transmission electron microscopy: Imaging and diffraction in nanoscience*. Springer, 2016. 56

[92] I. Lobato and D. V. Dyck, "An accurate parameterization for the scattering factors, electron densities and electrostatic potentials for neutral atoms that obey all physical constraints," *Acta Crystallographica Section A*, vol. 70, pp. 636–649, 2014. 57, 106

[93] D. Van Dyck, "Is the frozen phonon model adequate to describe inelastic phonon scattering?," *Ultramicroscopy*, vol. 109, pp. 677–682, may 2009. 57, 59

[94] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," 2013. 59

[95] T. Friedrich, C.-P. Yu, J. Verbeeck, and S. Van Aert, "Phase Object Reconstruction for 4D-STEM using Deep Learning, (4D-STEM Training Data)," Aug. 2022. 60

[96] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), (Cham), pp. 234–241, Springer International Publishing, 2015. 61, 62

[97] T. Friedrich, C.-P. Yu, J. Verbeek, T. Pennycook, and S. V. Aert, "Phase Retrieval From 4-Dimensional Electron Diffraction Datasets," in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 3453–3457, IEEE, 2021. 61

[98] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018. 61

[99] P. Virtue, S. X. Yu, and M. Lustig, "Better than real: Complex-valued neural nets for MRI fingerprinting," in *Proceedings - International Conference on Image Processing, ICIP*, vol. 2017-Septe, pp. 3953–3957, 2018. 61

[100] J. Munshi, A. Rakowski, B. H. Savitzky, S. E. Zeltmann, J. Ciston, M. Henderson, S. Cholia, A. M. Minor, M. K. Chan, and C. Ophus, "Disentangling multiple scattering with deep learning: application to strain mapping from electron diffraction patterns," *arXiv preprint arXiv:2202.00204*, 2022. 61

[101] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv preprint arXiv:1710.05941*, 2017. 62

[102] D. Jannis, C. Hofer, C. Gao, X. Xie, A. Béché, T. J. Pennycook, and J. Verbeeck, "Event driven 4D STEM acquisition with a Timepix3 detector: microsecond dwelltime and faster scans for high precision and low dose applications," July 2021. 65

[103] T. Friedrich, C.-P. Yu, J. Verbeeck, and S. Van Aert, "Phase Object Reconstruction for 4D-STEM using Deep Learning, (4D-STEM Example Data)," Aug. 2022. 65

[104] J. Rodenburg, B. McCallum, and P. Nellist, "Experimental tests on double-resolution coherent imaging via STEM," *Ultramicroscopy*, vol. 48, no. 3, pp. 304 – 314, 1993. 65

[105] I. Lobato and D. Van Dyck, "An accurate parameterization for scattering factors, electron densities and electrostatic potentials for neutral atoms that obey all physical constraints," *Acta Crystallographica Section A: Foundations and Advances*, vol. 70, no. 6, pp. 636–649, 2014. 70

[106] 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 N. PD, "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. 70

[107] 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. 70

[108] F. Roddier, *Adaptive Optics in Astronomy*. Cambridge University Press, 1999. 79

[109] C. Barnbaum and R. F. Bradley, "A new approach to interference excision in radio astronomy: Real-time adaptive cancellation," *The astronomical journal*, vol. 116, no. 5, p. 2598, 1998. 79

[110] L. Brennan and L. Reed, "Theory of adaptive radar," *IEEE Transactions on Aerospace and Electronic SySTEMs*, vol. AES-9, no. 2, pp. 237–252, 1973. 79

[111] U. Michel, "History of acoustic beamforming," in *1st. Berlin Beamforming Conference*, 2006. 79

[112] N. Davids, E. Thurston, and R. Mueser, "The design of optimum directional acoustic arrays," *The Journal of the Acoustical Society of America*, vol. 24, no. 1, pp. 50–56, 1952. 79

[113] J. Billingsley and R. Kinns, "The acoustic telescope," *Journal of Sound and Vibration*, vol. 48, no. 4, pp. 485–510, 1976. 79

[114] M. P. van den Ende and J.-P. Ampuero, "Evaluating seismic beamforming capabilities of distributed acoustic sensing arrays," *Solid Earth*, vol. 12, no. 4, pp. 915–934, 2021. 79

[115] P. Hall, "Review of radio frequency beamforming techniques for scanned and multiple beam antennas," *IEE Proceedings H (Microwaves, Antennas and Propagation)*, vol. 137, pp. 293–303(10), October 1990. 79

[116] H. Yu, L. Zhong, A. Sabharwal, and D. Kao, "Beamforming on mobile devices: A first study," in *Proceedings of the 17th annual international conference on Mobile computing and networking*, pp. 265–276, 2011. 79

[117] K. Bliokh, E. Karimi, M. Padgett, M. Alonso, M. Dennis, A. Dudley, A. Forbes, S. Zahedpour, S. Hancock, and H. Milchberg, "Roadmap on structured waves," *arXiv preprint arXiv:2301.05349*, 2023. 79

[118] A. Fisher, "A review of spatial light modulators," *Optical Computing*, p. TuC1, 1985. 79

[119] C. Maurer, A. Jesacher, S. Bernet, and M. Ritsch-Marte, "What spatial light modulators can do for optical microscopy," *Laser & Photonics Reviews*, vol. 5, no. 1, pp. 81–101, 2011. 79

[120] L. De Broglie, *Recherches sur la théorie des quanta (Research on the Theory of the Quanta)*. PhD thesis, Ph. D. thesis, Migration-université en cours d'affectation, 1924. 79

[121] H. Rose, "Outline of a spherically corrected semiaplanatic medium-voltage transmission electron microscope," *Optik*, vol. 85, pp. 19–24, 1990. 79

[122] M. Haider, G. Braunshausen, and E. Schwan, "Correction of the spherical aberration of a 200 kv TEM by means of a hexapole-corrector," *Optik*, vol. 99, pp. 167–179, 1995. 79

[123] M. Haider, S. Uhlemann, E. Schwan, H. H. Rose, B. C. Kabius, and K. W. Urban, "Electron microscopy image enhanced," *Nature*, vol. 392, pp. 768–769, 1998. 79

[124] O. Krivanek, N. Dellby, and A. Lupini, "Towards sub-å electron beams," *Ultramicroscopy*, vol. 78, no. 1-4, pp. 1–11, 1999. 79

[125] R. Danev and K. Nagayama, "Single particle analysis based on zernike phase contrast transmission electron microscopy," *Journal of structural biology*, vol. 161, no. 2, pp. 211–218, 2008. 80

[126] K. Schultheiss, F. Pérez-Willard, B. Barton, D. Gerthsen, and R. Schröder, "Fabrication of a boersch phase plate for phase contrast imaging in a transmission electron microscope," *Review of scientific instruments*, vol. 77, no. 3, p. 033701, 2006. 80

[127] K. Schultheiss, J. Zach, B. Gamm, M. Dries, N. Frindt, R. R. Schröder, and D. Gerthsen, "New electrostatic phase plate for phase-contrast transmission electron microscopy and its application for wave-function reconstruction," *Microscopy and Microanalysis*, vol. 16, no. 6, pp. 785–794, 2010. 80

[128] N. Frindt, M. Oster, S. Hettler, B. Gamm, L. Dieterle, W. Kowalsky, D. Gerthsen, and R. R. Schröder, "In-focus electrostatic zach phase plate imaging for transmission electron microscopy with tunable phase contrast of frozen hydrated biological samples," *Microscopy and Microanalysis*, vol. 20, no. 1, pp. 175–183, 2014. 80

[129] R. Danev and W. Baumeister, "Cryo-EM single particle analysis with the Volta phase plate," *elife*, vol. 5, p. e13046, 2016. 80

[130] R. Danev, B. Buijsse, M. Khoshouei, J. M. Plitzko, and W. Baumeister, "Volta potential phase plate for in-focus phase contrast transmission electron microscopy," *Proceedings of the National Academy of Sciences*, vol. 111, no. 44, pp. 15635–15640, 2014. 80

[131] B. Buijsse, P. Trompenaars, V. Altin, R. Danev, and R. M. Glaeser, "Spectral DQE of the volta phase plate," *Ultramicroscopy*, vol. 218, p. 113079, 2020. 80

[132] K. Y. Bliokh, I. P. Ivanov, G. Guzzinati, L. Clark, R. Van Boxem, A. Béché, R. Juchtmans, M. A. Alonso, P. Schattschneider, F. Nori, and J. Verbeeck, "Theory and applications of free-electron vortex states," *Physics Reports*, vol. 690, pp. 1–70, 2017. 80

[133] M. Kim, "Large electron vortex beams produced by mems-based device," 2022. 80

[134] H. Yang, P. Ercius, P. D. Nellist, and C. Ophus, "Enhanced phase contrast transfer using ptychography combined with a pre-specimen phase plate in a scanning transmission electron microscope," *Ultramicroscopy*, vol. 171, pp. 117–125, 2016. 80

[135] A. H. Tavabi, P. Rosi, E. Rotunno, A. Roncaglia, L. Belsito, S. Frabboni, G. Pozzi, G. C. Gazzadi, P.-H. Lu, R. Nijland, M. Ghosh, P. Tiemeijer, E. Karimi, R. E. Dunin-Borkowski, and V. Grillo, "Experimental demonstration of an electrostatic orbital angular momentum sorter for electron beams," *Phys. Rev. Lett.*, vol. 126, p. 094802, Mar 2021. 80

[136] V. Grillo, A. H. Tavabi, F. Venturi, H. Larocque, R. Balboni, G. C. Gazzadi, S. Frabboni, P.-H. Lu, E. Mafakheri, F. Bouchard, R. E. Dunin-Borkowski, R. W. Boyd, M. P. Lavery, M. J. Padgett, and E. Karimi, "Measuring the orbital angular momentum spectrum of an electron beam," *Nature Communications*, vol. 8, no. 1, p. 15536, 2017. 80

[137] J. Verbeeck, A. Béché, K. Müller-Caspary, G. Guzzinati, M. A. Luong, and M. Den Hertog, "Demonstration of a 2× 2 programmable phase plate for electrons," *Ultramicroscopy*, vol. 190, pp. 58–65, 2018. 80, 84

[138] B. Barwick and H. Batelaan, "Aharonov–Bohm phase shifts induced by laser pulses," *New Journal of Physics*, vol. 10, no. 8, p. 083036, 2008. 80

[139] M. C. C. Mihaila, P. Weber, M. Schneller, L. Grandits, S. Nimmrichter, and T. Juffmann, "Transverse electron-beam shaping with light," *Physical Review X*, vol. 12, no. 3, p. 031043, 2022. 80

[140] H. Mueller, J. Jin, R. Danev, J. Spence, H. Padmore, and R. M. Glaeser, "Design of an electron microscope phase plate using a focused continuous-wave laser," *New journal of physics*, vol. 12, no. 7, p. 073011, 2010. 80

[141] O. Schwartz, J. J. Axelrod, S. L. Campbell, C. Turnbaugh, R. M. Glaeser, and H. Müller, "Laser phase plate for transmission electron microscopy," *Nature methods*, vol. 16, no. 10, pp. 1016–1020, 2019. 80

[142] F. J. G. De Abajo and A. Konečná, "Optical modulation of electron beams in free space," *Physical Review Letters*, vol. 126, no. 12, p. 123901, 2021. 80

[143] A. H. Tavabi, P. Rosi, A. Roncaglia, E. Rotunno, M. Beleggia, P.-H. Lu, L. Belsito, G. Pozzi, S. Frabboni, P. Tiemeijer, R. E. Dunin-Borkowski, and V. Grillo, "Generation of electron vortex beams with over 1000 orbital angular momentum quanta using a tunable electrostatic spiral phase plate," *Applied Physics Letters*, vol. 121, no. 7, p. 073506, 2022. 80

[144] A. eu, "Adaptem," 2023. `https://adaptem.eu/` [Accessed: (25/04/23)]. 82

[145] T. Matsumoto and A. Tonomura, "The phase constancy of electron waves traveling through Boersch's electrostatic phase plate," *Ultramicroscopy*, vol. 63, no. 1, pp. 5–10, 1996. 82

[146] K. Strehl, "Aplanatische und fehlerhafte abbildung im fernrohr," *Zeitschrift für Instrumentenkunde*, vol. 15, no. 7, pp. 362–370, 1895. 84

[147] V. N. Mahajan, "Strehl ratio for primary aberrations in terms of their aberration variance," *JOSA*, vol. 73, no. 6, pp. 860–861, 1983. 84

[148] F. Vega Ibáñez, A. Béché, and J. Verbeeck, "Can a programmable phase plate serve as an aberration corrector in the transmission electron microscope (TEM)?," *Microscopy and Microanalysis*, vol. 29, no. 1, pp. 341–351, 2023. 84, 86, 88

[149] J. Hadamard, "Resolution d'une question relative aux determinants," *Bull. des sciences math.*, vol. 2, pp. 240–246, 1893. 84

[150] J. Verbeeck, H. Tian, and P. Schattschneider, "Production and application of electron vortex beams," *Nature*, vol. 467, no. 7313, pp. 301–304, 2010. 84

[151] E. Zauderer, "Complex argument Hermite–Gaussian and Laguerre–Gaussian beams," *JOSA A*, vol. 3, no. 4, pp. 465–469, 1986. 84

[152] M. De Graef, *Introduction to conventional transmission electron microscopy*. Cambridge university press, 2003. 84

[153] A. Tonomura, "Applications of electron holography," *Reviews of modern physics*, vol. 59, no. 3, p. 639, 1987. 86

[154] W. Coene, A. Thust, M. O. De Beeck, and D. Van Dyck, "Maximum-likelihood method for focus-variation image reconstruction in high resolution transmission electron microscopy," *Ultramicroscopy*, vol. 64, no. 1-4, pp. 109–135, 1996. 86

[155] C. Ozsoy-Keskinbora, C. Boothroyd, R. Dunin-Borkowski, P. Van Aken, and C. Koch, "Hybridization approach to in-line and off-axis (electron) holography for superior resolution and phase sensitivity," *Scientific reports*, vol. 4, no. 1, pp. 1–10, 2014. 86

[156] R. Hovden, H. L. Xin, and D. A. Muller, "Extended depth of field for high-resolution scanning transmission electron microscopy," *Microscopy and Microanalysis*, vol. 17, no. 1, pp. 75–80, 2011. 88

[157] K. van Benthem, A. R. Lupini, M. P. Oxley, S. D. Findlay, L. J. Allen, and S. J. Pennycook, "Three-dimensional ADF imaging of individual atoms by through-focal series scanning transmission electron microscopy," *Ultramicroscopy*, vol. 106, no. 11-12, pp. 1062–1068, 2006. 88

[158] H. Blom and J. Widengren, "Stimulated emission depletion microscopy," *Chemical reviews*, vol. 117, no. 11, pp. 7377–7427, 2017. 88

[159] B. Hein, K. I. Willig, and S. W. Hell, "Stimulated emission depletion (STED) nanoscopy of a fluorescent protein-labeled organelle inside a living cell," *Proceedings of the National Academy of Sciences*, vol. 105, no. 38, pp. 14271–14276, 2008. 88

[160] J. N. Farahani, M. J. Schibler, and L. A. Bentolila, "Stimulated emission depletion (STED) microscopy: from theory to practice," *Microscopy: science, technology, applications and education*, vol. 2, no. 4, pp. 1539–1547, 2010. 88

[161] L. Thibon, M. Piché, and Y. De Koninck, "Resolution enhancement in laser scanning microscopy with deconvolution switching laser modes (D-SLAM)," *Optics Express*, vol. 26, no. 19, pp. 24881–24903, 2018. 88

[162] H. Dehez, M. Piché, and Y. De Koninck, "Resolution and contrast enhancement in laser scanning microscopy using dark beam imaging," *Optics express*, vol. 21, no. 13, pp. 15912–15925, 2013. 88

[163] F. C. Groen, I. T. Young, and G. Ligthart, "A comparison of different focus functions for use in autofocus algorithms," *Cytometry: The Journal of the International Society for Analytical Cytology*, vol. 6, no. 2, pp. 81–91, 1985. 89

[164] G. Guzzinati, A. Béché, H. Lourenço-Martins, J. Martin, M. Kociak, and J. Verbeeck, "Probing the symmetry of the potential of localized surface plasmon resonances with phase-shaped electron beams," *Nature Communications*, vol. 8, no. 1, p. 14999, 2017. 90

[165] M. Odstrčil, M. Lebugle, M. Guizar-Sicairos, C. David, and M. Holler, "Towards optimized illumination for high-resolution ptychography," *Optics express*, vol. 27, no. 10, pp. 14981–14997, 2019. 91

[166] F. Allars, P.-H. Lu, M. Kruth, R. E. Dunin-Borkowski, J. M. Rodenburg, and A. M. Maiden, "Efficient large field of view electron phase imaging using near-field electron ptychography with a diffuser," *Ultramicroscopy*, vol. 231, p. 113257, 2021. 91

[167] W. Van den Broek, M. Schloz, T. Pekin, P. Pelz, P.-H. Lu, M. Kruth, V. Grillo, R. Dunin-Borkowski, R. Miller, and C. Koch, "Towards ptychography with structured illumination, and a derivative-based reconstruction algorithm," *Microscopy and Microanalysis*, vol. 25, no. S2, pp. 58–59, 2019. 91

[168] P. Pelz, H. DeVyldere, P. Ercius, and M. Scott, "Structured illumination electron ptychography at the atomic scale," *Microscopy and Microanalysis*, vol. 28, no. S1, pp. 388–390, 2022. 91

[169] A. Maiden, M. Sarahan, M. Stagg, S. Schramm, and M. Humphry, "Quantitative electron phase imaging with high sensitivity and an unlimited field of view," *Scientific reports*, vol. 5, no. 1, pp. 1–8, 2015. 91

[170] M. T. Otten, "High-angle annular dark-field imaging on a tem/STEM sySTEM," *Journal of electron microscopy technique*, vol. 17, no. 2, pp. 221–230, 1991. 93

[171] M. J. Weber, A. J. Mackus, M. A. Verheijen, C. van der Marel, and W. M. Kessels, "Supported core/shell bimetallic nanoparticles synthesis by atomic layer deposition," *Chemistry of Materials*, vol. 24, no. 15, pp. 2973–2977, 2012. 93

[172] G. Bai, S. Yuan, Y. Zhao, Z. Yang, S. Y. Choi, Y. Chai, S. F. Yu, S. P. Lau, and J. Hao, "2D layered materials of rare-earth er-doped mos2 with nir-to-nir down-and up-conversion photoluminescence," *Advanced Materials*, vol. 28, no. 34, pp. 7472–7477, 2016. 93

[173] C. W. Helstrom, *Quantum Detection and estimation theory*. Academic Press, 1976. 100

[174] T. R. Harvey, F. S. Yasin, J. J. Chess, J. S. Pierce, R. M. Dos Reis, V. B. Özdöl, P. Ercius, J. Ciston, W. Feng, N. A. Kotov, B. J. McMorran, and C. Ophus, "Interpretable and efficient interferometric contrast in scanning transmission electron microscopy with a diffraction-grating beam splitter," *Physical Review Applied*, vol. 10, no. 6, p. 061001, 2018. 107

[175] M. Hammel and H. Rose, "Optimum rotationally symmetric detector configurations for phase-contrast imaging in scanning transmission electron microscopy," *Ultramicroscopy*, vol. 58, no. 3-4, pp. 403–415, 1995. 107

[176] Z. Lee, U. Kaiser, and H. Rose, "Prospects of annular differential phase contrast applied for optical sectioning in STEM," *Ultramicroscopy*, vol. 196, pp. 58–66, 2019. 107

List of Figures

List of Tables

Acknowledgment

It was about five years ago that I received an offer from University of Antwerp, confirming that I have been accepted as a future PhD student. This letter also marked the beginning of a journey in science, research, and this country far away. All these, and the Dr. title, are probably meaningless to people who are not involved and not in the field, but I know one person is going to be proud of me for what I have achieved, regardless of how much or how little.

My mother Yi-Rong Ding, who almost single-handedly raised me, has always been checking on my progress. It is without a doubt hilarious to receive phone calls from Mom asking how it goes with the microscope training or publishing of a paper, but it is oddly wholesome and was once the only support I could find during the solitude of the pandemic. Thank you for being there for me, no matter when and where.

The amount of support and help I have received in this foreign land is unbelievable, and I am beyond grateful for everyone who has been a part of my life in Belgium. Allereerst wil ik mijn dankbaarheid uitspreken naar mijn beide promotors. Die zijn Jo en Sandra, van wie ik zoveel heb geleerd. Sandra heeft voor alles gezorgd wat nodig was voor het onderzoek, inclusief haar begeleiding toen ik problemen had met de projecten of twijfelde over de weg van mijn PhD. De discussies met jou waren zeer behulpzaam en ik heb er echt van genoten. Jo heeft altijd coole ideeën met ons te delen, en ik weet niet meer hoe vaak ik me heb afgevraagd waar deze ideeën vandaan komen. In jou zie ik een passie voor de wetenschap dat echt aanstekelijk is, en ik raak er diep door geïnspireerd.

I wish to thank members of my Doctoral committee: Prof. Dr. Bart Partoens, Prof. Dr. Jan De Beenhouwer, Prof. Dr. Peter Nellist, and Prof. Dr. Knut Müller-Caspary for reading my thesis and sharing your thoughts on how to improve it.

This thesis, as well as the days in Antwerpen, would certainly be very different, were it not for one specific Thomas Friedrich. You are a great scientist and researcher who made every project that we worked on a fruitful adventure. You have also inspired me to achieve more than I could have imagined four years ago. As a talented musician, an enthusiastic hiker, and a true connoisseur and generous giver of good beer, you have provided the much-needed excitement for the life of a mundane PhD student. I will forever be thankful and cherish our friendship.

The Pico group is one of the things I am most proud of. With Thomas, Christoph Hofer, Tereza Špičáková, Francisco Vega Ibañez, and Anna Katia Russi Millán, we have a rock/jazz/blue/indi/folk/oriental band, which provided the necessary fuel for writing this thesis. I cherish fondly every memory we had together, whether the sweet performance at the wedding or the ride through the bitter Belgium rain for rehearsals.

I would also like to thank everyone who has been a part of my work and personal life in Belgium and acknowledge the ones with the most profound influence on me:

- Zezhong, thank you for all the deep discussion about physics and life. I learned a lot from how you face difficulties and overcome them.
- Francisco, you are an amazing friend with contagious positivity. The level of fun you possess should be illegal for PhD students in physics.
- Ece, thank you for all the guidance along the way. You are always so friendly and make the daily office life something I can look forward to.
- Yansong, you were one of the first to show me around when I came to Belgium and many more places after that. I am very thankful and feel lucky to get to know you and Weitong.
- Daen, thank you for all the cheerful memories you have provided. I truly enjoy working and chatting with you.
- Michał, thanks for all the help you provided for the thesis. I owe you another "thank you" for accompanying me to learn more Dutch than simply "Een pintje, aub!"
- Nadine, you have provided precious and high-quality data for me to test my ideas. I want to thank you for being a loyal customer for me and Thomas for using our methods.
- Miek and Lydia, I could not imagine how my life in Antwerpen would be if it were not for your unlimited support. From occasional favors to serious visa issues, and from the very beginning of my PhD till the end of it, I am almost spoiled by the luxury of your help.
- Lars and Alex, you guys are amazing. Chatting with you, even though 90 percent of it is around microscopes, was always a treat and lightened up my days at the lab. I am also very grateful for the support you have given me for the many projects that we worked together on.
- Gert, thank you for all the aesthetic improvements you made to my presentation and thesis, they certainly took away much of my nerdiness and are very well appreciated. Also, a personal thanks for being the biggest rock head in the Pico concerts.
- All the members of Jo's group: Francisco, Lukas, Luka, Nikita, Nico, Andrey, Daen, Hoel, Paul, and Arno, thank you for all the support and discussion. The meeting on Monday morning is always refreshing because of you.
- All the members of Sandra's group: Annick, Ece, Gizem, Zezhong, Yansong, Ivan, Mohammed, Tom, Tiago, Noopur, Annelies, and Jarmo. These four years would certainly be much more challenging if it were not for your help.

琦琦，謝謝妳為我的生活帶來的温暖。無論是愉快的周末下午，還是偶爾來之的沮喪，都能有著妳的陪伴是我的幸運和幸福。在博士的尾聲，妳曾説希望我能在未來做自己喜歡做的事，但其實無論是甚麼事，我都希望能夠與妳一起。當然也包含了妳的，和我的，未來將發生的，亦或是始料未及、捉摸不定的一切，我們都會在彼此陪伴下努力，就像這段日子妳一直為我而做的一樣。謝謝。

喻竹平 Yu, Chu-Ping, 05.11.2023