Three-Dimensional Phase Contrast Electron Tomography For Multiple Scattering Samples

Yonghuan David Ren

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2021-250
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-250.html

December 4, 2021

Three-Dimensional Phase Contrast Electron Tomography
For
Multiple Scattering Samples

by

David Ren

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Ted Van Duzer Associate Professor Laura Waller, Chair
Assistant Professor Mary Cooper Scott
Professor Shimon Michael Lustig

Fall 2021

Three-Dimensional Phase Contrast Electron Tomography
For
Multiple Scattering Samples

Abstract

Three-Dimensional Phase Contrast Electron Tomography
For
Multiple Scattering Samples

by

David Ren

Doctor of Philosophy in Electrical Engineering

University of California, Berkeley

Ted Van Duzer Associate Professor Laura Waller, Chair

Three-dimensional (3D) electron tomography (ET) is used to understand the structure and properties of samples, for applications in chemistry, materials science, and biology. By illuminating the sample at many tilt angles using an electron probe and modelling the image formation model, 3D information can be reconstructed at a resolution beyond the optical diffraction limit. However, as samples become thicker and more scattering, simple image formation models assuming projections or single scattering are no longer valid, causing the reconstruction quality to degrade. In this work, we develop a framework that takes the non-linear image formation process into account by modelling multiple-scattering events between the electron probe and the sample. First, the general acquisition and inverse model to recover multiple-scattering samples is introduced. We mathematically derive both the forward multi-slice scattering method as well as the gradient calculations in order to solve the inverse problem with optimization. As well, with the addition of regularization, the framework is robust against low dose tomography applications. Second, we demonstrate in simulation the validity of our method by varying different experimental parameters such as tilt angles, defocus values and dosage. Next, we test our ET framework experimentally on a multiple-scattering Montemorillonite clay, a 2D material submerged in aqueous solution and vitrified under cryogenic temperature. The results demonstrate the ability to observe the electric double layer (EDL) of this material for the first time. Last but not least, because modern electron detectors have large pixel counts and current imaging applications require large volume reconstructions, we developed a distributed computing method that can be directly applied to our framework for seeing multiple-scattering samples. Instead of solving for the 3D sample on a single computer node, we utilize tens or hundreds of nodes on a compute cluster simultaneously, with each node solving for part of the volume. As a result, both high resolution sample features and macroscopic sample topology can be visualized at the same time.

To my past & future self.

Contents

Contents ii

List of Figures iv

List of Tables viii

1 Introduction 1
- 1.1 Transmission Electron Microscopy Layout 1
- 1.2 Motivation 2
- 1.3 Electron Imaging Theory 4
- 1.4 Electron Tomography 7
- 1.5 Dissertation Outline 11

2 Theory 13
- 2.1 Modelling a Multiple Scattering Sample 13
- 2.2 Inverse Problem 17
- 2.3 Summary 23
- 2.4 List of Symbols 23

3 Phase Contrast Atomic Electron Tomography with Multiple Scattering Phantoms 25
- 3.1 Phantom Generation 25
- 3.2 Imaging Modality 26
- 3.3 Results 28
- 3.4 Summary 38

4 Cryo Electron Tomography with Multple-scattering Montmorillonite 41
- 4.1 Introduction 41
- 4.2 Simulated Results 42
- 4.3 Experimental Results 44

5 Efficient Computation for Electron Tomography 54
- 5.1 Introduction 54

5.2 Slice-binning 55
5.3 Implementation of Multi-slice Algorithm 58
5.4 Distributed Compute for Electron Tomography 60
5.5 Summary 71

6 Conclusion and Future Directions 72

Bibliography 76

List of Figures

1.1 Layout of a typical transmission electron microscope. A simulated defocus image of a synthetic silicon oxide sample is shown. 2

1.2 Sample micrograph of Iron-Platinum crystal on amorphous substrate in phase contrast. Micrograph courtesy of Jihan Zhou. 5

1.3 Illustration of filtered back projection algorithm. Illustration courtesy of [113].(a) Forward projection from 2D to 1D at multiple angles. (b) Backprojection step during tomographic inversion. 9

1.4 2D reconstruction of a phantom to illustrate the missing wedge problem with 20°, 60°, 100° missing wedge. As the amount of missing wedge increases, the axial resolution (horizontal) decreases and artifacts start to dominate. 10

1.5 2D Illustration of tilt-series misalignment using a phantom and its set projections. (a) Lateral translation misalignment due to sample drifting. (b) Incorrect tilt axis assumption with respect to the imaging system can cause significant errors during the reconstruction. 11

2.1 Foward measurements for phase contrast atomic electron tomography experiment with a core-shell SiO_2 needle geometry. (a) The multislice forward model treats the 3D sample as a series of 2D slices separated by propagation, thus accounting for multiple scattering. (b) The sample is tilted with respect to the electron beam to capture plane-wave illuminated images at varying angles (up to 180°). For each tilt angle several HRTEM images are recorded at different focus planes. 18

3.1 HRTEM simulation of the SiO_2 model. (a) A slice of the atomic structure, perpendicular to the electron beam direction. (b) The summed 2D projected potential of the object at 0° and (c) 5° rotation, with intensity scaled to show the weakly scattering edges. (d),(e) Noise-free (infinite dose) HRTEM images at 100 nm defocus for (b) and (c), respectively. (f),(g) Noisy versions of the same images, simulating a dose of 40 electrons/Å^2. 27

3.2 Varying dose. Phase contrast AET reconstructions of 1 Å thick 2D atomic potential slices of a simulated Si-SiO$_2$ reconstruction in $x-y$ across multiple z depths, using 60 tilt angles and 3 defocus values per angle. (a) Infinite dose (no noise), (b) 50,000 electrons/Å^2, and (c) 7,000 electrons/Å^2 total dose. Lower dose results in more noise, which causes errors and artifacts in the reconstruction. Each slice shows the square root of the reconstructed potential from 0 to 80 volts and the tilt axis is along the vertical direction. White arrows show location of reconstruction slices for the following sections in FIG. 3.3. 30

3.3 Varying dose. Phase contrast AET reconstructions in $z-y$ direction for (a) infinite electron dose, (b) 50,000 electrons/Å^2, and (c) 7,000 electrons/Å^2 total dose. We show (top row) a slice of the normalized reconstructed potential and (bottom row) the corresponding estimated atomic coordinates. Lower dose results in more noise, causing errors in the volume reconstruction, atom identification, and atom classification. 31

3.4 Cost function vs iterations to show convergence for various dose budgets. Reconstruction becomes noisier as total dose is lowered, and cost function increases. For each reconstruction, we ensure convergence is achieved. 32

3.5 Varying the number of tilt angles and the number of defocus planes while keeping a constant total dose. Phase contrast AET reconstructions for (a) 20 tilt angles with 9 defocus planes linearly increasing from 20nm to 100nm, (b) 60 tilt angles with 3 defocus planes at 20nm, 45nm, and 100nm, and (c) 180 tilt angles with single defocus plane at 100nm. The case of 3 defocus planes and 60 tilt angles gives minimal error, offering a good trade-off between number of tilt angles and defocus planes. 33

3.6 The missing wedge problem in the measurements primarily affects the axial accuracy of our reconstructions. All scenarios have the same total dose. Phase contrast AET reconstructions for (a) full tomography data with no missing angle , (b) limited tomography data with 30° missing angle, and (c) limited tomography data with 60° missing angle. 34

3.7 Regularization is important for image reconstruction quality. Phase contrast AET reconstructions using (a) real & positivity constraints only, (b) LASSO regularization, and (c) total variation (TV) regularization. In this case, TV regularization provides the best performance. 35

3.8 Phase contrast AET reconstructions with vacancies. (a) Ground truth atomic potential with vacancies. Reconstruction when (b) no vacancies are present, and when (c) 5% of the atoms are removed in the crystalline region and amorphous region. The top row shows slices in $z-y$ direction, and the bottom row shows slices in $z-x$ direction. 36

3.9 Phase contrast AET reconstructions for partial coherence with finer sampling during image calculation formalism. Yellow arrows show reconstruction artifacts due to partial coherence. 38

3.10 Phase contrast AET reconstructions for tungsten crystalline and tungsten oxide amorphous structure. 39

4.1 Clay stack and ion configurations used for image and tomogram simulation. . . 42

4.2 Integrated ion density profiles in tomographically reconstructed simulated images from structures shown in Fig. 4.1. Differences in ion densities can be observed between the four cases in 2D slices of 3D reconstructions (top row). These differences are clear in 1D profiles adjacent to the single isolated layer (middle row) and are also present in the stacked region despite the presence of artifacts due to the small reconstruction volume. 44

4.3 Images taken with different defocus values at the same tilt angle as part of the tilt series from which the tomogram in Fig. 4.5 was reconstructed. -75 nm defocus (a-c) the contrast is weakest (a-b) but the information transfer is highest (c). At -550 nm defocus (g-i) the contrast is highest (g-h) but the information transfer is reduced (i). Scale bars are 100 nm in (a, d, g), 50 nm in (b, e, h) and 1.2 Å^{-1} (c, f, i). 46

4.4 All traces of gold nanoparticles as fiducial markers(in blue) overlayed on top of minimum projection of the tilt-series. A in-plane rotation of angle θ is needed to ensure that the tilt axis is parallel to the y-axis. 47

4.5 Li-Mt suspensions viewed with cryoET. Isosurfaces of Li-Mt layers in 0.1 M (a) and 0.75 M (d) lithium chloride. (b, e) Subvolumes from (a) and (d), showing osmotic hydrate stacks. Comparison of structure factors from X-ray scattering and cryoET in 0.1 M (c) and 0.75 M (f), showing larger average interlayer spacing and larger layer thickness at lower electrolyte concentration. 49

4.6 Effect of layer curvature on montmorilonite-electrolyte ion-density profiles in 0.1 M lithium chloride. (a) Density reconstruction of a single exfoliated layer with low curvature quantified by H, the reciprocal of the local radius of curvature. (b) Single layer with intermediate curvature. (c) Highly curved layer in an osmotic hydrate stack. A single layer is colored according to local curvature and the neighboring layers shown in transparent gray. (d) Symmetric average ion-density profiles from within 20 nm of convex and concave sides of the low-curvature layer. Shading indicates the variance from 128,180 individual profiles. (e) Asymmetric average ion-density profile from the intermediate-curvature layer. (f) Asymmetric, and higher magnitude, average ion-density profiles between stacked layers (note difference in scale). (g) Non-negative matrix factorization (NNMF) of all absorbance profiles in (d), showing first two factors, f1 (Mt) and f2 (Li). (h) Increasingly asymmetric f2 in NNMF profile of curved layer in (b). (i) Highly asymmetric NNMF f2 for curved stacked layer. 52

5.1 Plot of cost function vs iterations to show convergence for various binning factors (N_B). 57

5.2 Plot of relative time savings (left y-axis) and relative error of reconstruction (right y-axis) vs slice binning factors (N_B). 57

5.3 Brief software architecture for the proposed framework from high level to lower level. 59

5.4 Conceptual illustration of our distributed reconstruction algorithm. Two examples of parallel nodes are shown in this figure, in blue and green, respectively. Each correspond to a different sub-volume of the sample, the tilt-series of which are shifted by $\Delta x_\theta = x_0\cos(\theta) + z_0\sin(\theta)$ and cropped from the full tilt-series dataset depending on the position of the sub-volume with respect to the full volume. After all sub-volumes are reconstructed, they are stitched together to form the large-volume reconstruction. 61

5.5 Illustration for calculating the minimum overlap between sub-volumes that covers the entire volume, assuming $N_x = N_z$. The reconstructed volume has cylindrical support and the inscribed blue square with side length of $\sqrt{2}/2N_z$ is used for final volume stitching. As such, the minimum overlap required is approximately 30%. 63

5.6 Splitting the volume into more sub-volumes, each with smaller size, enables faster reconstructions, but also induces more artifacts. (a) Extra reconstruction error (MSE with respect to a full reconstruction) vs. sub-volume size. (b) Number of nodes needed and reconstruction time for each node vs. sub-volume size. (c)-(f) Example $z - x$ slice and (g)-(j) $y - x$ slice of reconstructed 3D volume with various sub-volume sizes. As the sub-volume size decreases, artifacts increase, as indicated by the white arrows. 66

5.7 Computation complexity as a function of ratio of reconstructed sub-volume size to full volume size and sub-volume overlap ratio. The complexity is normalized with respect to that of a single reconstruction. 66

5.8 Increasing the overlap between adjacent sub-volumes reduces reconstruction errors. (a) Extra reconstruction error (MSE with respect to a single reconstruction) vs. overlap ratio. (b) Number of nodes needed and reconstruction time for each node as overlap ratio varies. (c)-(f) Single $z - x$ slice of reconstructed volume with various overlap ratios. (g)-(j) Single $y - x$ slice of reconstructed volume with various overlap ratios. (f) and (j) are full single reconstructions without any overlaps. 68

5.9 Carbon footprint with (a) varying sub-volume sizes and (b) varying overlap ratio. The optimal combination should be made by balancing the two parameters, such as sub-volume size of 150^3 voxels and a 35% overlap ratio. 69

5.10 Full field-of-view 3D reconstruction of lithium-montmorillonite, with a total of 5500×3952×952 voxels. (a) Maximum depth projection of the absorption channel. Three smaller regions-of-interest (ROIs) are cropped and zoomed in, with sizes of 400^3 voxels. (b)-(c),(e)-(f),and (h)-(i) Show zoom-in lateral slices at different depth for ROI 1, 2, and 3, respectively. (d), (g), and (j) show 3D volume renders for each of the ROIs. 70

List of Tables

3.1 Summary of atom tracing results, out of 62,402 sites in the tip region with a radius $\leq$ 12 nm diameter. 40

Acknowledgments

Thanks to the great vibe at UC Berkeley, it has been truly a great ride with ups and downs in the past five years. I am forever grateful for my time exploring my academic interests with minimal pressure. I am lucky to be part of the Wallerlab and the Molecular Foundry at Lawrence Berkeley National Lab. I appreciate all the people and events that I encountered along the journey, and they have made me who I am today.

First and foremost, I would like to thank my advisor, Prof. Laura Waller, who gave me the first opportunity to take a stab at her ongoing projects in the lab in my first year of Ph.D. The research experience not only offered me a way to enter the field of computational imaging but, more importantly, helped me build a solid foundation for this dissertation work. Her trust in me and my work is my greatest motivation. I must also express my sincere gratitude to Dr. Colin Ophus, who acts as my second "advisor" throughout this Ph.D. journey. Without his guidance and inspiration, none of this work would have been possible. Being a strong advocate for our work, he helped establish many collaboration projects and invited me to talk on different occasions. I would also like to thank my qualifying and dissertation committee, Prof. Mary Scott, Prof. Michael Lustig, and Prof. Steve Conolly, for guiding my project. I learned many things from my co-authors and collaborators worldwide: Prof. Jihan Zhou, Prof. Jianwei Miao, Dr. Michael Whittaker, Dr. Philipp Pelz, Hannah DeVyldere, Dr. Emrah Bostan, and Prof. Yi Xue.

I enjoyed teaching at UC Berkeley, especially the two semesters of being the Graduate Student Instructor for Prof. Steve Conolly's medical imaging class. He provided me with the maximum degree of freedom to teach any related topics I am interested in. Through teaching the classes, I met many bright minds who further inspired me along the way. Even though my dissertation work mainly deals with electrons, I spent most of my time in an optics lab. In my first few years in the lab, I had countless conversations with Li-hao Yeh, Michael Chen, Hsiou-yuan Liu to learn about the theory of optics. I also had the company of my peers Gautam Gunjala, Stan Smith, Kristina Monakhova, Kyrollos Yanny, Linda Liu, and Stuart Sherwin to take classes together and to share all the stress. I had the pleasure to build an optical microscope with Prof. Shwetadwip Chowdhury to learn about optical system alignment. In the office, occasional chats with Zach Phillips, Nick Antipa, Grace Kuo, Regina Eckert, Henry Pinkard, Neerja Aggarwal, Michael Kellman, and Ruiming Cao were always fun to have.

In college, Prof. Farzad Kamalabadi and Prof. Lara Waldrop influenced and shaped my earlier research career, and they were always encouraging me to achieve more no matter the circumstances. I would also like to thank Jerome Ng for being such a great career mentor. It was one of the hardest decisions I had to make to leave Illinois and start my graduate school at Berkeley.

My fantastic high school and college friends Tu-hsiang Ting, Jui-ting Hsu, Aijue Liu, Zijian Li, Shun Yao, Yue Liu, Shiming Song, Dan Li, Tetsuo Oura, Koji Son, and Angela Wang consistently supported me and entertained me. I also credit the Berkeley gang for keeping me sane during the time: Qing Tian, Qianyi Xie, Hongyu Zhang, An Ju, Huazhe

Xu, Jiaxin Zhao, Shouping Chen, Zichao Ye, Yipin Wu, Zitao Liao, and Yun Hao. Equally important, the cats that I adopted during the pandemic, Youtiao & Doujiang, provided me daily joy.

After all, none of what I described above would have happened without the unconditional love from my family. Ever since I was a little child, having the best education was never the focus for my parents. Instead, they always taught me to be kind and caring to people around me. My brother Alain always has my best interest at heart, even though he does not talk much. My cousin Renee also influenced me significantly by teaching me life lessons and sharing personal stories with me. Finally, I owe a great debt of gratitude to my partner Amy, who was by my side every step of the way, who listened, cared, and helped me navigate through tough decisions, and most importantly, whose existence makes me feel calm.

Chapter 1

Introduction

1.1 Transmission Electron Microscopy Layout

Invented in the 1930s by Knoll and Ruska [68, 110], electron microscopy has since been developed to become a prevalent imaging technology in chemistry, materials science, geoscience, and biology [12, 141]. Transmission Electron microscopes (TEM) fundamentally rely on the interaction between the electron probe and matter, as well as magnetic lenses that are capable of bending the path of the fast traveling electrons to form an image.

A conventional TEM (CTEM) or high resolution TEM (HRTEM) has the layout shown in Fig. 1.1. Electrons are emitted from an electron gun and are accelerated through an anode to a desired speed, with a corresponding energy. According to the application and different planned experiments, the energy of the electrons could range from 20 KeV to 1 MeV, and the current of the electron source could fall anywhere between 1-10 nA. For a typical electron energy of 300 KeV, the corresponding de Broglie wavelength is around 2 pm. After that, the accelerated electrons pass through a set of magnetic lenses, which are made of electromagnets. Within the lens, the path of electrons are altered by the magnetic field created by the coils. One can adjust the current passing through the coil to change the strength of the magnetic field and hence the power of the lens. Through a set of condenser and objective lenses, the electron wave is collimated into a plane wave and ready to be illuminated on the sample, which sits on a stage that is sometimes capable of tilting [26]. After the electron beam interacts with the sample, it is magnified through another set of objective lens and projector lenses and an intensity image is formed onto the detector. Recent advances in direct electron detectors with high quantum efficiency [82] have enabled many discoveries. In structural biology, for instance, many beam-sensitive samples can now be seen and new protein structures have been solved [12].

Alternatively, an imaging mode also known as Scanning Transmission Electron Microscopy (STEM) can be configured on the same system. A beam is focused into a small probe rather than collimated as a plane wave. Then, the probe is rastered across the the field-of-view (FoV) to scan the entire sample. In the end, the scattered electrons are collected and processed. If a center disc is placed to collect the direct beam, then a bright field image can be formed. On the other hand, if an annular ring detector is placed that collects scattered electrons only, an annular dark field (ADF) image can be formed at the end of the raster scan. One popular imaging mode is called high-angle ADF (HAADF), which directly measures the signal from thermal diffuse scattering electrons [90, 99]. Due to its ease of data interpretation, it has been adopted for 3D atomic electron tomography [141], where linear tomographic inversion techniques can be applied.

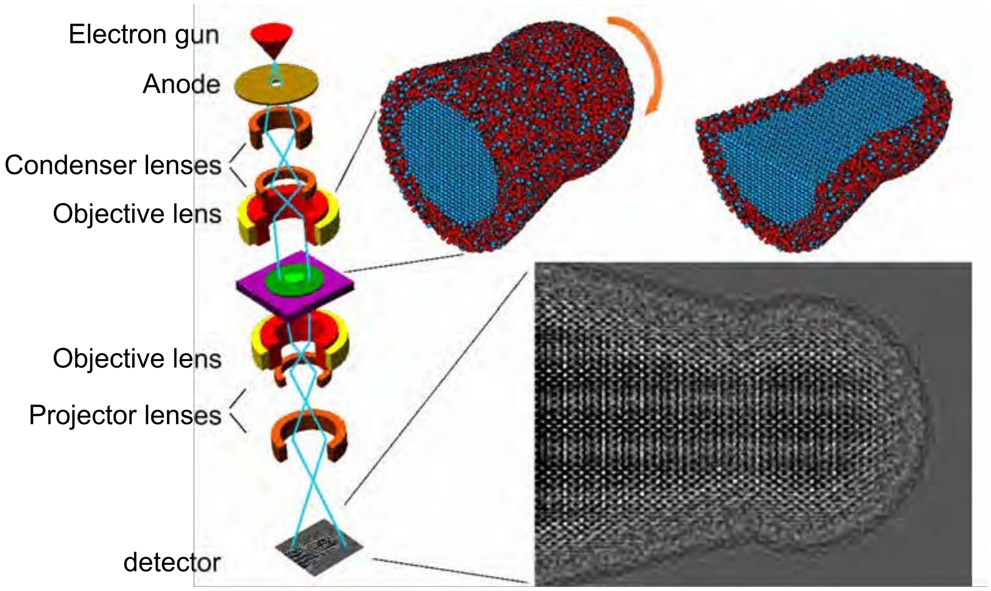

Figure 1.1: Layout of a typical transmission electron microscope. A simulated defocus image of a synthetic silicon oxide sample is shown.

1.2 Motivation

As described above, TEM offers unprecedented resolution for imaging applications in biology and materials science [79, 73]. Modern systems can quantitatively reconstruct 3D local structure, electrostatic and magnetic potentials, and local chemistry [83]. Recent progress enables locating the 3D position of individual atoms with high precision [128, 6, 140], and even determining both the 3D position and species of every atom in a nanoscale sample [141]. These atomic electron tomography (AET) studies use a TEM imaging mode called annular dark field (ADF) scanning transmission electron microscopy (STEM). ADF-STEM generates contrast that increases monotonically with the 2D projection of the 3D electrostatic potential of the sample along the beam direction. Such approximated linearity allows for traditional tomographic reconstruction algorithms [38, 102]. However, ADF-STEM requires large electron doses, as it is much less efficient than phase contrast imaging [93, 19]. Additionally, because the electron probe is focused to a small spot and scanned over the sample surface, sample motion during the experiment can cause artifacts [91].

The simplest phase contrast imaging mode used in TEM studies is plane-wave illumination, usually referred to as high-resolution transmission electron microscopy (HRTEM). However, at atomic resolution, HRTEM imaging produces highly nonlinear contrast for any sample thicker than a few atomic monolayers, making it difficult to interpret the results [127, 66]. For thin samples, comparing experiments to simulations can recover some quantitative 3D information [59, 4], but this is difficult or impossible for experiments with a high degree of multiple electron scattering. Thus, phase contrast imaging is not widely used in materials science electron tomography studies at atomic resolution.

By comparison, phase contrast HRTEM imaging in biology is simpler to interpret because most biological specimens can be approximated as *weak phase objects*, which I will derive in a later section, allowing for the sample's phase to be reconstructed from a single defocused intensity measurement [31]. This single-image requirement is important for biological samples because they tend to be extremely sensitive to electron beam damage and cannot tolerate the much higher electron doses used in materials science [33]. In structural biology, the introduction of direct electron detectors with high quantum efficiency [82] has rapidly expanded the number of solved protein structures, using 3D tomographic averaging of images of many identical or near-identical protein structures with random orientations. This technique is called single particle cryo-electron microscopy (cryo-EM) [30]. When imaging larger biological samples, averaging of sub-volumes can also produce high-resolution reconstructions [16].

Recent advances in computational methods have improved reconstruction accuracy even further, for example by introducing a correction for the microscope contrast transfer function (CTF) [12]. However, advanced algorithms generally make a weak object assumption and treat the measured signal as a linear sum of the projected potential [133]. This linearizes the physical model in order to provide a closed-form solution, but these assumptions usually only hold for very thin samples [92]. Nonlinear effects of multiple scattering are non-negligible for thick samples, which represent a large majority of materials science samples. Therefore, thick samples require both a nonlinear forward model and a reconstruction method that captures the dynamical scattering of the electron beam.

Nonlinear phase reconstruction in 2D for TEM includes algorithms for reconstruction of the sample potential phase contrast measurements [45, 11], maximum likelihood methods [23, 76, 75], and other iterative algorithms [84, 67, 1, 2]. These methods, however, are usually limited to samples that are either single scattering or satisfy crystal approximations.

Methods to correct for multiple scattering in 3D phase reconstructions have been proposed in optics [80, 117, 63, 134, 39, 112, 77, 124]. A typical strategy - the multislice or beam propagation method [25, 66, 129] - treats the 3D object as a series of 2D slices, each with it's own transmittance function, separated by small distances of free-space propagation. For TEM, the interaction of the electron beam with the sample can thus be modeled by two linear operators. The first is a multiplication by the transmittance function that describes the absorption and phase delay of the electron beam when interacting with that slice of the sample. The second is the free space propagation operator, which captures the dynamics of propagation. Unfortunately, these two operators do not commute, making the inverse scattering calculation both nonlinear and non-convex. Van den Broek and Koch have proposed an inversion method for multiple electron scattering, which uses multiple beam tilt projections for phase contrast TEM imaging to perform a 3D reconstruction with very few layers [130, 129, 60] similar to 3D Fourier Ptychographic Microscopy [124]. In simulation, they were able to reconstruct the atomic potential of a small nanoparticle in 3D from a small number of tilt angles, for strongly scattering atoms and a low TEM accelerating voltage of 40 kV, and assuming structural priors. However, the 3D transfer function for tilting the beam results in non-isotropic resolution [69, 86]; hence, the axial resolution is fundamentally limited when assuming no structural priors on the sample.

In this dissertation, I present an isotropic high resolution framework for 3D reconstruction from intensity-only images taken at varying tilt angles and defocus values. Our algorithm models multiple scattering of the electron beam and strong phase shifts induced by individual atoms. Along with efficient regularization and our carefully chosen inverse problem formulation, these improvements enable imaging of thicker samples and those that cannot withstand high electron doses. Biological cryo-EM studies may also benefit if they are performed on very large volumes (where the projection assumption breaks down) or contain multiple scattering regions.

1.3 Electron Imaging Theory

Phase Contrast Imaging

Samples in materials science and biology are transparent to electrons beams, meaning that they do not absorb electrons, but only scatter them and slightly change their paths, and the atomic potentials of the samples contribute to the phase delay of the electron wave. Because the electron detector can only measure the intensity of the complex wave function (square of the amplitude), the phase information is lost. As a result, under HRTEM imaging mode, when an electron plane wave is illuminated onto the phase sample, the image formed has a low contrast. Hence, to retrieve information of the phase and subsequently the sample, we need to introduce the phase information into the amplitude variation of the image captured, such as Fig.1.2. The amplitude variation shown in Fig.1.2 and Fig.1.1 are realized by taking deliberate defocused intensity images of the sample, and next we explain the relationship between the intensity variation of a defocused image and its phase information.

Without loss of generality, we derive the relationship assuming a thin 2D sample, as phase induced from 3D multiple scattering samples are non-linear with respect to the intensity measured and an analytical relation can be more convoluted. The transmittance function for a thin 2D object can be written as $t(\mathbf{r}) = \exp(i\sigma v_z(\mathbf{r}))$, where σ is the beam-sample interaction parameter that linearly scales with wavelength λ, $\mathbf{r} = (x, y)$ is the lateral

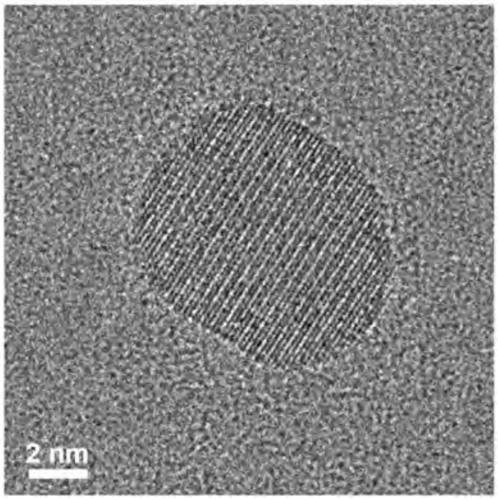

Figure 1.2: Sample micrograph of Iron-Platinum crystal on amorphous substrate in phase contrast. Micrograph courtesy of Jihan Zhou.

coordinates, and $v_z(\mathbf{r})$ is the projected 2D potentials. The combined term $\sigma v_z(\mathbf{r})$ forms the phase induced by the sample. Since it is a thin 2D sample and they typically induce very small phase values, we can apply the weak phase object approximation (WPOA) by only taking its fist-order Taylor series:

$$t(\mathbf{r}) \sim 1 + i\sigma v_z(\mathbf{r}) \tag{1.1}$$

When an incident electron wave $\psi_{inc}(\mathbf{r})$, satisfying the time-independent Schrodinger's equation, is illuminated, the resulting electron wave is the product of the two: $\psi_t(\mathbf{r}) = t(\mathbf{r})\psi_{inc}(\mathbf{r})$. The rest of the electron microscope can be modelled as a low-pass filter with a kernel of $h(\mathbf{r})$, with its Fourier transform of

$$H(\mathbf{k}) = A(\mathbf{k})\exp(i\chi(\mathbf{k})), \tag{1.2}$$

where $A(\mathbf{k})$ is the aperture function that models the attenuation of electrons when scattered to higher angles, and $\chi(\mathbf{k})$ is the aberration function that alters the path of the electrons at different angles $\mathbf{k} = (k_x, k_y)$. For instance, a perfect imaging system would imply $\chi(\mathbf{k}) = 0$. When deliberate defocus is introduced to the imaging process, $\chi(\mathbf{k}) \propto k^2 = (k_x^2 + k_y^2)$.

Hence, the image acquired is the intensity of the field convolved (denoted by $\star$) with the kernel:

$$\begin{aligned} I(\mathbf{r}) &= \left|\psi_t(\mathbf{r}) \star h(\mathbf{r})\right|^2 && (1.3) \\ &= \left|1 \star h(\mathbf{r}) + i\sigma v_z(\mathbf{r}) \star h(\mathbf{r})\right|^2. && (1.4) \end{aligned}$$

By expanding the intensity term and omitting the terms of second or higher order, we arrive at:

$$I(\mathbf{r}) \approx H(0) + 2\sigma v_z(\mathbf{r}) \star h_{WP}(\mathbf{r}), \tag{1.5}$$

and equivalently, the Fourier transform of the image is

$$\mathcal{F}\left\{I(\mathbf{r})\right\} \approx H(0)\delta(\mathbf{k}) + 2\sigma\mathcal{F}\left\{v_z(\mathbf{r})\right\} \cdot \sin\chi(\mathbf{k}), \tag{1.6}$$

where $\mathcal{F}\{\cdot\}$ is the Fourier transform operator. Equation (1.6) above provides a mathematical explanation for the phase of a sample contributing to the intensity variation. The spectrum of the phase information is modulated by a contrast transfer function (CTF) that is the sine of the aberration function. Therefore, when the imaging is conducted in-focus, very little contrast of phase is turned into amplitude variation. However, as defocus is introduced, the phase contrast becomes stronger. Notice that the transfer function is periodic and has many zero crossings. As such, some frequency information about the phase is lost, depending on the defocus distance. To recover a quantitatively accurate phase profile, one can capture multiple images with different defocus distances. In addition, since the aberration function introduced by defocus has very small values in the lower frequency region closer to the DC in Fourier space, low-frequency content is notoriously difficult to recover [61], and larger defocus distances are often required.

Noise Statistics

When imaging beam-sensitive samples, measurements are generally noisy due to small dosage of illumination, so it is important to take into the account the noise associated with the measurement processes. The two common noise factors are the detector readout noise and the electron counting noise.

The detector readout noise can be approximately modelled as white Gaussian noise with zero mean and a standard deviation of σ_w:

$$w \sim \mathcal{N}(0, \sigma_w^2). \tag{1.7}$$

The value of σ_w is detector dependent [82]. Notice that the noise per pixel is independent from the intensity captured at the pixel, and each pixel's noise can be considered as independent and identically distributed $(i.i.d)$.

The electron counting noise (also known as shot noise) is the noise associated with the the process of electron arrival at the detector. It follows a Poisson distribution:

$$p \sim \text{Poisson}(\lambda = I). \tag{1.8}$$

According to Poisson distribution, both the mean and variance of the noise p at a pixel are the true intensity at that pixel. As such, noise p is not an additive noise.

However, Poisson noise closely resembles a Gaussian distribution when the intensity is large [142]. Mathematically, we say that when the intensity is high, noise p is approximated by $\mathcal{N}(0, I)$. On the other hand, as the intensity (or equivalently dose) decreases, the distribution becomes less symmetric and more one-sided. Notice here that the noise for each pixel has a different distribution, and each depends on its own intensity. Therefore, correctly modelling the distribution of noise is a difficult, yet very important, task in electron tomography.

1.4 Electron Tomography

Fundamental Principles of Linear Electron Tomography

Tomography is a class of methods to recover 3D objects from a series of 2D measurements of the sample. It is a technique that is widely adopted in medical imaging, optics, materials science, remote sensing, and notably other areas of science [141, 107, 106, 8, 37, 63]. In an X-ray Computed Tomography (X-ray CT) scanner, for instance, the X-ray source and the detector panel rotate around the patient. While it is rotating, a series of 2D X-ray projections of the patients are taken. After the scan, an inversion algorithm recovers the 3D structure of the patient from the projections. In an electron tomography system, on the other hand, samples of interest are prepared and placed onto a tilting stage. Instead of rotating the source and detector around the sample, the sample itself is tilted to different angles while the microscope is fixed in place and acquires images.

The fundamental principle that tomography relies on is the projection of 3D object down to a 2D image, also known as the Radon transform [103, 104]. A single projection image is related to the 3D sample by:

$$I(x,y;\theta) = \iint f(x',y,z')\delta(x'\mathrm{cos}\theta + z'\mathrm{sin}\theta - x)\,\mathrm{d}x'\mathrm{d}z', \tag{1.9}$$

where $\delta(\cdot)$ is the Dirac delta function, θ is the tilt angle of the sample. The 3D sample $f(x',y,z')$ is tilted with respect to the y-axis to an angle θ, and the beam propagates axially along the z-axis to project the object down to a 2D image. The simple 2D version of projection is illustrated in Fig.1.3(a).

The formulation above is used to derive the Projection Slice Theorem, which relates the 3D object and the projected image in terms of the information of the sample imaged. By taking a 2D Fourier transform with respect to the coordinates x and y on both sides of Eq.(1.9), we have:

$$\tilde{I}(k_x,k_y;\theta) = \iiiint f(x',y,z')\delta(x'\mathrm{cos}\theta + z'\mathrm{sin}\theta - x)\mathrm{exp}(-2\pi i(k_x x + k_y y))\,\mathrm{d}x\mathrm{d}y\mathrm{d}x'\mathrm{d}z'. \tag{1.10}$$

By re-arranging the terms on the right hand side and separating out the variable x, we have:

$$\tilde{I}(k_x,k_y;\theta) = \iiint f(x',y,z')\int \delta(x'\mathrm{cos}\theta + z'\mathrm{sin}\theta - x)\mathrm{exp}(-2\pi i(k_x x + k_y y))\,\mathrm{d}x\mathrm{d}y\mathrm{d}x'\mathrm{d}z' \tag{1.11}$$

$$= \iiint f(x',y,z')\mathrm{exp}(-2\pi i(k_x(x'\mathrm{cos}\theta + z'\mathrm{sin}\theta) + k_y y))\,\mathrm{d}y\mathrm{d}x'\mathrm{d}z'. \tag{1.12}$$

Notice that the triple integral in Eq.(1.12) corresponds to a 3D Fourier transform of the function $f(x',y,z')$, and the 2D spectrum of the image is the 3D spectrum of the object evaluated on a plane. Therefore, the Fourier transform of a projection image is:

$$\tilde{I}(k_x,k_y;\theta) = \tilde{f}(k_{x'},k_y,k_{z'})\Big|_{k_{x'}=k_x\mathrm{cos}\theta,k_{z'}=k_x\mathrm{sin}\theta} = \tilde{f}(k_x\mathrm{cos}\theta,k_y,k_x\mathrm{sin}\theta). \tag{1.13}$$

Equation (1.13) is a version of the well-known Fourier slice theorem. The theorem suggests that the Fourier transform of the projection corresponds to a 2D slice within the 3D spectrum of the object. The derivation above is a special case of the general theorem in that it restricts the tilt axis to be aligned with the y-axis. In fact, the object can be rotated arbitrarily in 3D and the same result would hold. This theorem provides the intuition to understand almost any tomography system, and from it many reconstruction algorithms and artifacts can be understood, even though samples can be multiple-scattering and algorithms nonlinear. From a series of projections, we are effectively sampling the 3D spectrum of the object. And in the end a 3D object spectrum can be recovered by "stitching" all projections back together in the frequency space.

Tomographic Inversion Techniques

Depending on the particular imaging modality, a pre-processing step may be required before inverting the projections. For STEM tomography, since the measured signal is directly the accumulated high-angle scattering, which is proportional to Z^2, where Z is the atomic number, the forward model in Eq.(1.9) can be directly applied. However, if the projection images are acquired in phase contrast TEM mode through defocus, the image intensities are not linearly proportional with the projected potentials, shown in Eq.(1.6). As such, CTF correction is needed to be able to apply Eq.(1.9) [12, 133]. As we derived previously, such a model depends heavily on the linear approximation. If the sample becomes highly scattering, after CTF correction the phase may no longer be a linear projection of the sample's electrostatic potential. Alternatively, one can experimentally place a phase plate to avoid CTF correction and directly obtain an image that is proportional with respect to the projected potentials [57]. Note that this method also required linearity in order to recover the 3D sample.

When a quantity that is directly proportional to the projected potential is obtained, either through direct measurement or after CTF correction, the 3D volume is ready to be recovered. Filtered back projection (FBP) is one of the fundamental and original methods to stitch the projections back to a 3D volume [13]. As the name suggests, FBP consists of two steps: filtering and back projection. Since each projection is a 2D slice of the object spectrum, the low-frequency space of the 3D spectrum is inevitably over-sampled and high frequencies are under-sampled. Therefore, to weigh the balance in different regions, a filter is applied. Theoretically, a ramp filter is needed. In practice, however, a ramp filter may amplify the noise due to its high weights at high frequencies [13]; hence, other filters have been proposed to mitigate the effect of noise [13, 62]. The second step is known as the back projection step, where the filtered projection image (by the ramp filter or otherwise) is distributed across the line integral with respect to its tilt angle θ. After all projections are back projected on the 3D grid, the resulting 3D object is the sum of all back projections, as illustrated in Fig.1.3.

Another popular class of algorithms is iterative algorithms [100, 41, 106]. An iterative algorithm attempts to solve an optimization problem by iteratively updating the solution,

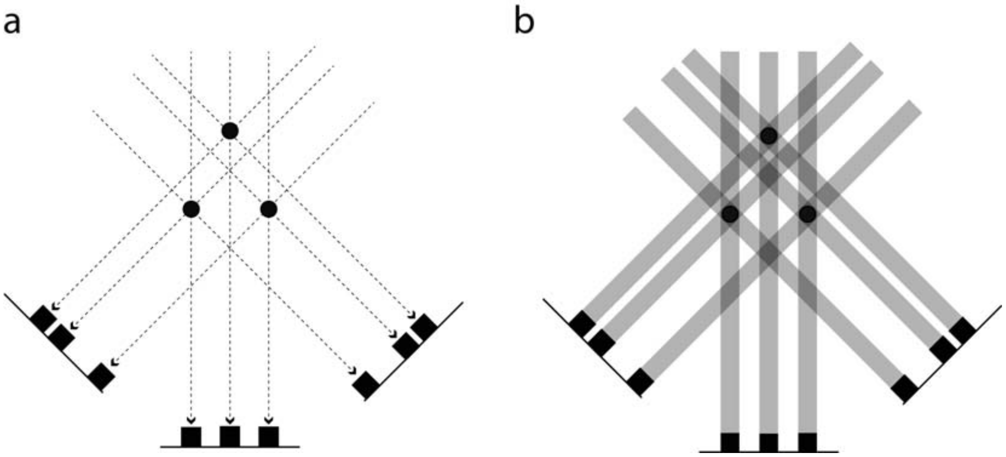

Figure 1.3: Illustration of filtered back projection algorithm. Illustration courtesy of [113].(a) Forward projection from 2D to 1D at multiple angles. (b) Backprojection step during tomographic inversion.

rather than arriving at a solution with a single calculation. The benefit of such methods is that it is easier to enforce sample priors and constraints, thus reducing artifacts. One drawback of such methods is that it tends to take more time in order for the algorithm to converge.

Reconstruction Complexity

Assume that a a volume is cubic and has N^3 voxels, and P projections are acquired. The first filtering step is performed in the Fourier space and the complexity per projection is dominated by the Fourier transform operation, which is $\mathcal{O}\left(N^2 \log N\right)$, and the complexity when all P projection combined is thus $\mathcal{O}\left(PN^2 \log N\right)$. The second step, if performed in real space as illustrated in Fig.1.3, is $\mathcal{O}\left(N^3\right)$ per projection, and the total complexity is $\mathcal{O}\left(PN^3\right)$. Alternatively, since each projection corresponds to a slice of the 3D spectrum, one can directly combine all slices in the 3D Fourier space and interpolate all frequency components onto a rectangular grid [13]. After that, a 3D inverse Fourier inverse transform is computed to obtain the 3D reconstructed sample. Depending on the interpolation method used, the dominating complexity in this operation is usually the Fourier transform, which has a complexity of $\mathcal{O}\left(N^3 \log N\right)$ [62, 50].

Common Issues in Electron Tomography

In this section, common practical issues that are known to be difficult to cope with in the general area of tomography will be introduced. If the issues are not dealt with properly, significant artifacts can result in the reconstruction and affect the interpretability of the reconstruction.

Missing wedge

We recall from Fourier slice theorem that each projection image corresponds to a 2D slice of the 3D sample spectrum. If the object can be tilted to a full range of 180° (-90° to 90°), the entire spectrum can be covered. However, due to experimental constraints, the samples usually need to be placed on a substrate that does not allow 180° tilt range [113]. In our case, the sample's tilt range is typically from -60° to 60° [137]. As a result, the sample spectrum is not fully measured, with significant under-sampling along the axial direction. Because the region of missing information closely resembles a wedge, this effect is also known as the "missing wedge" problem. The consequence of failing to fully sample the spectrum is that the axial resolution is significantly worse than the lateral resolution, as illustrated in Fig.1.4.

There are several methods to mitigate this problem. First, from a hardware perspective, depending on the type of sample, a different geometry can be adopted. In [106, 141], the sample is mounted on a tip that tilts during the acquisition. Since the sample is no longer placed on a stage, it can be tilted to a range of full 180 °. Secondly, many have explored software algorithms to mitigate the artifact [106, 100, 101]. However, all algorithms merely try to fill in the missing information with their best judgements regarding the spectrum, so extra caution is needed to validate the accuracy of such reconstructions.

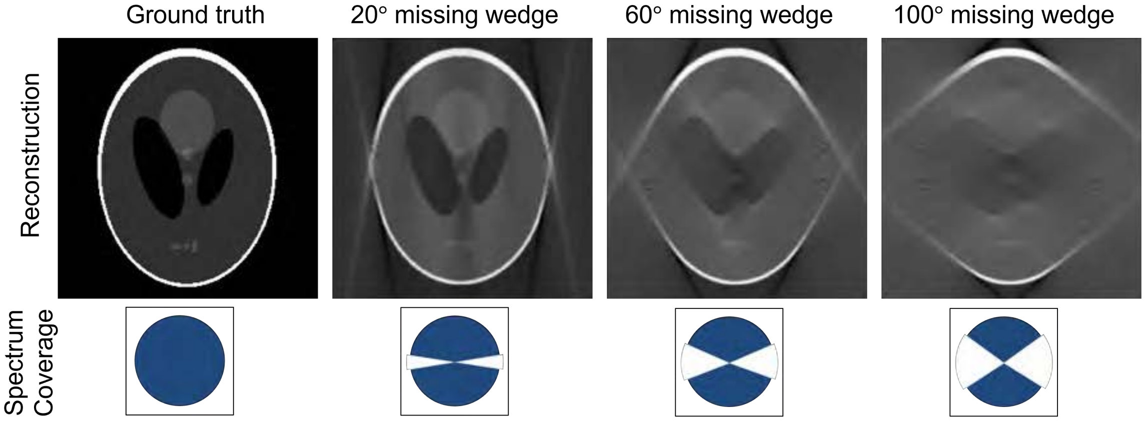

Figure 1.4: 2D reconstruction of a phantom to illustrate the missing wedge problem with 20°, 60°, 100° missing wedge. As the amount of missing wedge increases, the axial resolution (horizontal) decreases and artifacts start to dominate.

Tilt-series alignment

Image alignment plays a significant role in the quality of tomographic reconstruction. There are several reasons a tilt-series might not be well aligned. First, the sample drifts with respect to the imaging apparatus during the acquisition, as illustrated in Fig.1.5(a). For dose sensitive samples, it is impossible to adjust the focus while looking at the sample, as the electron beam is sample-damaging. Often times the focus is adjusted outside of the sample region of interest before translating the region of interest into the FoV. Additionally, tilting the sample can also drive it out of the FoV. Therefore, images need to be registered before passing to the reconstruction step.

Secondly, the tilt axis may be misaligned with respect to the image FoV. As shown in Fig.1.5(b), the expected tilt axis is parallel with the horizontal axis. However, the experimental tilt axis is in fact rotated with a certain angle, causing images to be misaligned. Correction is also needed before reconstructing the tilt-series. If not addressed properly, the reconstruction could suffer from lower resolution and inaccuracy. In a later section, methods used to align the tilt-series will be discussed, when markers in the images are available.

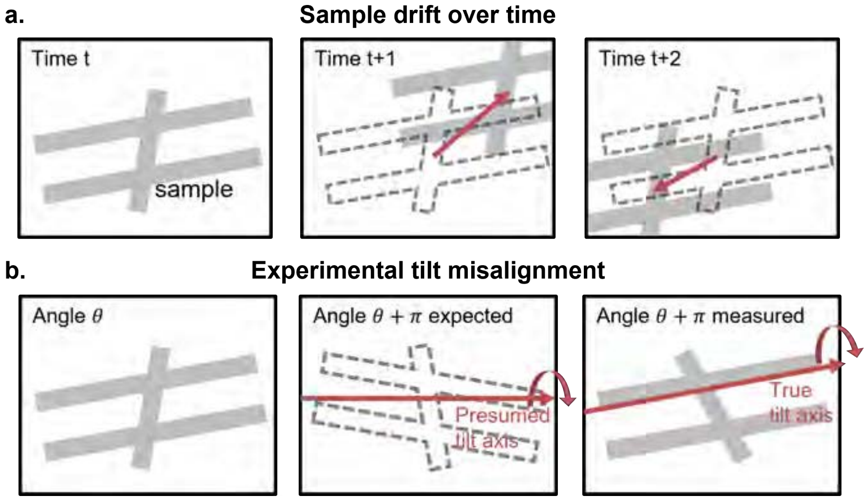

Figure 1.5: 2D Illustration of tilt-series misalignment using a phantom and its set projections. (a) Lateral translation misalignment due to sample drifting. (b) Incorrect tilt axis assumption with respect to the imaging system can cause significant errors during the reconstruction.

1.5 Dissertation Outline

Up until this point we have introduced the basic apparatus for imaging with electrons, the fundamental principles of tomography to recover 3D volumes of electric potential from a series of 2D projections, and prior contributions in the filed attempting to solve for various kinds of 3D samples. In the following chapters, new approaches will be introduced, followed by derivations, simulation as well as experimental validations, and discussions. Through the steps, we wish to illustrate the importance and power of modeling multiple scattering such that better reconstruction quality maybe attained. Specifically, the rest of the dissertation is organized as follows:

- Chapter 2 carries out the theoretical formulation of our tomography framework. We first derive the multi-slice scattering model that is capable of accounting for multiple-scattering events between the electron probe and the sample. Then, we write the image formation process that relates the image captured to the 3D sample on the tilt stage. In the end, we provide the method for solving the inverse problem.

- Chapter 3 shows our effort to validate the proposed framework via simulation. To be rigorous, we simulated a dense atomic model that contains both crystalline and amorphous structures of Silicon and Silicon Oxide with over 100,000 atoms. In the reconstructed volume, we traced all atoms and compared with ground truth to check the quality. Through varying a set of parameters such as dosage, missing wedge degrees, regularization techniques and tilt-defocus trade offs, we found the the best experimental parameters that provide practicality and superior quality simultaneously.

- Chapter 4 illustrates our successful application of the framework on experimental datasets for discoveries in materials science. The main subject in this study is a large volume of Montmorillonite clay immersed in aqueous solution. The material is extremely dose sensitive and the experiment was performed under Cryogenic temperature. Through a series of simulation and experimental validation, we demonstrated the possibility of inferring solution ion distributions from the absorbance profile in the 3D reconstruction. For the first time, we were able to observe the electric double layer (EDL) of this material in 3D.

- Chapter 5 explains the effort in building the driving force behind all the achievements – computation. Throughout the dissertation, many optimizations in the algorithm have been implemented that significantly improved the efficiency of the reconstruction algorithm and hence increased the overall throughput of the framework, allowing larger space-bandwidth product.

- Chapter 6 concludes the dissertation by summarizing my contributions as well as pointing out a few potential future directions.

Chapter 2

Theory

In this chapter, we present a new method for high-resolution 3D transmission electron microscopy (TEM) which reconstructs the electrostatic potential of a sample at atomic resolution in all three dimensions. We use phase contrast images captured through-focus and at varying tilt angles, along with an implicit phase retrieval algorithm that accounts for dynamical and strong scattering, to provide more accurate 3D reconstruction results with much lower electron doses than current atomic electron tomography methods. Particularly, we will mathematically derive the multi-slice scattering algorithm, outline the image formation process, and describe the overall optimization framework to solve for the inverse problem.

2.1 Modelling a Multiple Scattering Sample

Multi-slice Scattering Model

Throughout this dissertation, the multi-slice scattering model is used. Although it seems very intuitive and ad-hoc from the description in the previous chapter, in this section, we derive the this model formally from first principle. The derivation starts from the time independent Schrödinger equation describing fast travelling electrons in space and their interaction with matter:

$$\left[-\frac{\hbar}{2m}\nabla^2 - eV(x,y,z)\right]\psi_f(x,y,z) = E\psi_f(x,y,z) \tag{2.1}$$

where $\hbar$ is Planck's constant divided by 2π, $m = \gamma m_0$ is the relativistic mass of an accelerated electron with a de Broglie wavelength λ, $e = |e|$ is the unit electron charge, eV is the potential energy, and $E = \frac{h^2}{2m\lambda}$ is the total energy. The time-independent Schrödinger equation models the full wave function $\psi_f(x,y,z)$ of an electron travelling through a inhomogenous medium with electric potential $V(x,y,z)$. Since in a TEM the electrons are propagating in z with very high velocity, the full wave function can be written as a product of a plane wave travelling in z and a counterpart with a slowly varying z profile:

$$\psi_f(x,y,z) = \psi(x,y,z)\text{exp}\left(\frac{i2\pi z}{\lambda}\right) \tag{2.2}$$

As we shall see later, this factorization of the full wave function is beneficial for making certain assumptions of the model and further simplify. We next split the differential operator $\nabla^2 = (\nabla_\perp^2 + \frac{\partial^2}{\partial z^2})$. Then we can expand differential operator in the Schrödinger equation as

$$\begin{aligned}\nabla^2\psi_f(x,y,z) &= \left[\nabla_\perp^2 + \frac{\partial^2}{\partial z^2}\right]\psi_f(x,y,z) && (2.3)\\ &= \text{exp}\left(\frac{i2\pi z}{\lambda}\right)\nabla_\perp^2\psi(x,y,z) + \frac{\partial^2}{\partial z^2}\left[\psi(x,y,z)\text{exp}\left(\frac{i2\pi z}{\lambda}\right)\right]. && (2.4)\end{aligned}$$

We next focus on the second term in Eq.(2.4), using simple product rule:

$$\begin{aligned}\frac{\partial^2}{\partial z^2}\left[\psi(x,y,z)\text{exp}\left(\frac{i2\pi z}{\lambda}\right)\right] &= \frac{\partial}{\partial z}\left[\text{exp}\left(\frac{i2\pi z}{\lambda}\right)\left[\frac{\partial\psi}{\partial z} + \frac{i2\pi z}{\lambda}\psi\right]\right] && (2.5)\\ = \text{exp}\left(\frac{i2\pi z}{\lambda}\right)&\left[\frac{\partial^2\psi}{\partial z^2} + \frac{i4\pi}{\lambda}\frac{\partial\psi}{\partial z}\right] - \frac{i4\pi^2}{\lambda^2}\psi(x,y,z)\text{exp}\left(\frac{i2\pi z}{\lambda}\right) && (2.6)\\ = \text{exp}\left(\frac{i2\pi z}{\lambda}\right)&\left[\frac{\partial^2\psi}{\partial z^2} + \frac{i4\pi}{\lambda}\frac{\partial\psi}{\partial z}\right] - \frac{i4\pi^2}{\lambda^2}\psi_f(x,y,z). && (2.7)\end{aligned}$$

With all of the expansions derived in Eq.(2.4) and Eq.(2.7), we can plug them back into the original Schrödinger equation in Eq.(2.1). After the simplifications, we have:

$$-\frac{\hbar^2}{2m}\left[\nabla_\perp^2 + \frac{\partial^2}{\partial z^2} + i\frac{4\pi}{\lambda}\frac{\partial}{\partial z} + \frac{2meV(x,y,z)}{\hbar^2}\right]\psi(x,y,z) = 0. \tag{2.8}$$

Notice no assumptions have been made made at the moment and Eq.(2.8) still remains the general Schrödinger equation, but simply expanded. Also, notice that Eq.(2.8) is only a function of the slowly z-varying wave function. With this property we are ready to make our first assumption: $\left|\frac{\partial^2}{\partial z^2}\right| \ll \left|\frac{1}{\lambda}\frac{\partial}{\partial z}\right|$, which allows us to disregard the second order differential term along the z-direction. Then, as we regroup the terms and move the first order differential term to the left, we have:

$$\frac{\partial}{\partial z}\psi(x,y,z) = [\mathbf{A} + \mathbf{B}]\,\psi(x,y,z), \tag{2.9}$$

where $\mathbf{A} = i\frac{\lambda}{4\pi}\nabla_\perp^2$ and $\mathbf{B} = i\sigma V(x,y,z)$. Equations in the form like Eq. (2.9) have a close form solution:

$$\psi(x,y,z) = \text{exp}\left(\int_0^z [\mathbf{A}(z') + \mathbf{B}(z')]\text{d}z'\right)\psi(x,y,0). \tag{2.10}$$

One can easily plug this solution into Eq.(2.9) to verify that it indeed satisfies the differential equation. If the boundary value at $z = 0$ of the wave function at $\psi(x, y, 0)$ is known, any values of $\psi(x, y, z)$ can be calculated. This is especially useful in the context of HRTEM as we are able to control the incident beam on the sample, i.e. a plane wave illumination is used and $\psi(x, y, 0) = 1$. Although there will be some global phase offset associated to the incident beam, it will not affect the general intensity image contrast and thus can be neglected.

Despite the fact that $\psi(x, y, z)$ can be analytically related to $\psi(x, y, 0)$, it is not straight forward to compute the exponential of an integral of differential operators. Therefore, further simplifications are needed in order to compute Eq.(2.9) efficiently. If we assume that z is sufficiently small such that $V(x, y, 0)$ is almost constant between 0 and z, the exponential term in Eq.(2.9) can be evaluated as $\exp\left(i\frac{\lambda}{4\pi}z\nabla_\perp^2 + i\sigma zV(x, y, 0)\right)$. Then, we consider the Taylor series expansion of the exponential term:

$$\psi(x, y, z) = \exp\left(i\frac{\lambda z}{4\pi}\Delta_\perp^2\right)\exp\left(i\sigma zV(x, y, 0)\right)\psi(x, y, 0) + \mathcal{O}(z^2). \tag{2.11}$$

More details about the expansion details can be found in [66]. The exponential term involving lateral differential operators above can be performed in Fourier space by multiplying a quadratic term in the exponent and thus in real space becomes a convolution, and by neglecting the higher order terms, the field $\psi(x, y, z)$ can be expressed as:

$$\psi(x, y, z) = \mathcal{F}^{-1}\left\{\exp(-i\pi\lambda z(k_x^2 + k_y^2)) \cdot \mathcal{F}\left\{\exp\left(i\sigma zV(x, y, 0)\right)\psi(x, y, 0)\right\}\right\}, \tag{2.12}$$

where $\mathcal{F}\{\cdot\}$ and $\mathcal{F}^{-1}\{\cdot\}$ are Fourier transform and inverse Fourier transforms, respectively. After some approximations we have arrived at Eq.(2.12). Notice that Eq.(2.12) can be implemented very efficiently, and it can be split into two steps: a multiplication in real space and a convolution (equivalent to a multiplication in Fourier space). Intuitively, the first step corresponds to refraction where the wave function interacts with the inhomogenous medium. The second step corresponds to a free-space propagation of distance z, and this operation is also known as the Fresnel propagation. Notice that we arrived at this result by assuming that z is sufficiently small. This assumption is always true if we break the 3D sample into a series of slices with separation Δz sufficiently small. Only then, Eq.(2.12) is to be applied recursively to model to scattering process between the electron probe and the 3D sample, and hence the name *multi-slice*.

Because of the recursive nature of the model, the multi-slice method is capable of capturing the multiple scattering events between the probe and the sample, and thereby allowing a thicker sample to be modeled [106, 128, 129, 134, 63]. However, as briefly alluded to in the paragraph above, the accuracy of multi-slice depends on the thickness in between slices, and the electron beam has to propagate mostly forward in order for the approximations to be valid. Therefore, if a sample causes severe back-scattering, the multi-slice model will suffer a decrease in accuracy [20]. To deal with such scenarios, other scattering models have been proposed in the field of optics [20, 78], with a trade off in computation time. Luckily, due to the high speed of travelling electrons in a HRTEM, back-scattered electrons are extremely rare and most electrons are scattered to relatively small angles. Therefore, the multi-slice model is sufficient for most samples that we examine.

Image Formation

We now describe our computational model for the process of the incident beam interacting with the sample and forming each measurement; this *forward model* is used to simulate measurements and will also be crucial to our inverse problem reconstruction. It is composed of three parts: object rotation, complex-wave propagation and imaging. We model the 3D object as a series of projected 2D atomic potential functions $V \triangleq \{V_m(\mathbf{r})\}_{m=1}^{N_z}$, where $\mathbf{r} = (x, y)$ are the lateral coordinates and m is the slice index along the axial direction (z)[66], with slice separation described by a set $\{\Delta z_m\}_{m=1}^{N_z}$.

For each tilt angle, θ_k $(k = 1, 2, ..., N_\theta)$, we rotate the 3D object along the y-axis using a fast rotation algorithm [94]. The tilted object W_k is then $W_k = \mathcal{R}_{\theta_k}\{V\}$, where $\mathcal{R}_{\theta_k}$ denotes a linear rotation operator.

Then, we model the propagation of the complex wave, with relativistically-corrected electron wavelength λ, through the object. We use the multislice algorithm to account for multiple scattering events (see Fig. 2.1(a)). Each slice is converted from a 2D potential function to a 2D transmittance function $t_{k,m}(\mathbf{r}) = \exp[i\sigma W_{k,m}(\mathbf{r})]$, where σ is the beam-sample interaction parameter that depends linearly on λ. Example projected potentials are shown in Fig. 3.1(b),(c).

The complex electron wave function before reaching each slice is denoted by $\psi_{k,m}(\mathbf{r})$. As it passes through the slice, it will be multiplied by the corresponding 2D transmittance function at the corresponding z depth. After that, it is propagated in free space to the next slice using the angular spectrum method:

$$\psi_{k,m+1}(\mathbf{r}) = \mathcal{P}_{\Delta z_m}\{t_{k,m}(\mathbf{r})\psi_{k,m}(\mathbf{r})\}, \tag{2.13}$$

where

$$\mathcal{P}_{\Delta z_m}\{\cdot\} = \mathcal{F}^{-1}\left\{\exp\left[i2\pi\Delta z_m\sqrt{1/\lambda^2 - \|\mathbf{q}\|^2}\right] \cdot \mathcal{F}\{\cdot\}\right\} \tag{2.14}$$

is the linear operator for free-space propagation by distance Δz_m, $\mathbf{q} = (q_x, q_y)$ is the 2D Fourier space coordinates, and $\mathcal{F}\{\cdot\}$ and $\mathcal{F}^{-1}\{\cdot\}$ denote Fourier transform and its inverse, respectively.

The exit wave of a thin sample (in focus) will show primarily amplitude contrast, but most of the electron scattering information is encoded as phase shifts on the exit wave. Because defocus induces phase contrast, we use the free-space propagation operator to defocus the exit waves by distances of $\{\Delta f_j\}_{j=1}^{N_f}$ before capturing the intensity of the exit wave:

$$\hat{I}_{k,j}(\mathbf{r}) = \left|\mathcal{H}\left\{\mathcal{P}_{\Delta f_j}\left\{\psi_{k,N_z+1}(\mathbf{r})\right\}\right\}\right|^2 \triangleq \left|\psi_{\text{exit},k,j}(\mathbf{r})\right|^2, \tag{2.15}$$

where

$$\mathcal{H}\{\cdot\} = \mathcal{F}^{-1}\left\{H(\mathbf{q}) \cdot \mathcal{F}\left\{\cdot\right\}\right\}, \tag{2.16}$$

with $H(\mathbf{q})$ denoting the microscope's transfer function [66], similar to what we saw previously in chapter 1. After all tilt angles and defocus images are acquired, we obtain a series of images denoted as $\{\hat{I}_{k,j}(\mathbf{r})\}_{k=1,j=1}^{N_\theta,N_f}$, examples of which are shown in Fig. 3.1(d),(e). The multislice beam propagation method is outlined in Algorithm 1 and schematics are shown in Fig. 2.1(a).

Algorithm 1 Forward model computation

Input: Initial wave function $\psi_0(\mathbf{r})$, 3D rotated atomic potentials $W = \{W_m\}_{m=1}^{N_z}$, slice separations $\{\Delta z_m\}_{m=1}^{N_z}$, defocus angles $\{\Delta f_j\}_{j=1}^{N_f}$, and interaction parameter σ.

```
$\psi_1(\mathbf{r}) \leftarrow \psi_0(\mathbf{r})$
for $m \leftarrow 1$ to $N_z$ do                                              ▷ Beam propagation
    $t_m(\mathbf{r}) \leftarrow \exp\left[i\sigma W_m(\mathbf{r})\right]$
    $\psi_{m+1}(\mathbf{r}) \leftarrow \mathcal{P}_{\Delta z_m}\left\{t_m(\mathbf{r}) \cdot \psi_m(\mathbf{r})\right\}$
end for
for $j \leftarrow 1$ to $N_f$ do                                              ▷ Defocus and image
    $\psi_{\text{exit},j}(\mathbf{r}) \leftarrow \mathcal{H}\left\{\mathcal{P}_{\Delta f_j}\left\{\psi_{N_z+1}(\mathbf{r})\right\}\right\}$
end for
```

Return: Predicted exit wave $\{\psi_{\text{exit},j}(\mathbf{r})\}_{j=1}^{N_f}$ and intermediate wave function $\{\psi_m(\mathbf{r})\}_{m=1}^{N_z}$.

Figure 2.1(b) shows examples of simulated HRTEM plane-wave images when the sample is tilted at many angles and multiple defocus images are simulated.

2.2 Inverse Problem

Convex optimization

Given a set of intensity-only measurements, we estimate the potential, V, by solving an optimization problem. Starting with an estimated potential V, we use our forward model to generate a series of predicted measurements $\{\hat{I}_{k,j}(\mathbf{r})\}_{k=1,j=1}^{N_\theta,N_f}$. We formulate an error function to quantify the difference between predicted and actual measurements $\{I_{k,j}(\mathbf{r})\}_{k=1,j=1}^{N_\theta,N_f}$. The goal is to find the 3D atomic potential that fits the intensity measured and thus minimizes the error:

$$\begin{aligned} V &= \arg\min_V \sum_{k=1}^{N_\theta} \sum_{j=1}^{N_f} e_{k,j}^2 \\ &= \arg\min_V \sum_{k=1}^{N_\theta} \sum_{j=1}^{N_f} \left\| \sqrt{I_{k,j}(\mathbf{r})} - \sqrt{\hat{I}_{k,j}(\mathbf{r})} \right\|_2^2, \end{aligned} \tag{2.17}$$

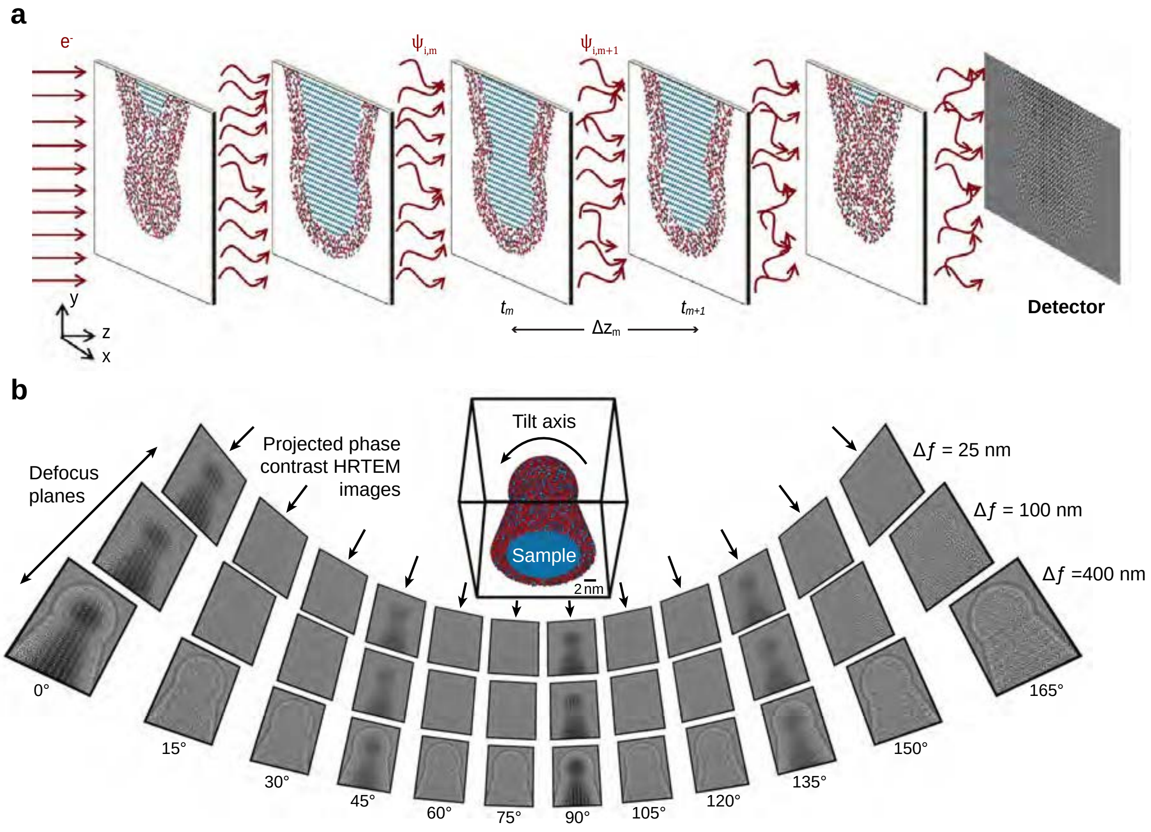

Figure 2.1: Foward measurements for phase contrast atomic electron tomography experiment with a core-shell SiO_2 needle geometry. (a) The multislice forward model treats the 3D sample as a series of 2D slices separated by propagation, thus accounting for multiple scattering. (b) The sample is tilted with respect to the electron beam to capture plane-wave illuminated images at varying angles (up to 180°). For each tilt angle several HRTEM images are recorded at different focus planes.

where $\|\cdot\|_2$ is the l_2 norm. Instead of directly comparing the difference between the predicted and actual intensity measurements, we compare the square roots of the intensity, which correspond to the amplitude of the exit waves. This is because the amplitude-based error function better accounts for Poisson-distributed noise (whereas a intensity-based error function would be ideal for Gaussian-distributed noise) [142]. In this study, the low electron dose means that Poisson noise dominates.

Algorithm 2 Error backpropagation for gradient computation

Input: Residual vectors $\{r_j(\mathbf{r})\}_{j=1}^{N_f}$, intermediate wave functions $\{\psi_m(\mathbf{r})\}_{m=1}^{N_z}$, 3D rotated atomic potentials W, slice separations $\{\Delta z_m\}_{m=1}^{N_z}$, defocus angles $\{\Delta f_j\}_{j=1}^{N_f}$, and interaction parameter σ.

```
$\phi_{N_z+1}(\mathbf{r}) \leftarrow 0$
for $j \leftarrow 1$ to $N_f$ do                                    ▷ Refocus to end of sample
    $\phi_{N_z+1}(\mathbf{r}) \leftarrow \phi_{N_z+1}(\mathbf{r}) + \mathcal{P}_{-\Delta f_j}\left\{\mathcal{H}^\dagger\left\{r_j(\mathbf{r})\right\}\right\}$
end for
for $m \leftarrow N_z$ to 1 do                                      ▷ Backpropagation
    $\phi_m(\mathbf{r}) \leftarrow \mathcal{P}_{-\Delta z_m}\left\{\phi_{m+1}(\mathbf{r})\right\}$
    $t_m^*(\mathbf{r}) \leftarrow \exp\left[-i\sigma W_m(\mathbf{r})\right]$
    $g_m(\mathbf{r}) \leftarrow -i\sigma t_m^*(\mathbf{r}) \cdot \psi_m^*(\mathbf{r}) \cdot \phi_m(\mathbf{r})$
    $\phi_m(\mathbf{r}) \leftarrow t_m^*(\mathbf{r}) \cdot \phi_m(\mathbf{r})$
end for
```

Return: Estimated gradient $\nabla_W e_i \triangleq \{g_m(\mathbf{r})\}_{m=1}^{N_z}$.

We solve the optimization problem with an accelerated gradient method outlined in Algorithm 3. For each tilt angle and defocus, We first tilt the estimated sample and predict the intensity using the multislice algorithm outlined in Algorithm 1. Next, we minimize Eq. (2.17) by differentiating the error with respect to each slice of V. This is done by recursively applying the chain rule to calculate the gradient, and we refer to this process as the backpropagation. The back propagation is illustrated in Algorithm 2, and it is derived in the appendix. Notice that the symmetry between Algorithms 1 and 2 is a key signature in many non-linear optimization methods. Then, we perform a regularization process that enforces prior knowledge we have about the sample (details discussed later). The last step in the loop is that we apply Nesterov's acceleration, which adds a momentum factor in the gradient update to improve convergence speed. By repeating these steps, we finally reach a converged estimate of V and terminate Algorithm 3. Notice that the reconstruction algorithm implicitly solves the phase retrieval problem in the gradient calculation. Line 7 of Algorithm 3 closely resembles the traditional Gerchberg-Saxton type phase retrieval method by applying an amplitude substitution to the residual error [40].

Algorithms that assume lattice types and occupancies inevitably preclude detection of small scale spatial variations. Notice that during the reconstruction, we do not assume any structural priors on the sample. Thus, our method is robust enough to show vacancies and defects when they are present in the sample. In contrast to [130, 129], we also do not assume specific shapes of the individual atoms.

Algorithm 3 Iterative reconstruction

```
Input: Tilt angles $\{\theta_k\}_{k=1}^{N_\theta}$, measured intensity images $\{I_{k,j}\}_{k=1,j=1}^{N_\theta,N_f}$, interaction parameter $\sigma$, step size $\alpha$, and maximum iteration $N_s$.
$U^{(1)} \leftarrow 0,\ V^{(0)} \leftarrow 0,\ \beta^{(1)} = 1$
for $s \leftarrow 1$ to $N_s$ do                                        ▷ Outer loop
    for $k \leftarrow 1$ to $N_\theta$ do                               ▷ Object rotation
        $W_k = \mathcal{B}_{N_B}\left\{\mathcal{R}_{\theta_k}\left[U^{(s)}\right]\right\}$
        $\left(\{\psi_{\text{exit},k,j}\}_{j=1}^{N_f}, \{\psi_{k,m}\}_{m=1}^{N_z}\right) \leftarrow$ run Algorithm 1 with $W_k$
        for $j \leftarrow 1$ to $N_f$ do                                ▷ Compute residual
            $r_{k,j} \leftarrow \psi_{\text{exit},k,j} - \sqrt{I_{k,j}}\frac{\psi_{\text{exit},k,j}}{\left|\psi_{\text{exit},k,j}\right|}$
        end for
        $\nabla_W e_k(U^{(s)}) \leftarrow$ run Algorithm 2 with $\{r_{k,j}\}_{j=1}^{N_f}$, $\{\psi_{k,m}\}_{m=1}^{N_z}$, and $W_k$
        $U^{(s)} \leftarrow U^{(s)} - \alpha\mathcal{R}_{\theta_k}^{\dagger}\left\{\mathcal{B}_{N_B}^{\dagger}\left[\nabla_W e_k(U^{(s)})\right]\right\}$
    end for
    $V^{(s)} \leftarrow \text{prox}\left(U^{(s)}\right)$                 ▷ Regularization
    $\beta^{(s+1)} \leftarrow \frac{1+\sqrt{1+4(\beta^{(s)})^2}}{2}$     ▷ Nesterov acceleration
    $U^{(s+1)} \leftarrow V^{(s)} + \left(\frac{\beta^{(s)}-1}{\beta^{(s+1)}}\right)\left(V^{(s)} - V^{(s-1)}\right)$
end for
Return: Estimated atomic potential $V^{(s)}$.
```

Gradient derivation

In this section, we derive the details of our approach to solve for the inverse problem in vectorized notation. First, we discretize the coordinate system into N_x and N_y pixels for $\mathbf{r} = (x, y)$ respectively. We sample all 2D functions at these discrete coordinates. Then, we raster-scanned the samples into column vectors in $\mathbb{R}^{N_x N_y}$. In addition, linear operators $\mathcal{H}, \mathcal{P}, \mathcal{F}$ can be represented by matrices $\mathbf{H}, \mathbf{P}, \mathbf{F} \in \mathbb{C}^{N_x N_y \times N_x N_y}$

For a given tilt angle θ_k and defocus f_j measurement, the error function in (2.17) can be expressed as:

$$e_{k,j}^2 = \mathbf{e}_{k,j}^{\dagger}\mathbf{e}_{k,j} \tag{2.18}$$

where $\mathbf{e}_{k,j} = \sqrt{\mathbf{I}_{k,j}} - \sqrt{\hat{\mathbf{I}}_{k,j}}$, and $(\cdot)^{\dagger}$ is the hermitian adjoint of a matrix or a vector. $\mathbf{I}_{k,j}$ is the measured intensity of the image, and $\hat{\mathbf{I}}_{k,j}$ is the estimated intensity through Algorithm 1.

Because the multislice propagation model assumes that the atomic potentials of each layer is independent of each other, we calculate the derivative of $e^2_{k,j}$ with respect to every layer of the potentials $\mathbf{W}_m$ separately by applying the chain rule:

$$\begin{aligned}\nabla_{\mathbf{W}_m} e^2_{k,j}(\mathbf{W}) = \left[\frac{\partial \mathbf{e}^\dagger_{k,j}\mathbf{e}_{k,j}}{\partial \mathbf{W}_m}\right]^\dagger &= \left[\frac{\partial \mathbf{e}^\dagger_{k,j}\mathbf{e}_{k,j}}{\partial \mathbf{e}_{k,j}}\frac{\partial \mathbf{e}_{k,j}}{\partial \mathbf{W}_m}\right]^\dagger \\ &= \left[-2\mathbf{e}_{k,j}\frac{\partial \mathbf{e}_{k,j}}{\partial \mathbf{W}_m}\right]^\dagger .\end{aligned} \tag{2.19}$$

Next, we show the calculation of $\frac{\partial \mathbf{e}_{k,j}}{\partial \mathbf{W}_m}$ using backpropagation. Following (2.13) and (2.15), the derivative of $\mathbf{e}_{k,j}$ with respect to the m^{th} layer $\mathbf{W}_m$ is:

$$\begin{aligned}\frac{\partial \mathbf{e}_{k,j}}{\partial \mathbf{W}_m} = &-\frac{\partial(|\psi_{\text{exit},k,j}|^2)^{1/2}}{\partial|\psi_{\text{exit},k,j}|^2}\frac{\partial \text{diag}(\psi^*_{\text{exit},k,j})\psi_{\text{exit},k,j}}{\partial \psi_{N_z+1}} \\ &\frac{\partial \psi_{N_z+1}}{\partial \psi_{N_z}}\cdots\frac{\partial \psi_{m+1}}{\partial \mathbf{t}_m}\frac{\partial \mathbf{t}_m}{\partial \mathbf{W}_m},\end{aligned} \tag{2.20}$$

where $(\cdot)^*$ denotes complex conjugate, $\text{diag}(\cdot)$ is an operator that puts a vector into the diagonal of a square matrix. Next, we list out the the differential terms in the chain rule in (2.20):

$$\frac{\partial(|\psi_{\text{exit},k,j}|^2)^{1/2}}{\partial(|\psi_{\text{exit},k,j}|^2)} = \frac{1}{2}\text{diag}\left(\frac{1}{|\psi_{\text{exit},k,j}|}\right), \tag{2.21}$$

$$\frac{\partial \text{diag}(\psi^*_{\text{exit},k,j})\psi_{\text{exit},k,j}}{\partial \psi_{N_z+1}} = \text{diag}(\psi^*_{\text{exit},k,j})\mathbf{H}\mathbf{P}_{\Delta f_j}, \tag{2.22}$$

$$\frac{\partial \psi_{N_z+1}}{\partial \psi_{N_z}} = \mathbf{P}_{\Delta z_{N_z}}\text{diag}(\mathbf{t}_{N_z}), \tag{2.23}$$

$$\frac{\partial \psi_{m+1}}{\partial \mathbf{t}_m} = \mathbf{P}_{\Delta z_m}\text{diag}(\psi_{N_z}), \text{and} \tag{2.24}$$

$$\frac{\partial \mathbf{t}_m}{\partial \mathbf{W}_m} = i\sigma\text{diag}(\mathbf{t}_m). \tag{2.25}$$

Combining the terms and apply the complex conjugate operator mentioned in (2.19), we arrive at the gradient of $e^2_{k,j}$ with respect to $\mathbf{W}_m$:

$$\begin{aligned}&\nabla_{\mathbf{W}_m} e^2_{k,j}(\mathbf{W}) = \\ &\quad i\sigma\text{diag}(\mathbf{t}^*_m \cdot \psi^*_{N_z})\mathbf{P}_{-\Delta z_m}\cdots\text{diag}(\mathbf{t}^*_{N_z})\mathbf{P}_{-\Delta z_{N_z}} \\ &\quad \mathbf{P}_{-\Delta f_j}\mathbf{H}^\dagger\text{diag}\left(\frac{\psi_{\text{exit},k,j}}{|\psi_{\text{exit},k,j}|}\right)\left(\sqrt{\hat{\mathbf{I}}_{k,j}} - \sqrt{\mathbf{I}_{k,j}}\right)\end{aligned} \tag{2.26}$$

Notice that computing the gradient is almost equivalent to applying the adjoint operators of the forward propagation to the residual error, hence the name *back propagation*.

If we consider all defocus measurements at tilt angle θ_k, the gradient then becomes:

$$\begin{aligned}\nabla_{\mathbf{W}_m} e_{k,j}^2(\mathbf{W}) = \\ & i\sigma \mathrm{diag}(\mathbf{t}_m^* \cdot \psi_{N_z}^*)\mathbf{P}_{-\Delta z_m} \cdots \mathrm{diag}(\mathbf{t}_{N_z}^*)\mathbf{P}_{-\Delta z_{N_z}} \\ & \sum_{j=1}^{N_f} \mathbf{P}_{-\Delta f_j}\mathbf{H}^\dagger \left(\psi_{\mathrm{exit},k,j} - \mathrm{diag}\left(\sqrt{\mathbf{I}_{k,j}}\right)\frac{\psi_{\mathrm{exit},k,j}}{|\psi_{\mathrm{exit},k,j}|}\right).\end{aligned} \tag{2.27}$$

Notice that in (2.27), the last term is equivalent to an amplitude substitution as a result of using amplitude-based cost function in (2.17), and it coincides with the well-known Gerchberg-Saxton type update term [40].

During back propagation, terms such as $\{\psi_m(\mathbf{r})\}_{m=1}^{N_z}$ and W will be used. However, since they were calculated once in the forward measurement, caching them in the forward propagation is recommended to avoid redundant computation. The specific steps for efficiently computing the gradient are in Algorithm 2 and 3.

Regularization

Although the objective function in Eq.(2.17) accounts for Poisson-distributed noise, the reconstruction quality will still suffer with increased noise. In addition, as we lower the number of measurements, the inverse problem becomes more ill-posed. We use a regularization scheme to incorporate *a priori* knowledge that can mitigate this problem. The regularized cost function is:

$$V = \arg\min_V \left\{ \sum_{k=1}^{N_\theta}\sum_{j=1}^{N_f} e_{k,j}^2 + \tau R(V) \right\}, \tag{2.28}$$

where $R(\cdot)$ is a general penalty function, and τ is a tuning parameter for the strength of regularization.

We tested several common types of regularization methods. LASSO (also known as l_1) regularization, where $R(V) = \|V\|_1$, promotes sparsity in the natural domain and is extensively used in statistical parameter estimations [95]. Total Variation (TV) regularization [109], where $R(V) = \|\mathcal{D}\{V\}\|_1$, with $\mathcal{D}\{\cdot\}$ denoting the finite difference operator, is a well-known denoising technique. TV enforces piece-wise smoothness between neighboring pixels by promoting sparsity in the finite difference domain. Since we know that the 3D atomic potential is a smoothly varying function, we choose to implement TV regularization here.

We use a proximal gradient implementation, outlined in Algorithm 3. First, we compute the gradient sequentially using through-focus intensities captured at different angles. Then, we evaluate the proximal operator of the regularization techniques. LASSO regularization has an efficient closed-form evaluation; however, the evaluation for the TV proximal operator is in itself another iterative algorithm [10]. In addition, since we assume in simulation that the atomic potential is purely real and positive (*i.e.* no absorption of the electron beam), we use a positivity constraint to refine our solution space, enforced by performing a projection of the estimate onto real and positive space. In the case where absorption is present, we can remove the constraint without changing the algorithm.

2.3 Summary

In this chapter we laid out the theoretical foundation of the multi-slice scattering model that is capable of capturing the multiple scattering events between the electron probe and the sample. At the same time, we proposed a imaging geometry for the tomography experiment as well as a reconstruction framework. When combined with the appropriate regularization, it is more robust even when low dose tomography problems are solved. In the next few chapters, we will delve in deeper to see the performance of the algorithm through a series of simulation as well as experimental datasets.

2.4 List of Symbols

In this second we list all of the symbols defined and used in the article for reader's convenience, shown in Table 2.1.

Symbol	Description
Coordinates	
$\mathbf{q} = (q_x, q_y)$	Frequency coordinates
$\mathbf{r} = (x, y)$	Spatial domain lateral coordinates
z	Spatial domain axial coordinate
Indices	
j	Defocus index
k	Tilt angle index
m	Slice index in axial direction
s	Iteration counter
Constants	
i	Imaginary unit, where $i^2 = -1$
$I_{k,j}$	2D true intensity measurement of k^{th} tilt and j^{th} defocus
N_B	Slice-binning factor
N_f	Number of defocus measurements per tilt
N_s	Number of optimization iterations
N_x, N_y	Number of lateral pixels
N_z	Number of slices in axial direction
N_θ	Number of tilt angles

α	Optimization step size
β	Optimization acceleration factor
τ	Optimization regularization parameter
λ	Electron wavelength
σ	Beam-sample interaction parameter
ψ_0	2D collimated electron beam

Variables

$e_{k,j}$	2D residual error between estimated and true measurement of k^{th} tilt and j^{th} defocus
f_j	j^{th} defocus distance
g_m	2D gradient update for W_m
$\hat{I}_{k,j}$	2D estimated intensity measurement of k^{th} tilt and j^{th} defocus
$r_{k,j}$	2D intermediate residual error
$t_m(\cdot)$	2D transmittance function corresponding to W_m
U	3D optimization acceleration momentum
V	3D volume of projected slices
V_m	2D m^{th} projected slice of V
W	Rotated 3D volume of projected slices
W_m	2D m^{th} projected slice of W
θ_k	k^{th} tilt angle
Δz_m	Separation distance between W_m and W_{m+1}
ϕ_m	2D residual error backpropagated to m^{th} layer
ψ_m	2D electron beam forward propagated to m^{th} layer
ψ_{exit}	2D exit wave

Operators

$\mathcal{B}\{\cdot\}$	Binning operator (subscript denotes binning factor)
$\mathcal{D}\{\cdot\}$	Finite difference
$\mathcal{F}\{\cdot\}$	Fourier transform
$\mathcal{H}\{\cdot\}$	System transfer function
$\mathcal{P}\{\cdot\}$	Free space propagation (subscript denotes distance)
$\mathcal{R}\{\cdot\}$	Rotation operator (subscript denotes tilt angle)

Chapter 3

Phase Contrast Atomic Electron Tomography with Multiple Scattering Phantoms

Here, we test our algorithm using simulated images of a synthetic needle geometry dataset composed of an amorphous silicon dioxide shell around a silicon core. By simulating various levels of electron dose, tilt and defocus, missing projections, and regularization methods, we identify a configuration that allows us to accurately determine both atomic properties by using an atom-tracing algorithm that is capable of identifying individual atoms as well as estimating their sub-voxel 3D positions and chemical species. We show that with an efficient regularization scheme that exploits the well-known structure of atoms, we can obtain a physically-accurate result, even with very low signal-to-noise ratio (SNR). We also test the ability of our method to recover randomly positioned vacancies in light elements such as silicon, and to accurately reconstruct strongly-scattering elements such as tungsten. After reconstruction, we use our proposed method to enable imaging samples that contain weakly scattering elements such as carbon, oxygen or even lithium, with either crystalline or amorphous structures, or a mix of both.

3.1 Phantom Generation

We consider a two-component sample structure, with a tip geometry similar to the experiment described in [140] (and shown in Fig. 3.1). The structure consists of a crystalline silicon core and a silicon dioxide outer layer. The crystalline core has a tip diameter of approximately 10 nm, as in experiments [119]. A 2 nm thick shell of SiO_2 surrounds the entire Si tip. The SiO_2 coordinates were taken from the SiO_2 structure given in [145], which were computed using Density Functional Theory (DFT). Additionally, a 1.2 Å minimum distance was enforced between the atomic positions of the Si core and SiO_2 shell. In total, 150,847 atoms are present in the structure. A slice of the atomic coordinates are plotted in

Fig. 3.1(a), showing the core-shell structure.

The overall structure of this sample is complex. It contains both fully crystalline and fully amorphous regions along the beam direction for all projection directions. Also, while silicon scatters the electron beam with a moderate cross-section, oxygen atoms scatter only weakly. Finally, the amorphous SiO_2 structure has an Si-O bond length of approximately 1.6 Å [28], making it challenging to resolve the individual atoms in this structure. Hence, this is a challenging test object with realistic length scales for AET reconstruction algorithms.

3.2 Imaging Modality

Data is captured using the simplest TEM measurement protocol: plane-wave illumination, typically referred to as HRTEM or phase contrast imaging. Using a modern TEM instrument equipped with hardware aberration correction, we can image the sample with very little aberrations and sufficient coherence for atomic resolution imaging [46, 9].

To capture phase information, we use through-focus HRTEM images at each tilt (rotation) angle. Defocusing the electron wave increases contrast and delocalizes the atomic signal (see Figure 2.1(b)). In this near-field, or Fresnel diffraction regime, each image is high-pass filtered by the microscope, and the measured signal is modulated by the CTF [66], which can lead to spatial frequency pass-bands or contrast inversions.

The sample is mounted on a tilt-rotation stage so that it can be rotated with respect to the electron beam. For the tip sample considered here, a full tilt range of 180° has been demonstrated [140] with the TEAM stage [26]. However, most electron tomography experiments have a “missing wedge” of tilt angles where the sample geometry or stage prevent measurements at some projection angles. Therefore, we consider both the full-angle and the missing wedge situations. When the tilt direction is closely aligned with the crystalline silicon region of the sample (the low-index zone axis imaging conditions), strong image contrast is observed (Fig. 3.1(d)).

To image the sample with minimal damage, a low dose is required, resulting in noisy measurements. The noise can be modeled by an electron counting process, with each pixel incurring Poisson noise with mean $\{\hat{I}_{k,j}(\mathbf{r})\}_{k=1,j=1}^{N_\theta,N_f}$. Figures 3.1(f) and (g) illustrate a measurement process with a total electron budget of 7,000 electrons/Å^2, which is equivalent to approximately 40 electrons/Å^2 when distributed across 60 tilt angles having 3 defocused images each.

In the meantime, we choose parameters that can be realistically achieved in experiments:

Electron energy: In order to achieve very high resolution, we use an electron accelerating voltage of 300 kV (de Broglie wavelength 0.0197 Å), as in [140, 141]. While SiO_2 is known to be sensitive to the electron beam, it has been imaged previously using 300 kV HRTEM [88, 54, 71, 48].

Voxel size: The voxel size of 0.5 Å (isotropic in all three dimensions) gives a good balance between resolution and field-of-view (FoV), with consideration for practical limits on computation. This voxel size can resolve individual atoms in the amorphous SiO_2 structure

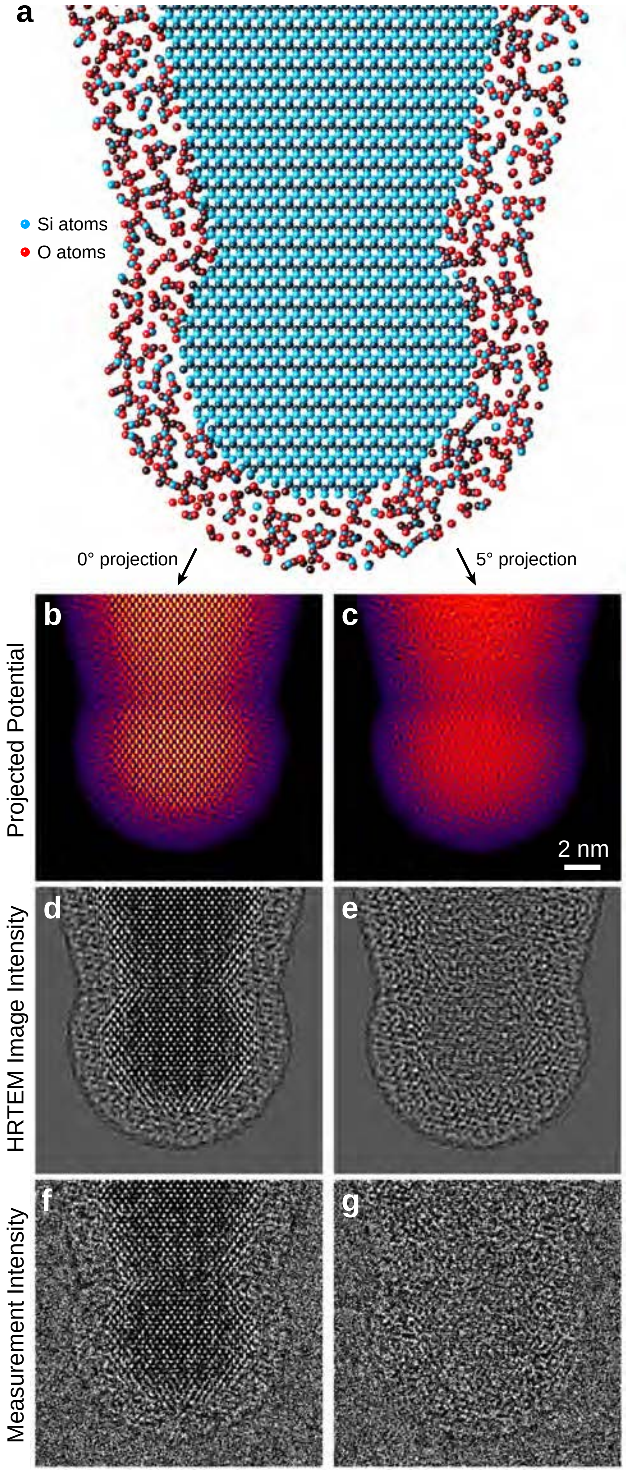

Figure 3.1: HRTEM simulation of the SiO_2 model. (a) A slice of the atomic structure, perpendicular to the electron beam direction. (b) The summed 2D projected potential of the object at 0° and (c) 5° rotation, with intensity scaled to show the weakly scattering edges. (d),(e) Noise-free (infinite dose) HRTEM images at 100 nm defocus for (b) and (c), respectively. (f),(g) Noisy versions of the same images, simulating a dose of 40 electrons/Å^2.

(average Si-O bond length of 1.6 Å). Our reconstruction volume is computationally limited to (24nm)^3, corresponding to $480^3 = 1.1 \cdot 10^8$ voxels, which requires 422 MB of storage space for each full array at single floating point precision. Because we operate in complex space, the storage size requirements double to 844 MB. Without loss of generality, our final reconstruction volume contains a large majority of the sample, which includes approximately 120,000 atoms. In the appendix we show that this voxel size is sufficient by reconstructing from measurements generated with a much smaller voxel size of 0.1 Å.

Tilt angles: Due to the nonlinearity of multiple scattering, choosing the optimal set of tilt angles analytically is not possible. However, we can get a good estimate by using a linear approximation (single scattering) from optical diffraction tomography [86, 62], which treats each tomographic measurement as coming from a particular subspace of the sample's 3D Fourier spectrum (specifically, a parameterized 2D surface). Crystalline samples have distinct preferred measurement directions, but amorphous materials do not [24]. Since our sample contains both, we choose tilt angles that are equally spaced, in order to evenly span Fourier space. In simulations, we mimic experimental limitations by simulating the effect of a missing wedge where some range of tilt angles are missing.

Defocus: As few as two measurements taken at different focus positions can provide phase information [122]. More images will improve the phase result, but must be traded off against dose, data size and capture time. Linearly-spaced focus steps have been shown to be an inefficient scheme for capturing all spatial frequencies; instead, we use exponentially-spaced focus steps [61]. Positive and negative defocus provide essentially identical information about the sample (up to a sign difference) for aberration-corrected microscopes, so we defocus the electron wave in one direction only. As a practical issue, we further restrict the defocus to small enough magnitudes to enable easy translation alignment of multiple images. Due to the increased signal delocalization, large defocus values also require a larger FoV and correction of any magnification or rotation errors, which would increase complexity.

3.3 Results

After using the algorithm described above to reconstruct the atomic potentials, the final step is to use the depth or size of the atomic potential wells to estimate the atomic coordinate positions and classify the atomic species. We have adopted a similar atomic refinement strategy as previous AET studies [140, 141], which is referred to as "atom tracing." First, the reconstructed volume is filtered with a smoothing kernel - a 3D Gaussian distribution with a standard deviation of 0.5 voxels minus another Gaussian distribution with a standard deviation of 1 voxel, normalized to zero total amplitude. Next, the local maxima are recorded as candidate atomic sites. These site positions are refined by fitting a 3D Gaussian function using nonlinear least squares. Next, the fitted intensities are subtracted from the reconstructed volume and candidate atomic sites are added by again filtering with a smoothing kernel and finding local maximum.

Next, an iterative fitting routine proceeds; for each atomic candidate, the nearest-neighbor site intensities are subtracted from the reconstructed volume. In this subtracted volume, nonlinear least squares is used to refine the 3D Gaussian function. After each of these iterations, several criteria were used to remove atomic coordinates. Any sites with a very low intensity (below 30 V, approximately 10% of the maximum sample potential) or size below 1 voxel were removed, and any sites within 2.25 voxels of another site were merged into a single site. After approximately 12 refinement steps, each reconstruction trial was removing less than 2 atomic sites per iteration, and the root-mean square (RMS) change in atomic positions was less than 0.005 voxels. Note that the thresholds were chosen to give good average performance across all datasets, and were not changed except in one specific instance described below.

To classify atom species, we first generate a histogram of atom intensities. We then fit the histogram curve with a bi-modal Gaussian distribution and choose the intersection of the two Gaussian distributions to be the species classification threshold. All atoms having intensities less than the threshold will be classified as oxygen, and the rest will be classified as silicon.

While the full reconstructed volume contains over 120,000 atoms, we select a smaller volume containing 62,402 atom sites to compare with the ground truth atomic configuration, in order to demonstrate accuracy in atom-tracing and atom identification.

The following sections show the results of varying several experimental or reconstruction parameters. For each, we show a single slice of the normalized reconstructed atomic potential that is perpendicular to the tilt axis. The slice was taken from the thickest part of the protrusion, where the diameter is approximately 12 nm. We plot the atomic coordinates that were correctly found for each slice, and the missing and false positives.

Additionally, we show tetrahedral shapes for each cluster of 5 atoms that formed a tetrahedron, with bond lengths of the 4 corner atoms to the center atom within 0.375 Å of the mean Si-O bond length of 1.6 Å. These tetrahedra help visualize how well the amorphous region of the sample was reconstructed, especially for reconstructions with a lot of noise or artifacts present. This feature classification is an example of the kind of classification measurement that could be performed even in the absence of clear atomic peaks, as is done in structural biology [64].

Next, we show two histograms that quantify how well we trace the individual atoms in Fig. 3.3. The first histogram shows the statistics of atomic potential intensities of identified atoms. The more resolved the two distributions are, the better we have classified the specific types of the atoms. The second histogram shows the errors of the 3D position estimation from the reconstruction. Here, for each identified atom we adopt the root-mean-square (RMS) from all coordinates:

$$\text{Position Error} = \sqrt{(x^* - \hat{x})^2 + (y^* - \hat{y})^2 + (z^* - \hat{z})^2}, \tag{3.1}$$

where x^*, y^*, z^* are the true coordinates and $\hat{x}, \hat{y}, \hat{z}$ are the estimated coordinates. A good reconstruction's histogram has a peak close to 0 and a narrow main lobe. We also show RMS error (ϵ) in all three dimensions.

All reconstructions, unless otherwise stated, are full-angle TV regularized, created from 60 uniformly spaced tilt angles, each with 3 defocus steps (25, 45, 100nm) with total incident electron count of 50,000 electrons / Å^2. The regularization parameter τ in Eq. (2.28) is chosen such that the background noise is suppressed, without over-smoothing (smearing) the adjacent atoms.

Reconstructions are computed on graphics processing unit (GPU) for accelerated computation (12GB NVIDIA Titan X GPU) and the algorithm converges within 40 iterations for all scenarios. The total computation time for the dataset mentioned above is less than 2 hours.

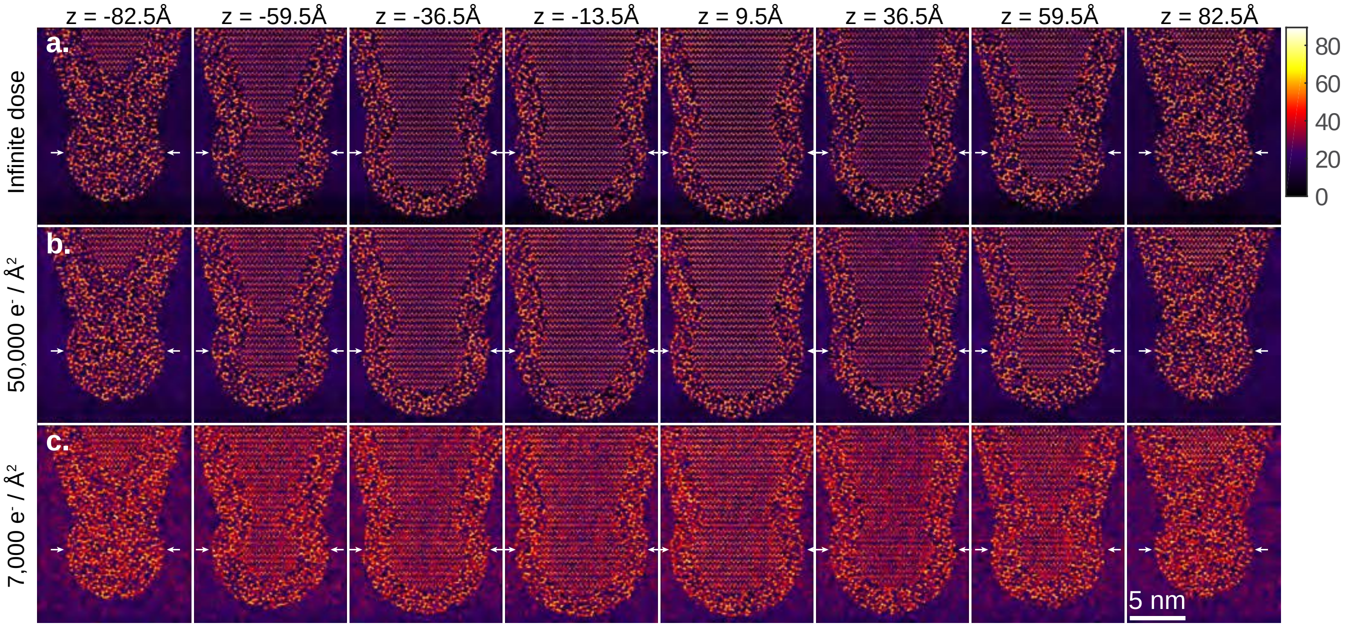

Figure 3.2: Varying dose. Phase contrast AET reconstructions of 1 Å thick 2D atomic potential slices of a simulated Si-SiO$_2$ reconstruction in $x - y$ across multiple z depths, using 60 tilt angles and 3 defocus values per angle. (a) Infinite dose (no noise), (b) 50,000 electrons/Å^2, and (c) 7,000 electrons/Å^2 total dose. Lower dose results in more noise, which causes errors and artifacts in the reconstruction. Each slice shows the square root of the reconstructed potential from 0 to 80 volts and the tilt axis is along the vertical direction. White arrows show location of reconstruction slices for the following sections in FIG. 3.3.

Effect of Electron Dose

In the first set of simulations, reconstructions using different dose budgets are compared to examine how noise affects the algorithm performance. We chose three doses: infinite (noiseless), 50,000 electrons/Å^2, and 7,000 electrons/Å^2. Figure 3.2(a)-(c) shows lateral slices

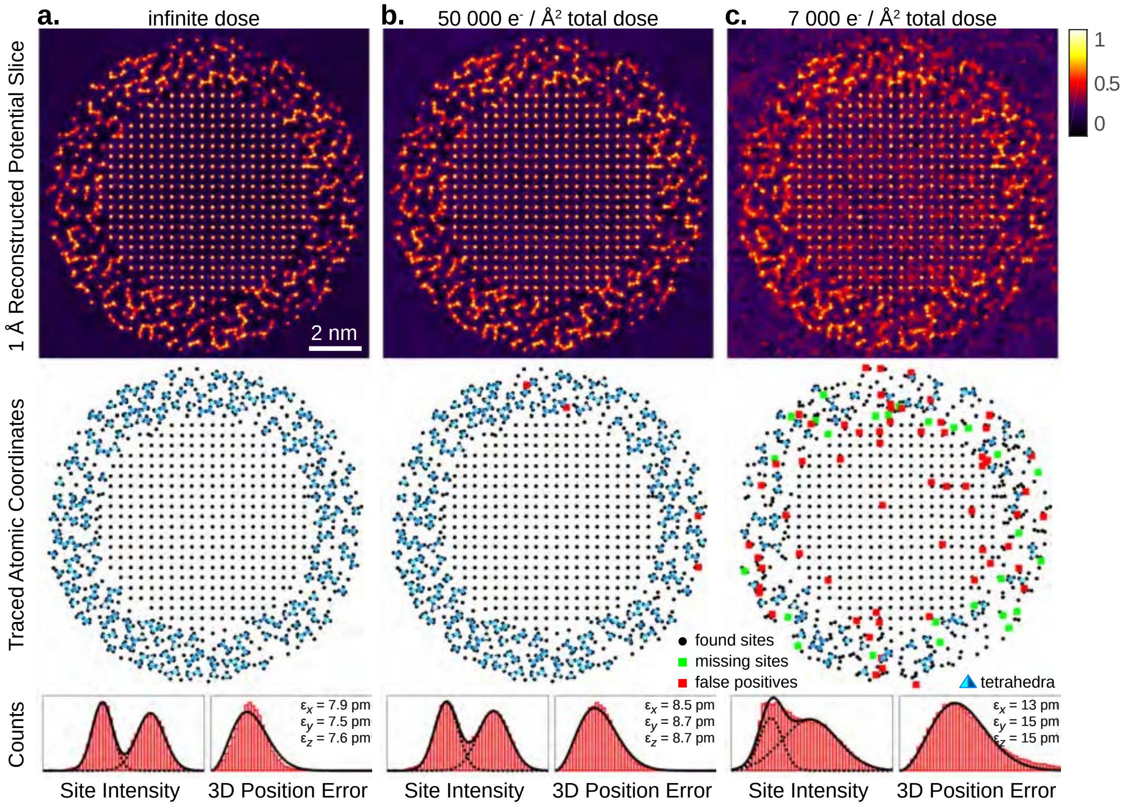

Figure 3.3: Varying dose. Phase contrast AET reconstructions in $z-y$ direction for (a) infinite electron dose, (b) 50,000 electrons/Å^2, and (c) 7,000 electrons/Å^2 total dose. We show (top row) a slice of the normalized reconstructed potential and (bottom row) the corresponding estimated atomic coordinates. Lower dose results in more noise, causing errors in the volume reconstruction, atom identification, and atom classification.

at multiple z depths, taken from simulations with different dose levels. Figure 3.3 shows a 1 Å thick z-x cross-section slice (intensity normalized), where the location is indicated by the white arrows in Fig. 3.2, and atom tracing results. In all reconstructions, the atomic peaks are easily identified. The reconstruction using 50,000 electrons/Å^2 total dose over all tilts and defocused images is nearly identical to the infinite dose reconstruction.

As expected, the reconstruction quality eventually deteriorates as we decrease the dose budget, with the background becoming noticeably more noisy. We cannot increase the regularization to compensate, as it will over-smooth the reconstruction. For the dose level of 7,000 electrons/Å^2, atoms that are too close to each other are smeared together and missing sites increase. Noisy fluctuations in the background lead to an increased number of false positive sites. The noise also causes loss of contrast in the atomic potential intensity, which can be seen from the intensity histogram; the distributions of two types of atoms are less

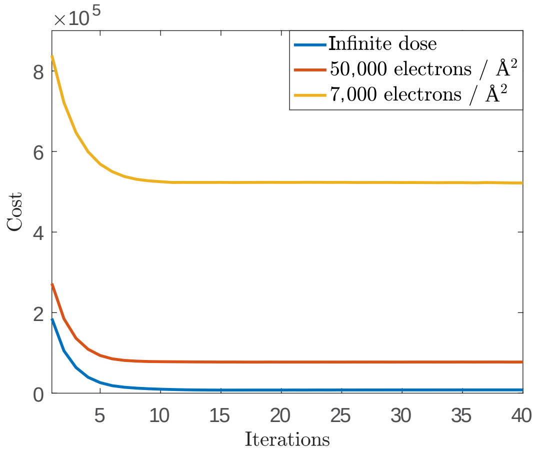

Figure 3.4: Cost function vs iterations to show convergence for various dose budgets. Reconstruction becomes noisier as total dose is lowered, and cost function increases. For each reconstruction, we ensure convergence is achieved.

resolved when dose is decreased, making it harder to classify the species of individual atoms. Finally, the RMS position estimation error increases isotropically as we decrease the dose level.

Figure 3.4 shows the example plots of cost function (Eq. (2.17)) vs iterations. Despite the convergence, as we lower the dose budget, the predicted intensity of the reconstruction has more mismatch with the measured intensity, causing the squared error to increase.

Effect of Number of Tilt and Defocus Measurements

Because total dose is distributed across measurements from all tilt angles and defocus distances, we face a trade-off between number of tilt angles (N_θ) and number of defocus planes (N_f). In this set of simulations, we compare the performance of our method as we vary N_θ and N_f, while keeping the total dose level constant (50,000 electrons/Å^2). Figure 3.5 shows reconstructions from three schenarios: 20 tilt angles (separated by 9°) with 9 defocus planes (20 nm-100 nm in steps of 10 nm), 60 tilt angles with 3 defocus planes (20 nm, 45 nm, and 100 nm), and 180 tilt angles with a single plane at 100 nm. These values give a good balance between using larger defocus values to produce more contrast, but not large enough to make image alignment difficult or lose resolution due to coherence limits.

Comparing Fig. 3.5(a) and (b), we find that using fewer defocus planes and more tilt angles results in a better reconstruction of the sample's structure and improved atom tracing. Particularly in the amorphous SiO_2 region, the number of missing sites is greatly reduced by using more tilt angles. Given that phase can be recovered from a few defocus planes [61], it is reasonable that 9 focus steps are not necessary. However, more focus steps should

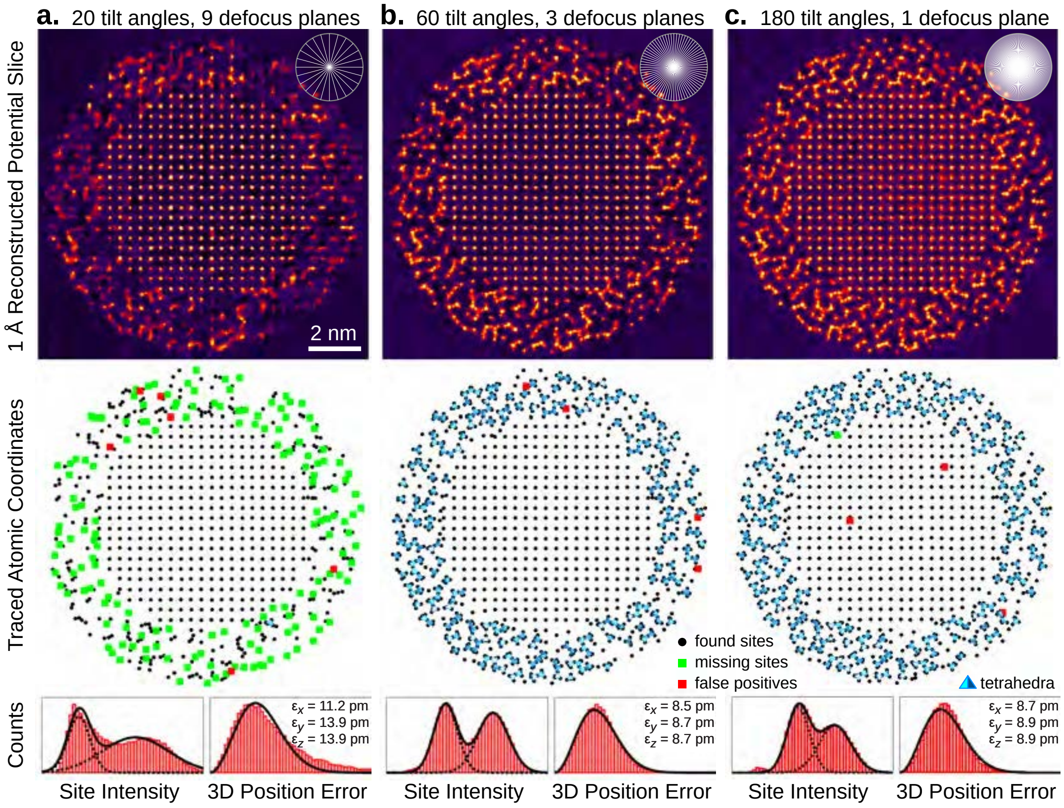

Figure 3.5: Varying the number of tilt angles and the number of defocus planes while keeping a constant total dose. Phase contrast AET reconstructions for (a) 20 tilt angles with 9 defocus planes linearly increasing from 20nm to 100nm, (b) 60 tilt angles with 3 defocus planes at 20nm, 45nm, and 100nm, and (c) 180 tilt angles with single defocus plane at 100nm. The case of 3 defocus planes and 60 tilt angles gives minimal error, offering a good trade-off between number of tilt angles and defocus planes.

help to better reconstruct the atomic potential [61]. For the case of only one defocus plane (Fig. 3.5(c)), the site intensity histograms show that the distributions of the silicon and oxygen atoms are not as well resolved. Hence, the case in Fig. 3.5(b) gives a good tradeoff between accurate structure recovery and good atom classification.

Effect of Missing Tilt Angles

When the tilt-rotation stage is capable of full-angle tomography, isotropic resolution can be achieved in x, y, and z. However, often projection angles are missing due to sample geometry or stage limitations. This means that the coverage of the object's Fourier spectrum is incomplete [86], often described as a "missing wedge". In this section, we test our algorithm

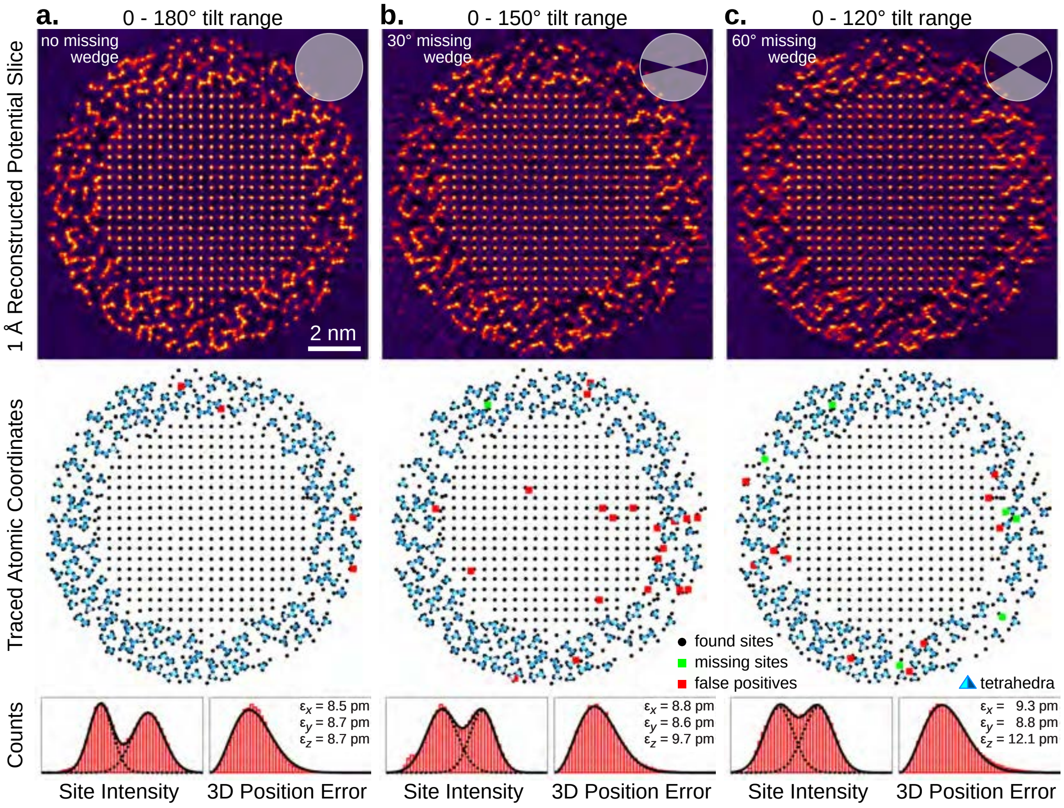

Figure 3.6: The missing wedge problem in the measurements primarily affects the axial accuracy of our reconstructions. All scenarios have the same total dose. Phase contrast AET reconstructions for (a) full tomography data with no missing angle , (b) limited tomography data with 30° missing angle, and (c) limited tomography data with 60° missing angle.

with missing wedges of 30° and 60° (see Fig. 3.6(b) and (c), respectively). Across the accessible tilt angles, the angle separation is constant, such that with constant total dose (50,000 e / Å^2) distributed across all acquisitions, the dose per image increases with the size of the missing wedge.

We find that the missing wedge problem primarily impacts axial resolution. As more angles are missed, the axial resolution deteriorates along the missing wedge direction, increasing errors in atom tracing and identification. Comparing the reconstructions in Fig. 3.6(a) and (c), the portion of missing sites increases from 0.06% to 0.98%. Not only is it harder to identify atoms, it is also more challenging to correctly identify the 3D positions of each atom. The position error histogram in Fig. 3.6(c) suggests that position estimate is less accurate in the axial direction as we increase the missing wedge, while the accuracy in the lateral directions are maintained.

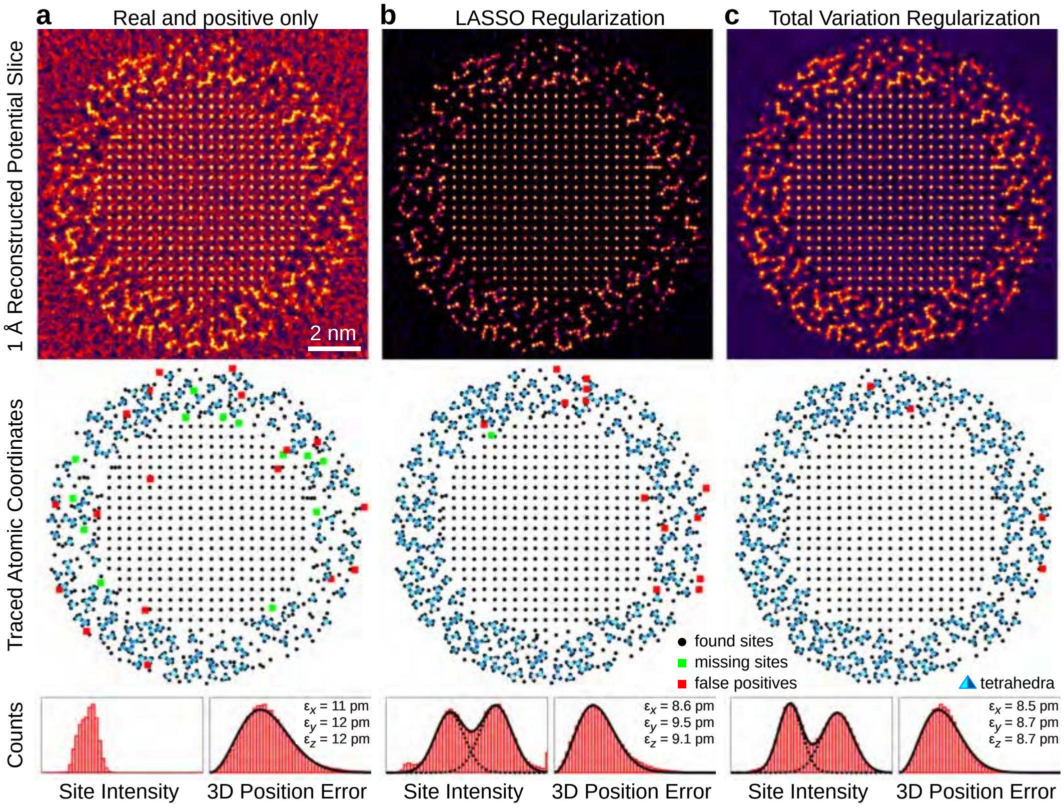

Figure 3.7: Regularization is important for image reconstruction quality. Phase contrast AET reconstructions using (a) real & positivity constraints only, (b) LASSO regularization, and (c) total variation (TV) regularization. In this case, TV regularization provides the best performance.

Effect of Regularization

Regularization allows us to use prior knowledge about the object to refine the solution space and produce better quality reconstructions, even with noisy data. Because low dose is required in order to preserve sample structure during imaging, our raw data suffers from significant (Poisson-distributed) noise. Here, we examine the effectiveness of three different regularization techniques: pure positivity & real constraint, LASSO regularization, and total variation (TV) regularization, as introduced previously.

The results, shown in Fig. 3.7, suggest that regularization plays a significant role in denoising with low-dose measurements. With only real & positivity constraints, the background is too noisy to perform accurate atom tracing and the position estimation error is large in all dimensions. The intensity histogram in Fig. 3.7(a) shows that it also fails to provide two resolved peaks that are needed to perform atom classification.

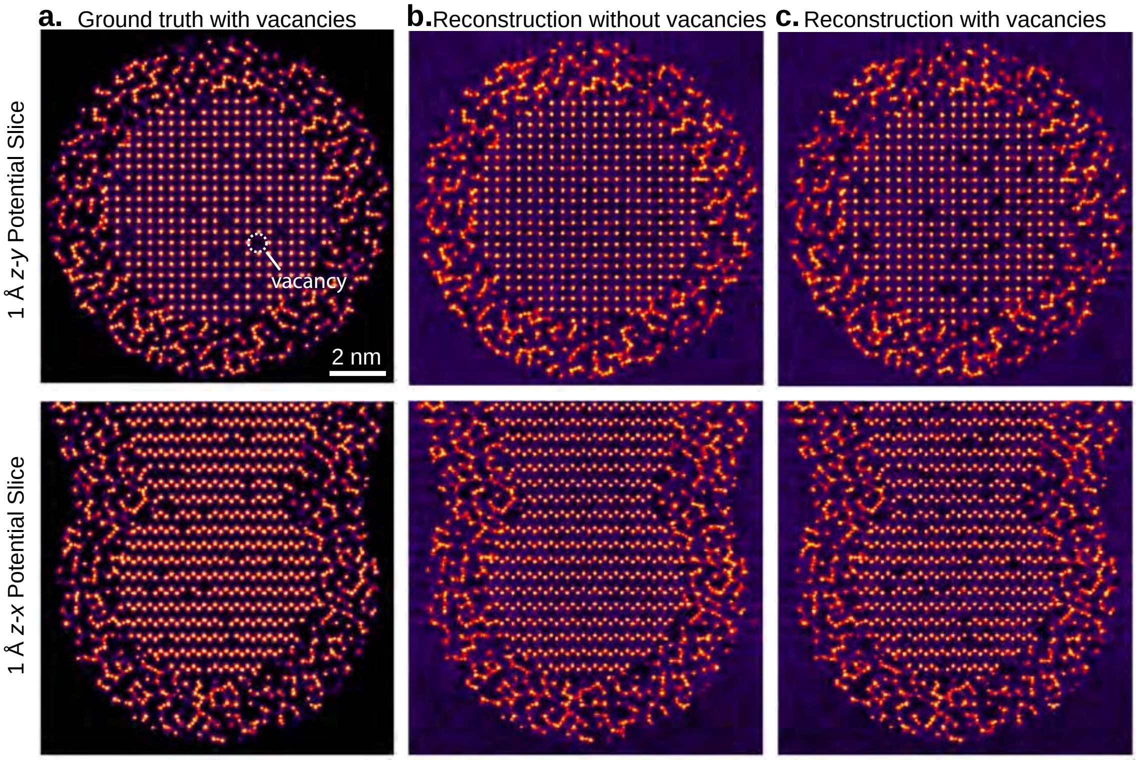

Figure 3.8: Phase contrast AET reconstructions with vacancies. (a) Ground truth atomic potential with vacancies. Reconstruction when (b) no vacancies are present, and when (c) 5% of the atoms are removed in the crystalline region and amorphous region. The top row shows slices in $z - y$ direction, and the bottom row shows slices in $z - x$ direction.

Both LASSO (Fig. 3.7(b)) and TV (Fig. 3.7(c)) regularization significantly improve the quality of the reconstruction. The LASSO reconstruction produces sharp peaks, but shrinks some peak intensities as well as sizes of the potential wells. This leads to a worse distribution of peak intensities, making atomic species classification less accurate. The peak position estimation results are also less accurate for LASSO than TV, as shown in Table 3.1. Therefore we choose to use TV regularization for our reconstructions.

Vacancies in crystalline Si and amorphous SiO_2

Our algorithm is capable of identifying single-atom defects or vacancies in the sample. Here, we validate this claim by simulating a Si-SiO_2 tip sample that contains vacancies. We simulate the vacancies and defects by randomly removing approximately 5% of the atoms in the original sample. Then, with the same geometry and experimental configuration as in Fig. 3.3(b), we reconstruct the atomic potentials of the defected sample. Figure 3.8(a) shows the ground truth atomic potential after the atoms have been removed. The reconstruction result is shown in Fig. 3.8(c). We also refer to Fig. 3.8(b) for the case where no atoms are removed. Samples can still be reconstructed when there are single-atom defects present, because the algorithm does not assume any structural priors.

Robustness against Partial Coherence

In this section, we show the robustness of our algorithm by adding several system imperfections that are frequently encountered in real experiments. Then, we use the proposed framework to reconstruct the 3D atomic potentials of the sample. Specifically, based on the tilt and defocus configuration we have shown in Fig. 3.3(b), we upsample the object and add partially coherent illumination to the measurements.

Upsampling: First, we use a voxel size (0.1 Å) finer than the sensor pixel size (0.5 Å) to generate simulated measurements with accurate diffraction effects. For an object of the same volume $(24\text{nm})^3$, this increases the number of voxels from 480^3 to 2400^3. We then use the multislice model to propagate the electron wave through the finer-grid volume. At the image plane, we bin the pixels to the pixel size of 0.5 Å.

Defocus spread: Next, we simulate the effect of chromatic aberration in the electron beam. In particular, we use a Gaussian spread of focal planes for each tilt and defocus to approximate the effect. A defocus spread of 8 Å is reported in [65], so we choose a somewhat larger Gaussian defocus spread of 20 Å with standard deviation of 10 Å. At the image plane, we use Gaussian weighting to incoherently sum the measurements.

Spatial coherence: We incorporate spatial partial coherence by simulating a 2D Gaussian spread of input scattering angles for each tilt and defocus. Referring to the work in [2], which reported a angular spread of 200 μrad, we choose a Gaussian angle spread of 400 μrad with standard deviation of 200 μrad. At the image plane, we use a 2D Gaussian weighting to incoherently sum the measurements.

Combining all of the effects above, we simulate a series of measurements, which we then use to reconstruct with the fully-coherent framework outlined in 3. Figure 3.9 shows a reconstructed slice. Despite some reconstruction artifacts, we are able to achieve similar atom identification accuracy comparing with the case in Fig. 3.3(b), as shown in Table 3.1. However, clearly the reconstruction artifacts indicated by yellow arrows contribute to the higher false positive rate during atom tracing, so caution should be taken when dealing with real measurements in the future.

Robustness against Heavy Atoms

In this section, we demonstrate that the proposed framework can also be generalized to recover the electrostatic potential distribution of samples that contain both light and heavy atoms. Without loss of generality, we replaced the silicon atoms in the previously synthesized sample with Tungsten atoms, which have larger electrostatic potentials, and thus induces stronger dynamical scattering. The sample closely resembles the one demonstrated in [140].

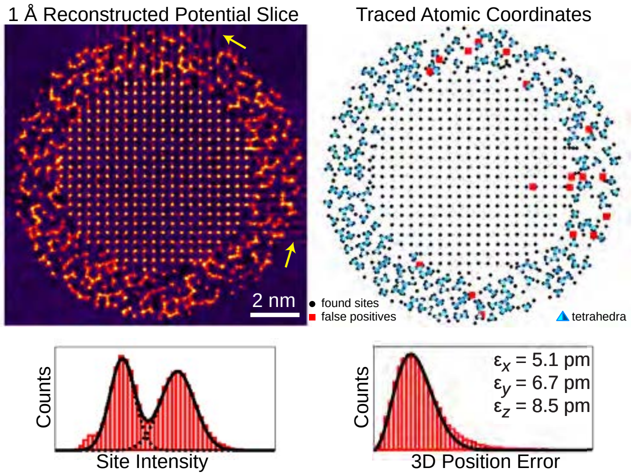

Figure 3.9: Phase contrast AET reconstructions for partial coherence with finer sampling during image calculation formalism. Yellow arrows show reconstruction artifacts due to partial coherence.

From the sample, we simulated the measurements using the same configuration as that of Fig. 3.3(b). Figure 3.10 shows the reconstructed potentials of the Tungsten sample.

As shown in Fig. 3.10(b), while we recover most of the atoms in the Tungsten tip, the reconstruction quality degrades towards the center of the tip, which corresponds to the thickest region of the needle. These artifacts are due to the large amount of accumulated dynamical scattering. As a result, these artifacts in the reconstruction will contribute to error in future atom localization and identification.

Summary of Reconstruction Results

Table 3.1 summarizes all atom tracing and classification results. Note that Fig. 3.3(b), Fig. 3.5(b), 3.6(a), and Fig. 3.7(c) are equivalent and are repeated for convenience. We report mean 3D position error, portion of the atoms correctly found, portion of false positives, and the portion of atoms where the species are correctly labeled.

3.4 Summary

We have tested our reconstruction algorithm for atomic electron tomography applications, from a tilt series of defocused plane-wave HRTEM images. Our nonlinear model takes into account multiple scattering of the electron beam and uses slice-binning and fast rotation and propagation algorithms to decrease the reconstruction time. We show that TV regularization improves the reconstruction quality. Using a simulated sample with both crystalline Si

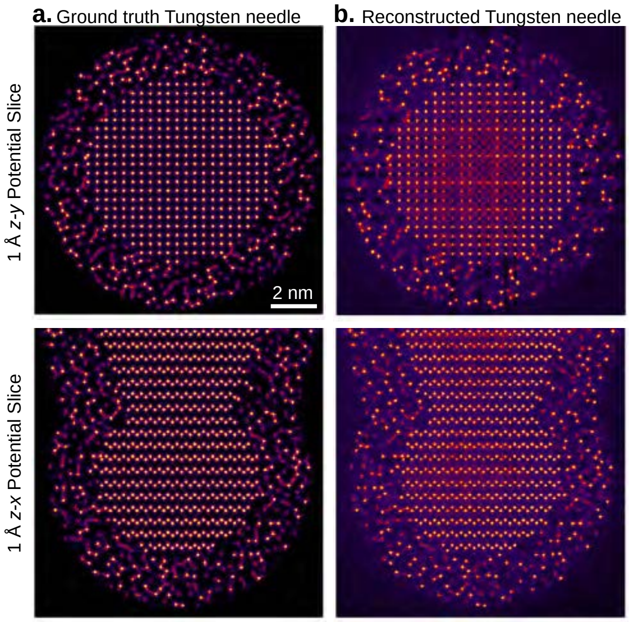
a. Ground truth Tungsten needle
b. Reconstructed Tungsten needle
1 Å z-y Potential Slice
2 nm
1 Å z-x Potential Slice

Figure 3.10: Phase contrast AET reconstructions for tungsten crystalline and tungsten oxide amorphous structure.

and amorphous SiO_2 in a core-shell tip geometry, we have demonstrated accurate atomic reconstructions of more than 60,000 atoms in a sample with a diameter up to 12 nm. Our method is robust to low-dose measurements, works for a small number of defocused images and can handle a large missing wedge of tilt angles. Furthermore, we show that our fully coherent model also works with partial coherent data, both temporally and spatially. The end result is atomic-resolution tomographic reconstruction of nanoscale samples containing both strongly and weakly-scattering elements, with either crystalline or amorphous structures.

Table 3.1: Summary of atom tracing results, out of 62,402 sites in the tip region with a radius ≤ 12 nm diameter.

Figure(s)	Total Dose	N_θ	N_f	Tilt Span	Regularizer	Position Error	Atoms Found	False Positives	Correct Species
3.3(a)	Infinite	60	3	180°	TV	12.51 pm	99.98%	0.00%	98.63%
3.3(b)	50,000 e / Å^2	60	3	180°	TV	13.91 pm	99.94%	0.25%	96.44%
3.3(c)	7,000 e / Å^2	60	3	180°	TV	21.62 pm	95.24%	9.72%	79.79%
3.5(a)	50,000 e / Å^2	20	9	180°	TV	19.11 pm	72.48%	1.26%	82.71%
3.5(b)	50,000 e / Å^2	60	3	180°	TV	13.91 pm	99.94%	0.25%	96.44%
3.5(c)	50,000 e / Å	180	1	180°	TV	14.30 pm	99.97%	0.90%	91.15%
3.6(a)	50,000 e / Å^2	60	3	180°	TV	13.91 pm	99.94%	0.25%	96.44%
3.6(b)	50,000 e / Å^2	60	3	150°	TV	14.34 pm	99.75%	0.44%	94.67%
3.6(c)	50,000 e / Å	60	3	120°	TV	15.84 pm	99.02%	2.02%	90.50%
3.7(a)	50,000 e / Å^2	60	3	180°	Positive	18.65 pm	97.81%	1.82%	46.81%
3.7(b)	50,000 e / Å^2	60	3	180°	Lasso	14.17 pm	99.78%	0.73%	92.53%
3.7(c)	50,000 e / Å^2	60	3	180°	TV	13.91 pm	99.94%	0.25%	96.44%
3.9	50,000 e / Å^2	60	3	180°	TV	10.59 pm	99.95%	1.59%	97.54%

Chapter 4

Cryo Electron Tomography with Multple-scattering Montmorillonite

In this chapter, we test our algorithm proposed in chapter 2 using experimental data of clay minerals. Visualizing the structure of hydrated interfaces is of nearly ubiquitous interest across the physical sciences and is a particularly acute need for layered minerals, whose properties are governed by electrolyte structure, frequently referred to as the electric double layer (EDL), at solution-mineral interfaces. We will show that cryo electron tomography enables direct imaging of the EDL at lithium- and sodium-montmorillonite interfaces with angstrom resolution over micron length scales. Our proposed method reveals ions bound asymmetrically on opposite sides of curved layers, forming a regular structure that we term an ion complexation wave.

4.1 Introduction

Layered minerals control carbon, water, and nutrient transport in the lithosphere [115, 105, 49, 125], promote cloud formation [5] and lubricate fault slip [17, 56, 55] through interactions among charged, hydrated interfaces [53]. Consequently, interactions at the mineral-aqueous interface have been widely investigated using scanning probe microscope (SPM), surface force apparatus (SFA), and X-ray reflectivity measurements from the interfaces of single crystals of mica [58, 15, 74, 81], or atomistic simulations [14]. These techniques inform electric double layer (EDL) models for the distribution of electrolyte among interface-associated (complexed) and diffuse locations that ultimately control the properties of these materials. However, no current model based on the planar geometries typically probed by these techniques generalizes to many natural [27, 139] or colloidal [89, 121, 111, 3] layered mineral systems because layers can swell and exfoliate, creating microstructures that are disordered and evolve over time [85, 116, 126, 138]. While measurements of the potential drop at the interface [35], the change in orientation and density of water molecules [132, 143, 136], and ordering of water and counterions [96, 18, 108, 29, 144] have revealed aspects of planar interfacial structures in increasing detail, direct images of electrolyte distributions where liquids and solids meet that are independent of the intermolecular forces imposed by SPM and SFA, the strongest constraint on theories of hydrated interfaces, have remained elusive.

Direct imaging of interfacial ion complexation in hydrated layered minerals was achieved by cryo electron tomography (cryoET) of lithium-montmorillonite (Li-Mt) suspensions. Mineral and electrolyte distributions were resolved in unprecedented three-dimensional (3D) detail by using the electron tomography framework proposed in 2 that accounts for multiple electron scattering from low-dose images acquired at multiple defocus values at each tilt angle. This enabled the recovery of both the amplitude and phase of the electron exit wave in three dimensions, revealing interfacial structures across the thousands of mineral layers with an isotropic real-space resolution of 3.64 Å, over a 1.02 μm × 0.79 μm × 0.36 μm field of view, shown later in Fig.4.5.

4.2 Simulated Results

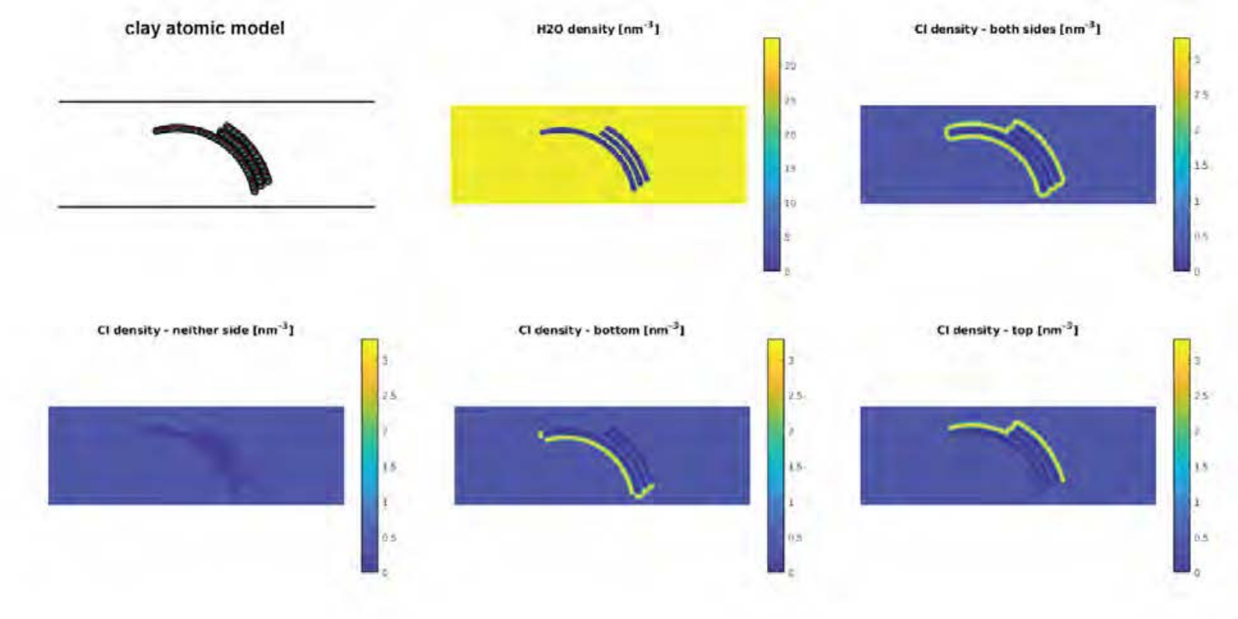

Figure 4.1: Clay stack and ion configurations used for image and tomogram simulation.

To demonstrate the existence of relationship between ion concentration and absorbance of the 3D absorbance of the sample and hence being able to capture relevant information, we first perform a series of tomographic reconstruction simulations of clay structures from their known atomic structures under different scenarios. In the atomic model, not only did we add the crystal clay structures, we also included the solution of LiCl with concentration of 0.75M dissolved in water as a more realistic background.

The clay structure's 3D projection (top view) is shown in Fig. 4.1. The top left inset sub-figure shows the general shape of the simulated structure - it is a stack of 3 layers of clay, and curved to a certain degree to match those observed in real experiments. Since the induced curvature of the clay layers will induce uneven distributions of ions on the two sides of the surface (inner surface and outer surface), we test our algorithm by creating multiple scenarios where excess ions are intentionally added to different sides of the clay surface. Four cases are created: (1) excess Cl ions surrounding all surfaces, (2) no excess Cl ions are added, (3) excess Cl ions added only on the inner surface, and finally (4) excess Cl ions added only on the outer surface. Under these situations, the resulting electric potentials are different and perturbed by the additional electric potential contributed from Cl atoms near the surface.

For each of the 4 scenarios, we create a tilt-series with 3 different defocus values to match what we measure in the real experiment, at an electron energy of 300kV. The three defocus values are -75, -200, and -550nm, respectively, and the tilt range is from -60° to 60° with 1° of increment. Combined together, a total of 363 intensity images are simulated for each case. To make the simulations physically accurate yet computationally feasible, we choose an isotropic voxel size of 0.2Å to render the volume. The simulated volume has a physical dimension of 40 nm(x) × 10 nm(y) × 40 nm (z), corresponding to a total of 2000 × 516 × 2000 voxels.

Since the electron beam travels along the z direction, each intensity image has a size of 512 × 2000 pixels. Then, to match the pixel size of the physical experiment, the simulated images are downsampled 8 times to 64 × 250 pixels, corresponding to a pixel size of 1.6 Å.

The downsampled intensity images are then reconstructed using the method proposed in chapter 2, and the results for 4 individual cases are shown in Fig. 4.2. The first row shows the mean projected image of the volume along the y-axis, and the second and third rows are averaged line tracing profile perpendicular to the surfaces at the isolated and stacked regions, respectively. The integrated density profiles do indeed produce the asymmetric distributions as planned on the isolated layer. These trends were also present in the stacked region, although slight sample-size artifacts reduced interpretability. The implication is that similar analysis can be done and similar phenomenon can be observed with the tomographic reconstruction should there be asymmetric ion distribution along the opposite sides of the clay surfaces.

A total of 60° missing wedge is included in the simulation to match the experiment, causing a lack of isotropic reconstruction resolution. As a result, the artifacts in the reconstructions due to the missing wedge problem is apparent. Also, an absolute relationship between the absorption profile and the ion distribution cannot be mapped out, as the specific absorbance is highly related to the experimental parameters, such as tilt range, sample orientation, and sample geometry. As such, even though the technique is capable of observing the asymmetry, we do not draw a quantitative conclusion.

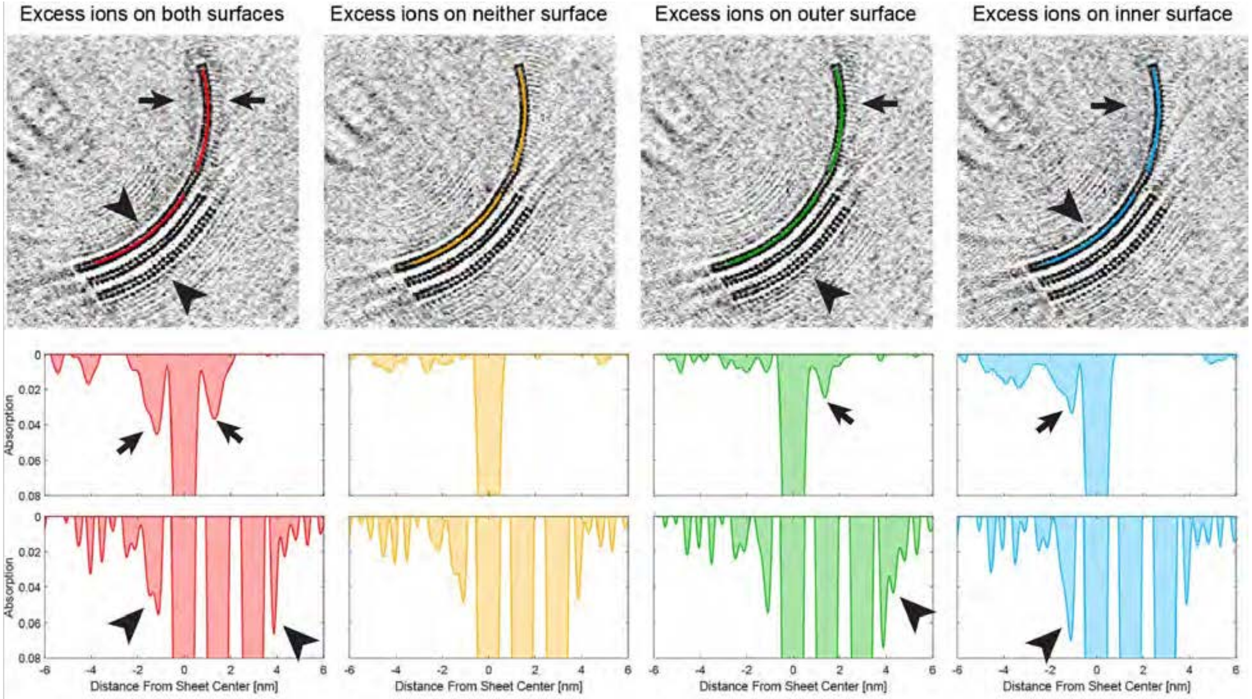

Figure 4.2: Integrated ion density profiles in tomographically reconstructed simulated images from structures shown in Fig. 4.1. Differences in ion densities can be observed between the four cases in 2D slices of 3D reconstructions (top row). These differences are clear in 1D profiles adjacent to the single isolated layer (middle row) and are also present in the stacked region despite the presence of artifacts due to the small reconstruction volume.

4.3 Experimental Results

Next, we demonstrate our 3D volume reconstruction on real lithium-montmorillonite clay samples. In the experimental samples, the clay layers are closely stacked together, and since the stacked clays are no longer isolated, one can influence another. However, this is the exact benefit of the the volumetric tomography as they offer dynamic behavior that other methods such as x-ray crystallography do. As a result, we can see both the nanoscopic structure as well as the implication of it on a macroscopic level. In this section, we first go over the sample preparation and dataset acquisition process. Then, we present our pre-processing techniques such as image normalization and coarse image alignment. After that, we show the volumetric reconstruction along with our scientific findings on EDL observation through clay surface analyses.

Sample Preparation and Image Acquisition

Suspensions of Li-Mt and Na-Mt with mineral concentrations of 5 mg/mL were deposited as 3 µL aliquots onto 200-mesh lacy carbon Cu grids (Electron Microscopy Sciences) which had been glow-discharged in air plasma for 15 seconds. Excess solution was removed by automatic blotting (1 blot for 10 s, blot force 10 at 95% relative humidity) before plunge-freezing in liquid ethane using an automated vitrification system (FEI Vitrobot). Imaging was performed with a Titan Krios TEM operated at 300 kV, equipped with a BIO Quantum energy filter. Images were recorded on a Gatan K3 direct electron detecting camera with a pixel size of 0.91 Å/pixel in superresolution mode for cryoE. Imaging was performed under cryogenic conditions using a low electron dose of 1100 $e^{-}/\text{Å}^{2}$. Dose-fractionated movies with a total dose of 3 $e^{-}/\text{Å}^{2}$ were acquired at tilt angles ranging from -60° in 1° increments and defocus values of -75, -200, and -550 nm at each tilt angle in a dose-symmetric scheme starting at 0° and -75 nm defocus using a custom script in SerialEM software.

Pre-processing

After acquisition, the images are pre-processed before sending into the reconstruction framework. There are two prior steps that are required for this particular dataset. First, all images need to be normalized to ensure a uniform background. Second, to cope with the sample drift as well as tilt axis mislignment problem mentioned in chapter 1, we utilize the fiducial markers present in the sample to coarsely align the images in terms of both translation and in-plane rotation. After the second step, the image stack shall have all images aligned and the tilt axis in the center of the field of view. Sample intensity images at different defocus distances are shown in Fig. 4.3, along with zoomed-in regions and their corresponding Fourier transforms.

Image Normalization

The proposed framework in chapter 2 assumes a uniform plane wave illumination onto the sample. However, often times in real experiments the illumination would not have uniform intensity over the entire field of view. Illumination is usually the strongest at the center and relatively weaker at the side or corner of the field of view. To avoid such model mismatch, the illumination intensity needs to be corrected.

For each 2D intensity image, we estimate the illumination intensity over the entire field of view by fitting a 2D binomial-polynomial surface to the image, with the Bézier surface basis [34]. A 2D Bézier surface of degree m and n can be parameterized as:

$$p(x,y)=\sum_{i=0}^{n}\sum_{j=0}^{m}B_i^n(x)B_j^m(y) \tag{4.1}$$

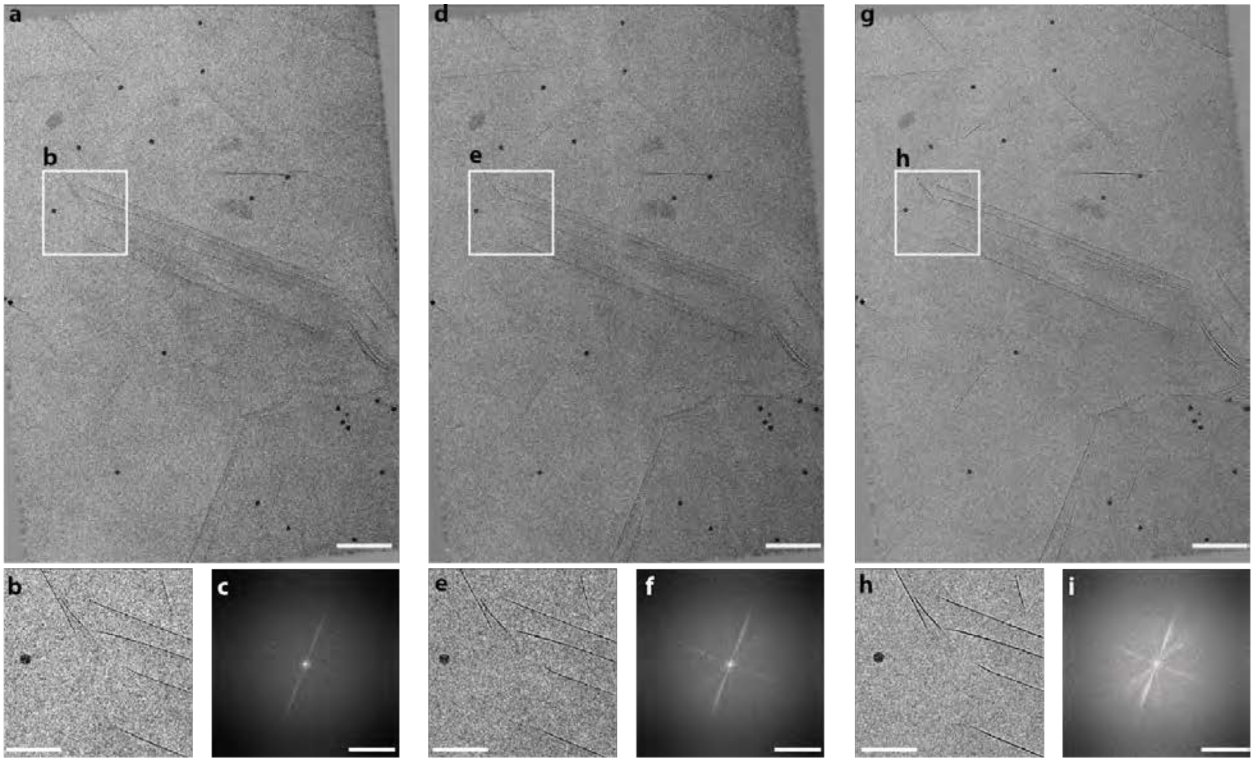

Figure 4.3: Images taken with different defocus values at the same tilt angle as part of the tilt series from which the tomogram in Fig. 4.5 was reconstructed. -75 nm defocus (a-c) the contrast is weakest (a-b) but the information transfer is highest (c). At -550 nm defocus (g-i) the contrast is highest (g-h) but the information transfer is reduced (i). Scale bars are 100 nm in (a, d, g), 50 nm in (b, e, h) and 1.2 $\mathring{A}^{-1}$ (c, f, i).

where $B_i^n(x)$ is the binomial polynomial evaluated over the unit square

$$B_i^n(x) = \binom{n}{i} x^i (1-x)^{n-i}. \tag{4.2}$$

For Bézier surfaces of lower degrees, the height profile $p(x, y)$ is slowly varying and can be used as background estimation to ensure uniform illumination. Since the experimental parameters (tilt angle and defocus distance) vary when each image is taken, we perform the same task for the images independently in parallel. After fitting the surface to the images, the image is divided by the estimated slowly varying background. Notice that this process may amplify the noise in the dimmer lit areas of the field of view, causing poor reconstruction qualities in those areas.

After ensuring that the illumination is uniform across the field of view, we further normalize the image by dividing out its mean, so that the background has a normalized value of 1, which is also assumed by the physical model - a physical plane wave has an intensity 1. Notice that this step assumes no electron loss when traveling through the homogeneous medium.

Image Alignment with Fiducial Markers

10nm Gold nanoparticles are used as fiducial markers during the tomography experiment to align the tilt series. It is easy to show that when a fiducial marker is placed in a volume, the tilt-series of it can be traced by a sum of sinusoidal functions that is parameterized by the tilt angle and the spatial positions of the marker inside the volume.

The algorithm first detects and tracks all markers in the tilt series, requiring certain degree of manual effort. Then, images are translated to ensure that the markers follows the sinusoidal trajectory according to the nominal tilt angles. After that, when all markers are plotted together (using minimum projection), it is clear in Fig. 4.4 that the traces of all fiducial markers are slanted. The slanted traces of markers indicate a misaligned tilt series. The red solid line in Fig. 4.4 is the experimental tilt axis, whereas the dashed line is the nominal tilt axis. Therefore, an in-plane rotation of angle θ is needed such that the experimental tilt axis coincides with the nominal tilt axis that is centered and parallel to the $y-axis$. The registration process requires padding of values 1 (again, intensity of a plane wave) to avoid any boundary artifacts. As such, the boundaries may seem unnatural in Fig. 4.3.

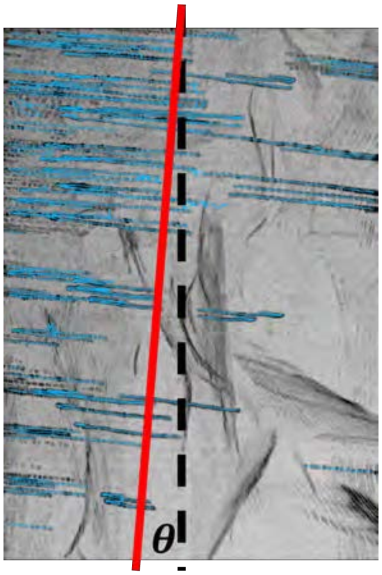

Figure 4.4: All traces of gold nanoparticles as fiducial markers(in blue) overlayed on top of minimum projection of the tilt-series. A in-plane rotation of angle θ is needed to ensure that the tilt axis is parallel to the y-axis.

Notice that some marker traces are shorter than others, and there are a few reason for that. First the markers can go out of the FoV at certain sample tilt agnles, especially those closer to the boundary of the volume. Second, when the markers are locally dense and partial or full occlusion occurs, our tracking algorithm fails to detect the correct marker, as it can be seen on the top left corner of Fig. 4.4. However, since the in-plane rotation angle θ is a global parameter shared among the entire tilt-series, error from tracking can be tolerated and treated as noise as long as sufficiently many markers are found.

For finer registration and alignments, the tilt series are passed into IMOD, an open-source software widely adopted in the community [70].

Reconstructed Volume

When the tilt-series is normalized and registered, it is passed into the multi-slice framework for reconstruction. All volumetric reconstructions are performed on the Lawrence Berkeley National Lab Lawrencium High Performance Computing Clusters. Due to the large size of the volume, the reconstruction process is carried out in a distributed fashion by splitting the tilt-series along the direction of the tilt axis. Splitting the tilt-series along the single tilt axis is intuitive and most accurate, as long as enough of pixels along the tilt axis is cropped and sufficient overlaps are used to ensure the diffraction effects and multiple scattering effects are preserved.

At a mineral volume fraction of 2% and electrolyte concentrations of C = 0.1M and 0.75M the lithium-montmorillonite (Li-Mt) suspensions are composed of mixtures of exfoliated layers and stacked layers separated by electrolyte solution, termed osmotic hydrates (Fig. 4.5a,d). All Mt layers are curved, despite having chemical compositions that are nominally symmetrical about the layer midplane and therefore exhibiting no spontaneous curvature due to, e.g., residual strain. In contrast to the common assumption that osmotic hydrates are planar stacks, we find that they are stacks of curved layers with coaligned curvature axes (Fig. 4.5 b, e). Furthermore, we find that the degree of curvature differs between the two electrolyte concentrations (Fig. 4.5b, e), which we show below is a result of the strong coupling between curvature and the potential drop at the mineral interface, Ψ, due to differing counterion complexation profiles (i.e, charge distribution) on opposing sides.

The large field of view in cryoET enables comparison to structural information accessible by in situ X-ray scattering of equivalent samples over two orders of spatial magnitude. We observe that the average interlayer spacing, $\langle D \rangle$, reflected by a peak in the reciprocal-space structure factor that is commonly used to characterize the structures and interaction forces in layered mineral suspensions [89], differs between electron- and X-ray-based techniques (Fig. 4.5c, f). This indicates that the structure sampled by cryoET, with a volume of 0.29 μm^3, differs from the ensemble structure of a suspension within the approximately 10^8 μm^3 volume sampled by the X-ray beam. Therefore, X-ray profiles are not adequately representative of individual suspension microstates from which interaction forces arise. In other words, the X-ray structure factor cannot be uniquely inverted to produce the irregular distribution of layer spacings that is obviously present in Fig. 4.5a. Only from cryoET can

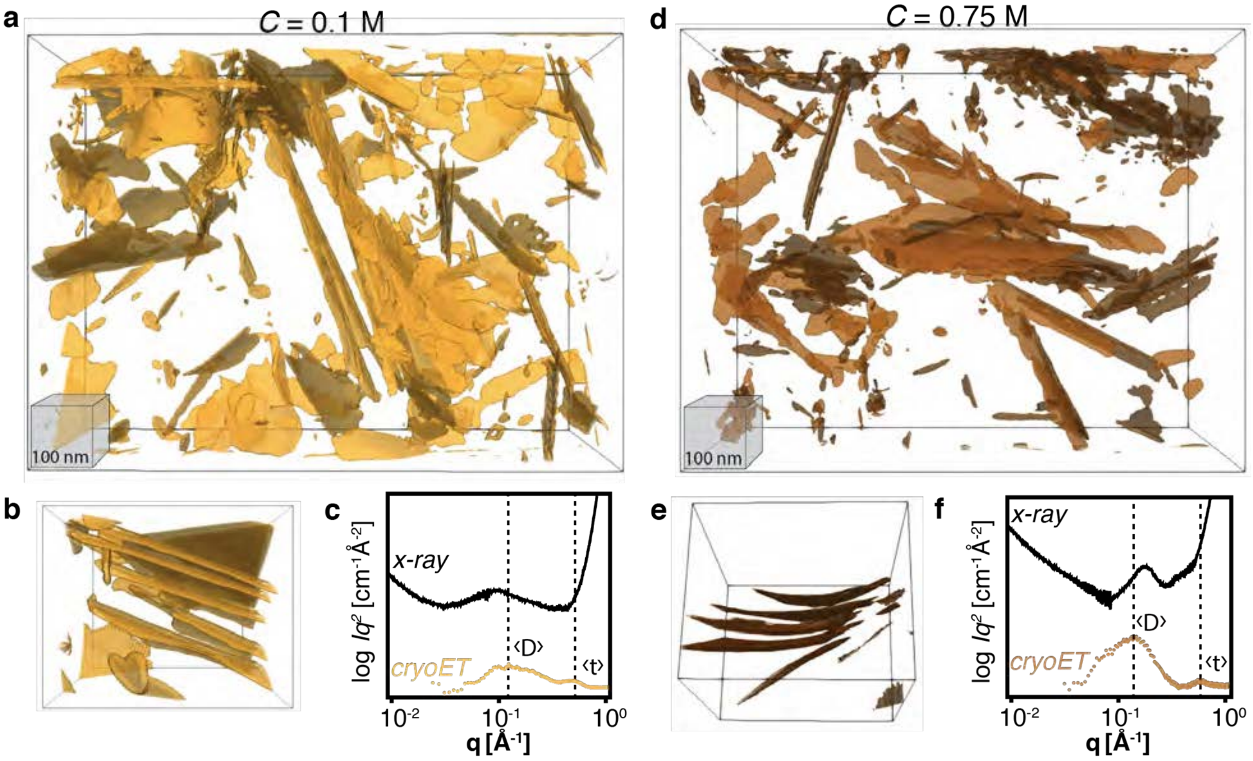

Figure 4.5: Li-Mt suspensions viewed with cryoET. Isosurfaces of Li-Mt layers in 0.1 M (a) and 0.75 M (d) lithium chloride. (b, e) Subvolumes from (a) and (d), showing osmotic hydrate stacks. Comparison of structure factors from X-ray scattering and cryoET in 0.1 M (c) and 0.75 M (f), showing larger average interlayer spacing and larger layer thickness at lower electrolyte concentration.

we determine that positionally ordered regions like that observed in Fig. 4.5b exist in some places but not others, and that not all layers participate in such ordered stacks.

Novel analyses of the 3D cryoET datasets in real and reciprocal space show how lithium complexation varies with electrolyte concentration. While the integrated and cryoET data in Fourier space are directly comparable to X-ray scattering, the difference is that both the phase and amplitude of the cryoET data are known by virtue of recording the data in real space first, allowing us to unambiguously interpret the structure factor peaks. For example, a peak in the cryoET structure factor between scattering vectors q = 0.50-0.57 Å^{-1} (Fig. 4.5c, f) corresponds to an average layer thickness, $\langle t \rangle$, that includes the aluminosilicate layer and hydrated lithium counterions at the interface, but not bulk electrolyte. The thickness of an exfoliated layer is 12.6 +1.7/-1.8 Å in 0.1 M lithium chloride and 11.0 Å +1.0/-1.9 in 0.75 M lithium chloride. Thicker layers at low electrolyte concentration are thus indicative of a greater fraction of fully hydrated lithium ions that reside further from the layer midplane, while conversely, thinner layers at elevated electrolyte concentration are direct evidence of a higher fraction of lithium ions that make inner-sphere complexes with the mineral interface. This is in accordance with the expectation of EDL models, which predict that higher electrolyte concentration drives complexation equilibria towards partially dehydrated inner sphere complexes.

Montmorillonite Clay Analysis

Segmentation

The clay sheets in the reconstructed cryoET absorption volumes were segmented using custom codes written in Matlab. The below variables were defined for a bin 2 reconstruction size. The goal was to reduce the set of voxels to a set of “sheets” which were defined ass a cloud of points representing a 2D sheet embedded in 3D space, each with an associated normal vector representing the sheet surface normal. First, we generated a list of unit vectors $\mathbf{n}_i$ with roughly even angular spacing over a hemispherical surface, representing the possible sheet orientations. Next, for each orientation we generated a 3D kernel $\mathbf{k}_i$ defined by the function

$$\mathbf{k}_i = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{t/2 - |\mathbf{r} \cdot \mathbf{n}_i|}{w}\right), \tag{4.3}$$

where $\mathbf{r}$ is the real space coordinates centered on the origin, $t = 3$ is the estimated sheet thickness, and $w = 1$ is the estimated sheet interfacial width. This kernel was normalized by applying a 3D Gaussian envelope function and subtracting the mean, i.e. the formula

$$\mathbf{k}_i^{norm} = \left[\mathbf{k}_i - \left\langle \mathbf{k}_i \text{exp}\left(-\frac{|\mathbf{r}|^2}{2\sigma^2}\right)\right\rangle\right] \text{exp}\left(-\frac{|\mathbf{r}|^2}{2\sigma^2}\right) \tag{4.4}$$

where $\sigma = 8$ is the Gaussian envelope standard deviation, and $\langle \cdot \rangle$ represents the expected value. For each potential orientation, we use Fast Fourier Transforms to efficiently compute the correlation of the kernel with the reconstructed volume.

Classification

We then take the maximum correlation value in each voxel over all orientations, while also storing the best-match orientation. We then classified the sheet voxels by applying two thresholds: (1) a global threshold by using a minimum value for the correlation signal, and (2) voxels with correlation signals greater than at least 18 neighboring voxels (out of a possible 26 neighbors).

Next, we computed the nearest neighbor network for the set of all sheet voxels. This network was used to segment the set into separate sheets using two matching rules: (1) locally connected voxels, and (2) those with orientations within 30° of each other. At this stage we also discarded sheets which consisted of less than 1000 voxels, since these were either false positives or sheets too small to make accurate measurements of the surface topology.

Surface Profile

The final analysis steps consisted of measurements performed on the segmented surfaces. After individual voxels are grouped together to form larger clay sheets, local change curvature within the clay sheet can be estimated. For each voxel in clay sheet, together with all its neighboring voxels within a certain distance d, the local surface curvature is parameterized by

$$S(x, y) = a_1 + a_2x + a_3y + a_4x^2 + a_5y^2 + a_6xy, \tag{4.5}$$

where S is the surface height (also the z coordinate) given coordinates x and y, and $a_1, ..., a_6$ are coefficients describing the local surface. The coordinates of each voxels (x, y, z) are known a priori, and a linear regression is run to solve for the best coefficients $a_1, ..., a_6$ that describes the local parabolic surface.

Once the parabolic surface is fitted, mean surface curvature $2H$ was then calculated by using the following expression

$$2H = \frac{\left(1 + S_x^2\right) S_{yy} - 2S_xS_yS_{xy} + \left(1 + S_y^2\right) S_xx}{\left(1 + S_x^2 + S_y^2\right)^{3/2}}, \tag{4.6}$$

where S_x and S_y are the first order spatial derivatives of the surface, and S_{xx}, S_{yy}, and S_{xy} are the second order derivatives.

From the parabolic surfaces, surface normal vectors are also defined, and so are line traces crossing the surface in the normal direction.

Discussion

The real-space structure of the EDL and its dependence on curvature were first quantified by extracting and averaging the reconstructed absorbance magnitudes normal to the surface of Mt layers, with representative exfoliated Mt layers with low and intermediate curvature (Fig. 4.6a,b) compared with Mt layers exhibiting higher curvature in an osmotic hydrate (Fig. 4.6c). Statistical analysis of between 5.3×10^4 and 1.3×10^5 absorbance profiles taken normal to the layers at each midplane voxel revealed features below the nominal voxel resolution (Fig. 4.6d-f), in analogy with the common practice of particle- or sub-tomogram averaging in cryoEM of biological macromolecules [21]. The resulting averaged ion-density profiles reflect some expected aspects of EDL models of layer silicates. For example, a region of low absorbance extends approximately 5 nm from the layer midplane (Fig. 4.6d-f), arising from the depletion of chloride ions that are repelled from the negatively charged mineral. Thus, absorbance profiles starting from the layer midplane and moving into the bulk solution can be attributed to dominant contributions from mineral, lithium, and chlorine, respectively. In contrast to existing models, however, real-space ion-density profiles reveal a prominent role for layer curvature, H, in modulating interfacial ion distributions. Compared to the low-curvature layer (Fig. 4.6d), both high-curvature layers show asymmetry in the

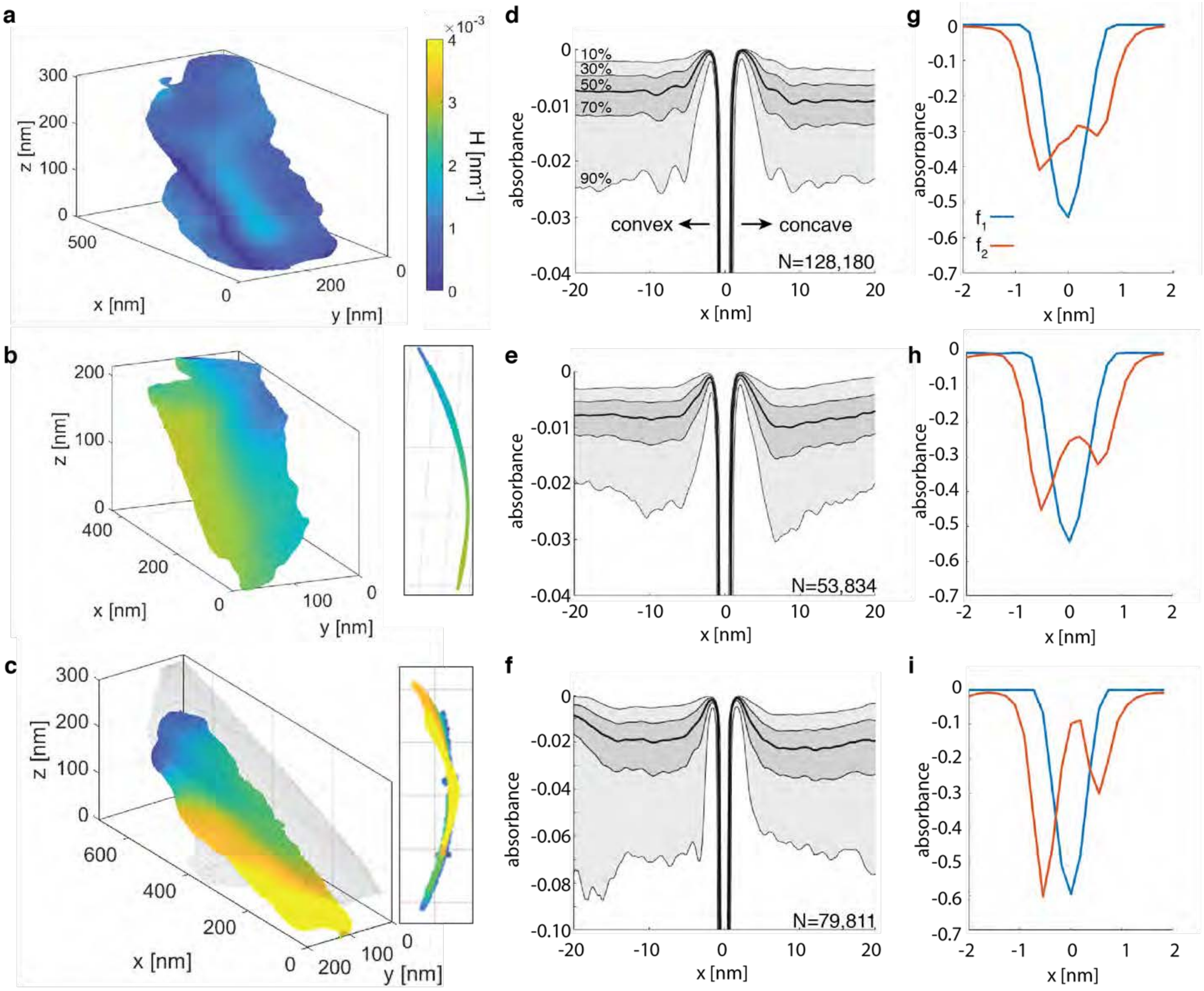

Figure 4.6: Effect of layer curvature on montmorilonite-electrolyte ion-density profiles in 0.1 M lithium chloride. (a) Density reconstruction of a single exfoliated layer with low curvature quantified by H, the reciprocal of the local radius of curvature. (b) Single layer with intermediate curvature. (c) Highly curved layer in an osmotic hydrate stack. A single layer is colored according to local curvature and the neighboring layers shown in transparent gray. (d) Symmetric average ion-density profiles from within 20 nm of convex and concave sides of the low-curvature layer. Shading indicates the variance from 128,180 individual profiles. (e) Asymmetric average ion-density profile from the intermediate-curvature layer. (f) Asymmetric, and higher magnitude, average ion-density profiles between stacked layers (note difference in scale). (g) Non-negative matrix factorization (NNMF) of all absorbance profiles in (d), showing first two factors, f1 (Mt) and f2 (Li). (h) Increasingly asymmetric f2 in NNMF profile of curved layer in (b). (i) Highly asymmetric NNMF f2 for curved stacked layer.

anion depletion region directly adjacent to the mineral surface and in the anion distributions at distances up to 15 nm from the mineral (Fig. 4.6e,f). Thus, concavity sequesters counterions over appreciable distances near exfoliated layers.

Often, profiles do not converge at large distances because of the presence of neighboring layers. Such a case is presented in Fig. 4.6c, in which the neighboring layers are shown in gray. The interlayer distance is not uniform between neighboring layers, and profiles taken normal to the midplane of a layer (over which the degree of curvature, and thus the orientation of the normal vector, vary considerably) extend over a large range of distances before encountering a neighboring layer (and the interfacial ion distribution associated with that layer). The use of normal vectors to extract ion profiles in Fig. 4.6d-f leads to reproducible trends in ion distributions, despite capturing broad distributions of interfacial curvature and interlayer distances. We also applied non-negative matrix factorization (NNMF) to the collection of normal profiles in order to separate these many convoluted contributions into quantifiable trends in the ion distributions, shown in Fig. 4.6g-i.

We observe that outer-sphere lithium complexation is asymmetrically distributed, with greater concentrations on the convex side relative to the concave side. Due to the strong absorbance from the layers, contributions from inner-sphere Li complexes could not be directly quantified. However, asymmetric outer-sphere complexation increases with increasing curvature (Fig. 4.6g-i), a clear demonstration that both inner- and outer-sphere complexation states coexist and that their relative proportions are dependent on layer curvature.

Summary

In this chapter, not only have we validated our proposed framework to achieve high resolution 3D reconstruction of multiply scattering samples, but also have new findings to demonstrate that complexation waves occur over a wide range of conditions in layered mineral systems that arise from the exchange of elastic, electrostatic and hydration energy as ions partition from the bulk electrolyte, complex with the mineral layer and induce it to bend. A new interaction force, carried by complexation waves, emerges through the delocalizing effect of curvature, which spans length scales ranging from hundreds of nanometers to angstroms, and strongly couples the temporal response of layered mineral systems to chemical or mechanical perturbations. This work has opened a new window into similar analysis of structure of other materials using cryoET.

Chapter 5

Efficient Computation for Electron Tomography

With the recent development of high-resolution detectors and algorithms that can account for multiple-scattering events,thicker samples can be examined at finer resolution, resulting in larger reconstruction volumes than previously possible. In this chapter, we introduce a series of optimizations toward efficient 3D reconstruction.

5.1 Introduction

In a brightfield Transmission Electron Microscopy (TEM) system, a plane wave illuminates the sample and phase delays are induced by the sample's electric potential. Phase contrast can be obtained by slightly defocusing the image of the sample, and the image contrast will be linear with respect to the cumulative phase as long as the sample is weakly-scattering. Many biological samples and thin foils meet this criteria; as such, classical tomography methods that rely on the projection slice theorem[41, 86, 102] can be directly applied to solve for 3D structures. However, to ensure the validity of the weakly-scattering assumption, the sample should be thin, and this poses a great challenge to sample preparation. For thicker samples or materials with larger Z number, nonlinear scattering effects become non-negligible.

In addition, many samples cannot tolerate high electron dose – these samples are often vitrified in medium at cryogenic temperature to avoid sample damage from the electron beam, known as Cryo-EM for 2D imaging or Cryo-ET for 3D imaging [12]. Electron beam damage is a complex process that is not fully understood, but is roughly inversely correlated with atomic number and occurs more rapidly at surfaces and defects. Therefore, most matter on Earth's surface is generally beam sensitive in an electron microscope, being composed of light elements that are frequently hydrated and imaged with cryoEM/cryoET [53, 137]. Being able to model the multiply-scattering events between the sample and the beam probe and more efficiently 'use' the electrons counted in the images can reduce the amount of dose needed to achieve the same reconstructed SNR [106].

To image thicker and more diverse samples, multiple scattering and the non-linearity in the image formation process should be taken into account [131, 118, 63, 106, 98, 78]. This can be done by incorporating such phenomena into the forward model, for example by implementing a multi-slice (beam propagation) model [131, 63, 106], which represents the 3D scattering process using a sequence of 2D layers of transmittance functions that cause beam absorption and phase delay. The input plane wave is propagated through the layers, assuming a fixed distance of free space in between. This method is intuitive, robust, and efficient to implement. Multi-slice methods are powerful, but more computationally expensive than traditional projection-based methods, and often need to be run dozens or hundreds of times inside an iterative reconstruction loop.

Computational requirements are further increased by recent advances in direct electron detector technology which have greatly improved imaging throughput, both in frame rate and pixel count. Many detectors have demonstrated capacity to capture images with up to $8k \times 8k$ pixels [47, 87, 120, 82]. In our study, for example, a Gatan K3 detector is used, and the images captured have 5760×4092 pixels. Given that 3D techniques generally capture dozens or hundreds of images in a tilt-series, the result is that very large volumes (in terms of voxels) can be reconstructed.

With these hardware improvements, more computational resources are required to process the increasing amount of tomography data. The scale of the datasets usually exceeds the capacity of a modern single-node computer; hence, parallelization is needed to decompose the reconstruction into multiple parallel sub-problems. Previous work in linear tomography has extensively used parallel computing [36, 43, 7, 8, 51, 52, 135], but requires linear projections or tomographic matrix sparsity. Since multiple scattering is a nonlinear interaction between the wave and the sample, a matrix cannot be written to represent the scattering process. Consequently, these methods are not compatible with existing distributed computing strategies.

Particular in the following sections, we first describe a slice-binning mechanism we use to reduce the number of axial slices needed per scattering calculation while maintaining accuracy. Secondly, a GPU-enabled python tomography solver is presented. With GPU acceleration, the reconstruction speed is significantly improved. Third, a distributed computing framework is presented that reconstructs large volumes by decomposing a projected tilt-series into smaller datasets such that sub-volumes can be simultaneously reconstructed on separate compute nodes using a cluster. We demonstrate our method by reconstructing a multiple-scattering montmorillonite sample at high resolution from a large field-of-view tilt-series dataset.

5.2 Slice-binning

In both the forward and back propagation of the multislice scattering model, the major bottleneck in computation is the Fourier transform. The number of Fourier transform performed is proportional to the number of slices in z. Since complete tomography without missing angles achieves isotropic resolution, the number of slices in z should match the number of pixels reconstructed in x and y, so the number of slices along the beam direction should be equally as dense, causing very heavy computation.

In this section, we describe our application of the slice-binning in the tomography algorithm [114]. With slice-binning, at every tilt angle we increase the thickness of each slice (i.e. reducing axial resolution per angle). As a result, while total thickness of the sample remains constant, the total number of slices is reduced, along with the computation time. However, because tomography allows us to capture information about each voxel from multiple angles, the redundant information from the other tilt angles allows us to still reconstruct the object at the original resolution isotropically.

In particular, we sum the 2D projected potentials of N_B consecutive layers at each angle:

$$\mathcal{B}_{N_B}\{V\} = \left\{\sum_{m=1}^{M} V_{nN_B+m}(\mathbf{r})\right\}_{n=0}^{\lceil N_z/N_B\rceil-1}, \tag{5.1}$$

where $\lceil\cdot\rceil$ is the ceiling function, and

$$M = \begin{cases} N_B & ,\text{if } n < \lceil N_z/N_B\rceil - 1 \\ N_z - nN_B & ,\text{if } n = \lceil N_z/N_B\rceil - 1 \end{cases}. \tag{5.2}$$

We then compute both the forward model and back propagation using this binned potential. After the gradient is calculated, we distribute the gradient to the full volume by applying the adjoint operator, $\mathcal{B}^\dagger$:

$$\mathcal{B}^\dagger_{N_B}\{V_B\} = \left\{V_{B,\lceil\frac{m}{N_B}\rceil}(\mathbf{r})\right\}_{m=1}^{N_z}. \tag{5.3}$$

In the simulations shown in the results section, we bin every 10 slices. Since the pixel size in z is 0.5 Å, the effective slice separation becomes 5 Å, which is sufficient to recover atomic resolution in the 2D parallel directions. This combined with many tilt angles will produce atomic resolution in 3D with pixel size of $(0.5\text{Å})^3$.

However, the reconstruction quality deteriorates as we gradually increase the number of slices being binned N_B. Therefore, the extent to which we can bin the slices is of special interest. The precise mathematical error analysis is not available due to the non-linearity of the multislice method, and so to estimate an upper bound for slice-burring we use the 3D CTF of the imaging system by assuming single or weakly scattering [124]. Then, we are able to linearize the problem to obtain an estimate of the error. In a traditional imaging system with numerical aperture $\text{NA} = \lambda/\Delta x$, where Δx is the pixel size, the axial resolution can be characterized as:

$$\Delta z = \lambda/(1-\sqrt{1-\text{NA}^2}). \tag{5.4}$$

Based on Nyquist sampling criterion, the maximum thickness for every slice should be less than Δz to support the axial resolution at every angle.

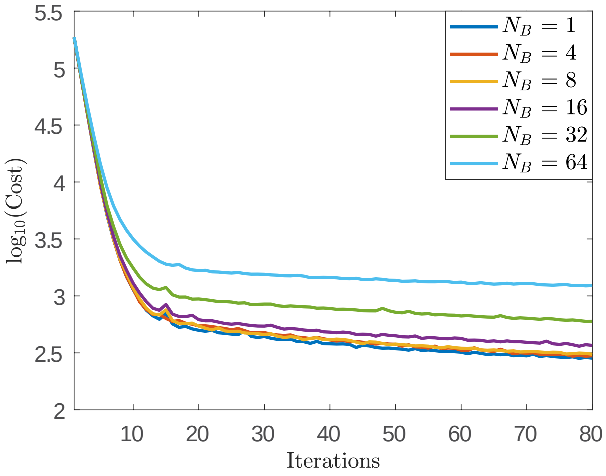

Figure 5.1: Plot of cost function vs iterations to show convergence for various binning factors (N_B).

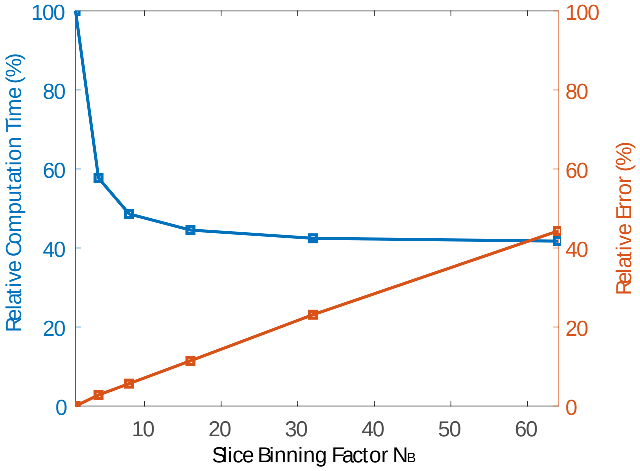

Figure 5.2: Plot of relative time savings (left y-axis) and relative error of reconstruction (right y-axis) vs slice binning factors (N_B).

We test the effectiveness and fundamental limit of the proposed slice-binning method. Here, we exponentially increase N_B to examine the effect it has on reconstruction error, computation time, and convergence behavior of the algorithm.

To simplify our discussion, all synthetic datasets in the validation process are generated from 60 uniformly separated tilt angles with 3 defocus planes, assuming infinite dose, and the phantom is the same as that described in 3. We do not apply any regularization methods as they alter the convergence behavior depending on the choice of the regularization parameter.

5.3 Implementation of Multi-slice Algorithm

The underlying reconstruction framework is powered by a Phase Contrast Tomography Solver library that is custom written in Python. This solver repository allows one to input the tilt-series and a list of configuration parameters and outputs the reconstructed volume. The details could be found online at GitHub.

The solver is written in a modularized fashion, into many different classes, as illustrated Fig. 5.3. First, an object of the highest level class **TorchTomographySolver** is constructed by passing in all relevant measurement parameters such as tilt-series, tilt angles, voxel sizes, as well as reconstruction parameters such as number of iterations, step size. After passing in the data and configuration parameters, the constructor of **TorchTomographySolver** class will subsequently in a hierarchical fashion create lower level objects of the following classes:

Class name	**Description**
ImageRotation	Rotates a 3D sample to an arbitrary degree.
PhaseContrastScattering	Models 2D image formation process from 3D sample.
ImageTransformOpticalFlow	Registers measurement images.
Regularizer	Enforces prior information about the sample.
AETDataset	Manages all measurements and corresponding parameters.

The configuration parameters will be parsed and distributed to the constructors of individual classes accordingly to initiate the objects.

The object of **TorchTomographySolver** class has a method called **run()** to start the reconstruction process when called. A flag *forward_only*(Default false) is used to switch between a single forward pass or full reconstruction. If the *forward_only* is true, then a single forward scattering operation is done to form the tilt-series given the input 3D sample. On the other hand, if the flag is false, a full reconstruction will be done. The reconstruction is conducted iteratively. In each iteration, the predicted measurements from a 3D estimated sample are formed by rotating the sample and propagating the electron wave through it using multi-slice. Then, the predicted tilt-series is registered with the experimental tilt-series to correct for translation and in-plane rotation error. The registration step is optional and is built based on an open source library named pyStackReg [123]. After that, a gradient step with respect to the error between the predicted and actual tilt-series is taken to update the

3D sample. The gradient is automatically calculated by Pytorch Autograd functionality and thus is transparent to us. Regularization is enforced after all tilt-angles have been visited.

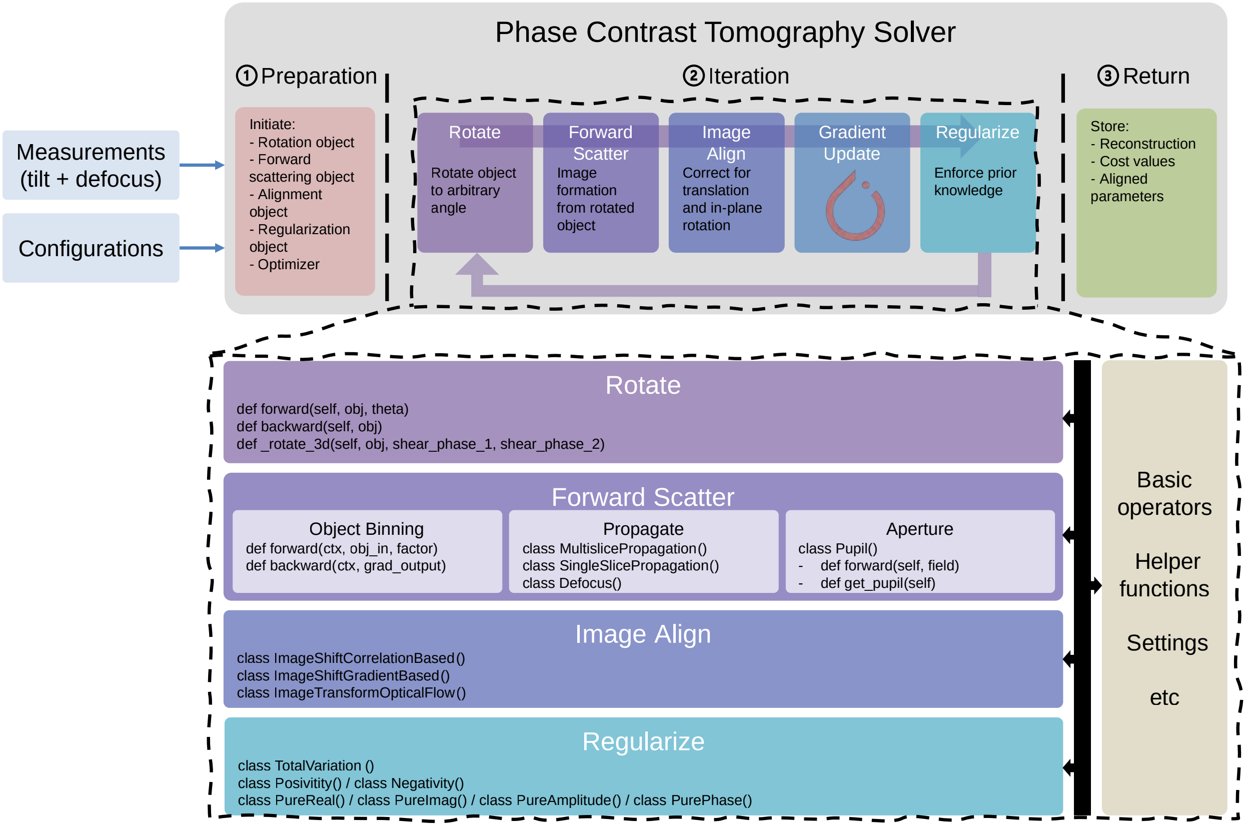
Phase Contrast Tomography Solver
Measurements (tilt + defocus)
Configurations
① Preparation
② Iteration
③ Return
Initiate:
- Rotation object
- Forward scattering object
- Alignment object
- Regularization object
- Optimizer
Rotate
Rotate object to arbitrary angle
Forward Scatter
Image formation from rotated object
Image Align
Correct for translation and in-plane rotation
Gradient Update
Regularize
Enforce prior knowledge
Store:
- Reconstruction
- Cost values
- Aligned parameters
Rotate
def forward(self, obj, theta)
def backward(self, obj)
def _rotate_3d(self, obj, shear_phase_1, shear_phase_2)
Forward Scatter
Object Binning
def forward(ctx, obj_in, factor)
def backward(ctx, grad_output)
Propagate
class MultislicePropagation()
class SingleSlicePropagation()
class Defocus()
Aperture
class Pupil()
- def forward(self, field)
- def get_pupil(self)
Image Align
class ImageShiftCorrelationBased()
class ImageShiftGradientBased()
class ImageTransformOpticalFlow()
Regularize
class TotalVariation ()
class Posivitiy() / class Negativity()
class PureReal() / class PureImag() / class PureAmplitude() / class PurePhase()
Basic operators
Helper functions
Settings
etc

Figure 5.3: Brief software architecture for the proposed framework from high level to lower level.

To ensure fast recovery of the 3D sample, GPU (Graphics processing unit) -accelerated implementation of the multi-slice algorithm is needed. Different from CPU, GPU helps parallelizing standard operations such as element-wise multiplication and Fourier transforms, thus vastly reducing the computation time. In our work, we used a python package named Pytorch [97] that has automatic low-level parallelization for complex numbers. Therefore, without changing much of the python code, one can move the computation entirely from CPU to GPU.

Notice, however, that the capacity of GPU memory is limited comparing to CPU RAM. Therefore, one needs to be cautious and efficient about the content saved in the GPU memory. Frequent transfer of arrays between CPU and GPU memory can become a bottleneck of the computation efficiency. In our work, the 3D sample is always saved on the GPU memory once the reconstruction begins. In the meantime, the tilt-series is saved on the CPU as not all of them are used at the a time. When a particular tilt-angle is visited, the corresponding measurement will be transferred from CPU memory to GPU memory. Total variation regularization mentioned in chapter 2 requires large amount of memory (3 times or more, depending on the implementation) in order to save intermediate variables. As such, largest size of 3D sample is limited. An alternative solution would be to use other classes of regularizers, or simply run the total variation regularization on the CPU, naturally coming with an impact on computation efficiency.

5.4 Distributed Compute for Electron Tomography

In this section, we propose a distributed algorithm that preserves the coherent diffraction effects in the tilt-series, and thus works for multiple-scattering forward models. We first show that by carefully cropping out particular regions of each raw image in the original tilt-series, any given sub-volume within the full volume can be reconstructed, though artifacts may occur due to contributions from outside of the sub-volume. Thus, we can define many sub-volumes within the full volume and reconstruct them all in parallel on separate compute nodes, while still accounting for multiple-scattering effects. The sub-volumes are then stitched together to form the full-volume reconstruction. We demonstrate the performance of the algorithm experimentally by solving small tomography problems while varying different parameters. Both reconstruction time and mean square error (MSE) are reported. Finally, we demonstrate the tomographic reconstruction of a large volume consisting of clay minerals vitrified in aqueous solution. To our knowledge, this is the largest volume (in number of voxels) ever reconstructed in Cryo-ET, with a size of $(0.73(x) \times 1.00(y) \times 1.73(z)\mu\mathrm{m}^3)$ and resolution of 1.82 $\mathring{\mathrm{A}}^3$/voxel. At this resolution, not only do we see unprecedented microscopic features, but also we can visualize and understand macroscopic sample structure immersed in the solution. In all, our method offers the following advantages:

- It is model independent as we only manipulate the tilt-series before the reconstruction. The choice of tomography algorithm is thus decoupled, ensuring compatibility with any multiple-scattering tomography model.

- It requires no inter-node communication – during the reconstruction, all parallel compute nodes are completely independent from each other. Therefore, the reconstruction speed of one node does not impact the speed of others, and results in less overall process idle time.

- It does not require one to possess deep knowledge of computer architecture, because implementing it does not need low-level architecture-specific optimization such as exploiting the matrix sparsity and other structural properties of the linear inverse problem.

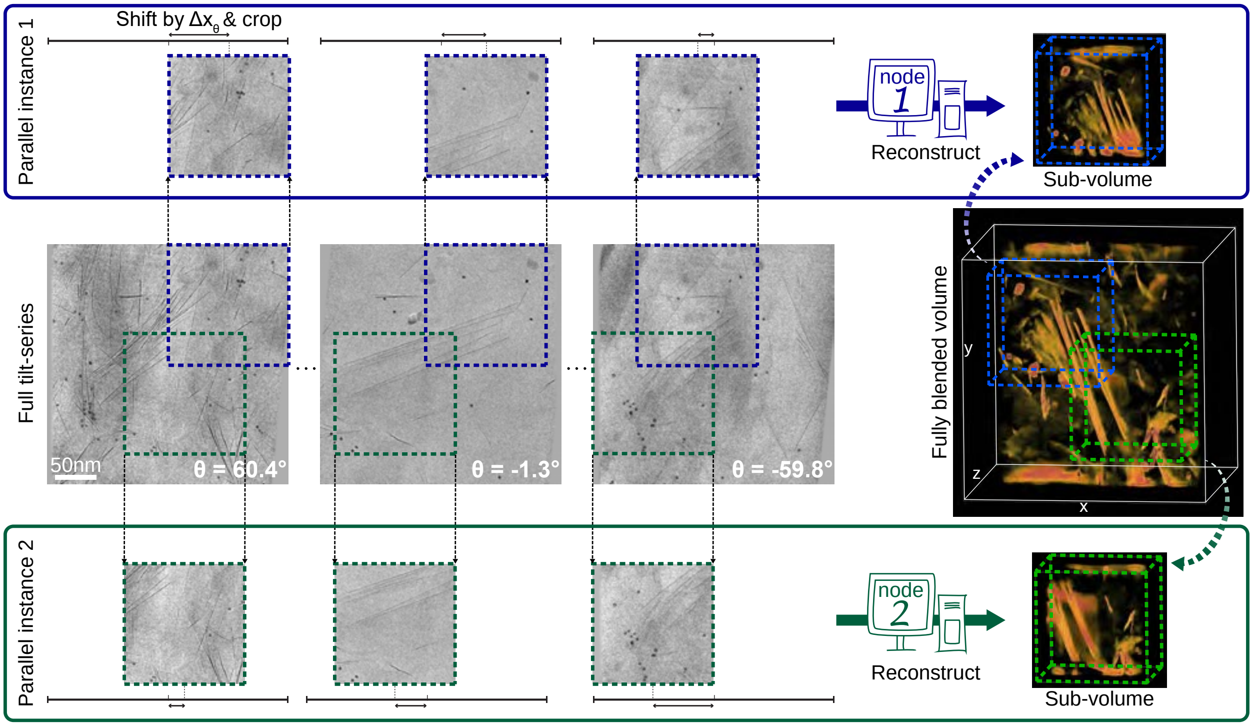

Figure 5.4: Conceptual illustration of our distributed reconstruction algorithm. Two examples of parallel nodes are shown in this figure, in blue and green, respectively. Each correspond to a different sub-volume of the sample, the tilt-series of which are shifted by $\Delta x_\theta = x_0\cos(\theta) + z_0\sin(\theta)$ and cropped from the full tilt-series dataset depending on the position of the sub-volume with respect to the full volume. After all sub-volumes are reconstructed, they are stitched together to form the large-volume reconstruction.

Methods

Reconstructing a sub-volume

In this section, we will show that a subset of the full volume can be reconstructed independently by shifting and cropping each raw image in the tilt-series appropriately. Each sub-volume within the full 3D sample volume can then be reconstructed in a distributed fashion, as shown in Fig. 5.4. To calculate which parts of each raw image to crop for a given sub-volume, we use the mathematics of linear projection, also known as the Radon transform [103, 104]. We first look at the 2D projection image formulation for single-axis tilt along the y-axis. The 2D projected image $I(x, y)$ at tilt angle θ is related to the volume-of-interest through a Radon transform:

$$\begin{aligned} I(x, y; \theta) = \iint & f(x', y, z') \times \\ & \delta(x'\mathrm{cos}(\theta) + z'\mathrm{sin}(\theta) - x)\, \mathrm{d}x'\mathrm{d}z', \end{aligned} \tag{5.5}$$

where $f(x, y, z)$ is the 3D sample and $\delta(\cdot)$ denotes Dirac delta function. A full projection dataset is formed by rotating the 3D sample to different angles θ. The tilt axis is assumed to be at the *center* of the volume. However, when a sub-volume of the 3D sample is considered, the center of it does not necessarily coincide with the center of the full 3D volume, and further manipulation is required to relate the tilt-series of the sub-volume to that of the full volume.

Consider the tilt axis of a sub-volume. It is parallel with the true tilt axis of the full volume during the experiment, and translated along x and z by x_0 and z_0, respectively. This is equivalent to the sample being shifted in opposite directions. The new projected images $I'(x, y; \theta)$ are then:

$$\begin{aligned} I'(x, y; \theta) = \iint & f(x' - x_0, y, z' - z_0) \times \\ & \delta(x'\mathrm{cos}(\theta) + z'\mathrm{sin}(\theta) - x)\, \mathrm{d}x'\mathrm{d}z'. \end{aligned} \tag{5.6}$$

After performing a change of variables ($x'' = x' - x_0$ and $z'' = z' - z_0$) and simplification, the relationship becomes

$$\begin{aligned} I'(x, y; \theta) = \iint & f(x'', y, z'') \times \\ & \delta[x''\mathrm{cos}(\theta) + z''\mathrm{sin}(\theta) - \\ & (x - x_0\mathrm{cos}(\theta) - z_0\mathrm{sin}(\theta))]\, \mathrm{d}x''\mathrm{d}z''. \end{aligned} \tag{5.7}$$

The new projections can now be related to the original projection as:

$$I'(x, y; \theta) = I(x - x_0\mathrm{cos}(\theta) - z_0\mathrm{sin}(\theta), y; \theta). \tag{5.8}$$

Given a tilt angle θ, the new projection corresponding to the tilt axis of the sub-volume is simply a shift along the x-axis with an amount of $\Delta x_\theta = x_0\mathrm{cos}(\theta) + z_0\mathrm{sin}(\theta)$. With the new tilt-series $I'(x, y; \theta)$, the set can be cropped while ensuring that the subset is centered around the tilt axis of the sub-volume. Because the result suggests a global shift of the projection along the x-axis for each tilt angle θ, it preserves the multiple scattering and diffraction effects in the original tilt-series. As such, new sets of tilt-series corresponding to different sub-volumes can be calculated independently from the original tilt-series. After that, each new set of tilt-series can be used to reconstruct the sub-volumes simultaneously on different compute nodes, in order to improve compute speed.

Sub-volume Overlap

When splitting a full volume into smaller sub-volumes, overlap between adjacent sub-volumes is necessary both to avoid empty areas in the volume, as well as to avoid artifacts due to diffraction and multiple scattering, where light scatters outside the edges of the sub-volume [22].

We can derive the minimum amount of overlap in order to avoid empty areas and ensure coverage of the full 3D volume. From a tilt series of size $N_x \times N_y$ pixels, the size of the reconstructed volume is $N_x \times N_y \times N_z$ voxels, where each is the number of pixels along each of the three axes, respectively. Since the sample is rotating with respect to the y-axis, the support of the reconstructed sample within the volume is an inscribed elliptic cylinder with semimajor axis $N_x/2$, semiminor axis $N_z/2$ and height N_y. If $N_x = N_z$, the support becomes a cylinder with radius $N_x/2$. When we concatenate the sub-volume cylinders together to form a 3D volume, there will still be areas in the volume that remain empty (i.e. the four corners) unless there is sufficient overlap between the cylinders.

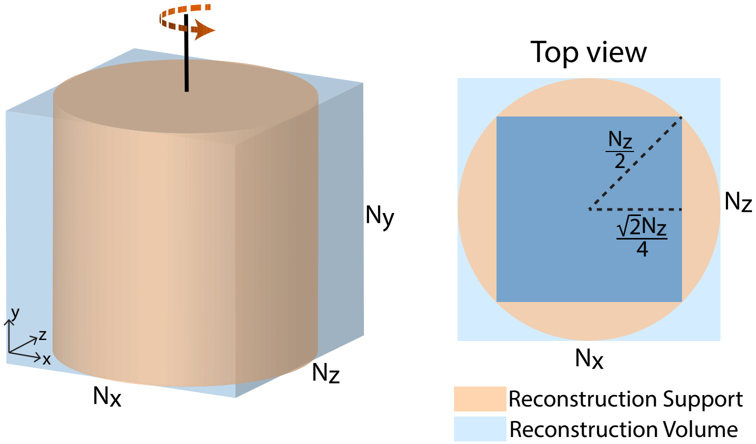

Figure 5.5: Illustration for calculating the minimum overlap between sub-volumes that covers the entire volume, assuming $N_x = N_z$. The reconstructed volume has cylindrical support and the inscribed blue square with side length of $\sqrt{2}/2N_z$ is used for final volume stitching. As such, the minimum overlap required is approximately 30%.

In our method, we overlap the adjacent sub-volumes, and only consider the cuboid inscribed within the cylinder (Fig. 5.5) for each sub-volume reconstruction. By geometry, the inscribed cuboid has length and width of $\sqrt{2}/2N_x$ and $\sqrt{2}/2N_z$, respectively. This corresponds to a minimum required overlap of 30% along both x-axis and z-axis. In Section 5.4, we show the effects of varying the overlap parameters. If overlap is not sufficient, major artifacts start to appear in the reconstructions.

Given the size of each sub-volume and the amount of overlap desired, the number of parallel nodes (M) needed to cover the full volume is $M = M_x M_y M_z$, where

$$M_x = \left\lceil \frac{1 - r_{vx} r_o}{r_{vx} - r_{vx} r_o} \right\rceil . \tag{5.9}$$

$0 \leq r_{vx} \leq 1$ is the ratio of size of reconstructed volume to that of full volume in x, and $0 \leq r_o \leq 1$ is the overlap ratio between consecutive sub-volumes. M_y and M_z can be derived similarly. Notice that this is also the largest number of parallel nodes needed on the cluster to reconstruct the full volume.

Another benefit that follows from being able to solve for sub-volumes is that one can easily reconstruct a custom volume-of-interest in order to quickly zoom in on particular features in 3D, or to avoid reconstructing areas that the sample does not occupy without requiring unnecessary computation. For instance, flat samples that have disproportionate ratio of length and width to depth can have reconstructions with larger M_x, M_y and smaller M_z.

Distributed algorithm

We now describe the overall algorithm for our distributed reconstruction method and show the pseudo code in algorithm 4. With sub-volumes having size $\{r_{vx}N_x, r_{vy}N_y, r_{vz}N_z\}$ and an overlap ratio r_o, we first calculate the centers of each sub-volume. As shown previously, along the x-axis, there are M_x nodes, so for each node $m (1 \leq m \leq M_x)$, the center x_m is:

$$x_m = \frac{N_x}{2} \left(1 + (r_{vx} - r_o r_{vx})(2m - M_x - 1)\right). \tag{5.10}$$

The centers are defined to be uniformly spaced. Similarly, volume centers along the y-axis and z-axis can be calculated as y_l and z_n, respectively, where $1 \leq l \leq M_y$ and $1 \leq n \leq M_z$.

Next, we calculate the sub-volume center points' deviation from the center of the full volume, and apply Eq. (5.8) to obtain the cropped tilt-series corresponding to each sub-volume from the original tilt-series (see Fig. 5.4). Each sub-volume may then be reconstructed independently of all other sub-volumes with any choice of tomographic reconstruction algorithm. As mentioned previously, we use the multi-slice algorithm [106] for multiple scattering. After all sub-volumes are reconstructed, we stitch them together into the full volume reconstruction using a volume blending method described in [20, 22].

Results

To test our distributed algorithm, we used the experimental data described in Whittaker et al. [137]. The sample is clay minerals suspended in a vitrified solution of electrolyte. The sample is imaged using a Titan Krios TEM operated at 300 KeV. A Gatan K3 direct electron detector is used, which has an effective pixel size of 0.91 Å under superresolution mode. The images are then binned down by a factor of 2 after registration. After binning and removing

Algorithm 4 Distributed algorithm

```
Input: Tilt angles $\{\theta_k\}_{k=1}^{N_\theta}$, measured intensity images $\{I_{k,j}\}_{k=1}^{N_\theta}$, sub-volume size $\{r_{vx}N_x, r_{vy}N_y, r_{vz}N_z\}$, overlap ratio $r_o$, reconstruction algorithm $\text{alg}_{\text{recon}}$, and stitching algorithm $\text{alg}_{\text{stitch}}$.
$\left(\{x_m\}_{m=1}^{M_x}, \{y_l\}_{l=1}^{M_y}, \{z_n\}_{n=1}^{M_z}\right) \leftarrow$ Eq. (5.10)
                                                    ▷ Pre-calculate centers
for each sub-volume center $i$ do
    $x_0 \leftarrow N_x/2 - x_m$
    $z_0 \leftarrow N_z/2 - z_m$
    for $k \leftarrow 1$ to $N_\theta$ do                     ▷ Shift and crop projections
        $\Delta x \leftarrow x_0 \cos\theta_k + z_0 \sin\theta_k$
        $\Delta y \leftarrow y_m$
        $I'_k \leftarrow \text{shift}\left(I_k, [\Delta x, \Delta y]\right)$
        $I'_k \leftarrow \text{crop\_center}\left(I'_k, \text{size} = [r_{vx}N_x, r_{vy}N_y]\right)$
    end for
    $f_i^* \leftarrow \text{alg}_{\text{recon}}\left(\{I'_k\}_{k=1}^{N_\theta}\right)$
end for
$f^* \leftarrow \text{alg}_{\text{stitch}}\left(\{f_i^*\}_{i=1}^{M}\right)$
Return: Estimated full volume $f^*$.
```

the boundaries, each image has 5760×4092 pixels. The full tilt series has 121 tilt angles and 3 defocus images per tilt, with a total applied dosage of 1100 $e^-/\text{Å}$. The large field-of-view 3D reconstruction of the sample allows not only a direct comparison with the previously determined average static structure factor by *in-situ* X-ray scattering, but also more insight into the dynamic mesoscale properties of the sample [137].

To reconstruct each sub-volume, we used the algorithm outlined in [106] with a multi-slice forward model. Thus, it includes nonlinear modelling of the 3D sample that captures the multiple-scattering events between the probe and the sample. We then solve for the sample's 3D electric potential by formulating a nonconvex optimization algorithm. The multi-slice reconstruction algorithm is implemented using PyTorch [97], including GPU acceleration. The distributed algorithm, on the other hand, is implemented in Python using only CPU.

The complexity of the multi-slice algorithm [106] on a single volume is $\mathcal{O}\left(N \log N\right)$, where $N = N_x N_y N_z$ is the number of voxels in the reconstruction. By following the calculations in Section 5.4, the total computation complexity across all compute nodes is therefore $\mathcal{O}\left(M(r_v N) \log\left(r_v N\right)\right)$, with sub-volume ratio $r_v = r_{vx} r_{vy} r_{vz}$ (ratio of the size of a sub-volume to that of the full volume) and M nodes required to cover the full volume. Figure 5.7 illustrates the complexity of the proposed algorithm in comparison with the full tomographic reconstruction. For all overlap ratios (r_o) and sub-volume ratios (r_v) tested, our algorithm has higher complexity than if a single multi-slice reconstruction is run on the entire volume. However, since all M tomographic reconstructions can be carried out simultaneously on dif-

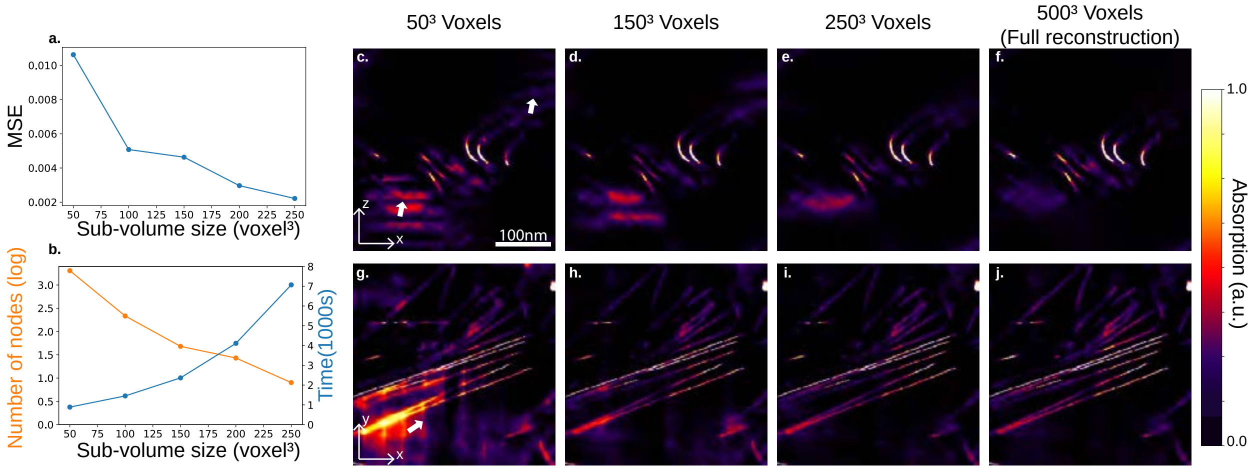

Figure 5.6: Splitting the volume into more sub-volumes, each with smaller size, enables faster reconstructions, but also induces more artifacts. (a) Extra reconstruction error (MSE with respect to a full reconstruction) vs. sub-volume size. (b) Number of nodes needed and reconstruction time for each node vs. sub-volume size. (c)-(f) Example $z - x$ slice and (g)-(j) $y - x$ slice of reconstructed 3D volume with various sub-volume sizes. As the sub-volume size decreases, artifacts increase, as indicated by the white arrows.

ferent nodes of the compute cluster, the total run time for our algorithm is notably faster than for the full reconstruction.

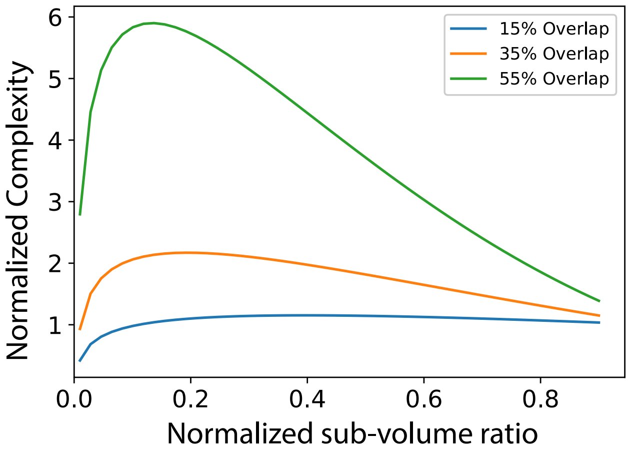

Figure 5.7: Computation complexity as a function of ratio of reconstructed sub-volume size to full volume size and sub-volume overlap ratio. The complexity is normalized with respect to that of a single reconstruction.

All parallel reconstructions were performed using the Lawrencium high-performance computing facility at Lawrence Berkeley National Lab. The cluster contains nodes with one Intel 8-core Xeon Silver-4112 CPU and two NVIDIA V100 GPUs per node. The raw intensity measurements were first loaded into the RAM of a single node that then split it into multiple parallel job instances. Then, the jobs were distributed such that each individual node reconstructed one sub-volume. Finally, all results were aggregated on a single node to form the full volume using a linear stitching algorithm [20, 22].

We tested the performance of the proposed algorithm by varying sub-volume size (Fig. 5.6) as well as overlap ratio (Fig. 5.8) and calculating the Mean Square Error (MSE) with respect to a single full tomographic reconstruction (f^*), which we treated as the ground truth in the error calculation:

$$MSE(f) = \frac{1}{N} \sum_{x,y,z} \left| f(x,y,z) - f^*(x,y,z) \right|^2 . \tag{5.11}$$

For these comparisons, the raw images were down-sampled by a factor of 4× in order to make a full tomographic reconstruction feasible on a single NVIDIA V100 GPU, and parallel reconstructions were carried out with only one reconstruction on a GPU at a time for benchmarking purposes.

Varying Sub-volume Size

First, we compared the performance for different sizes of individual sub-volumes, r_v, which trade off parallelization (compute speed) for reconstruction quality. Specifically, we increased volume sizes from 50^3 voxels to 250^3 voxels, with the overlap ratio being fixed at 25%. Figure 5.6 shows the result. As expected, decreasing the sub-volume size causes the reconstruction quality to deteriorate and artifacts to appear. This is because the proposed algorithm requires cropping of the projected image to reconstruct a reduced volume. However, electrons that scatter or diffract outside of the designated volume in 3D could still contribute to the projection data, and hence create artifacts during reconstruction. For larger sub-volume sizes, there is less contribution from outside of the sub-volume, and hence artifacts are reduced.

As predicted by the complexity analysis, reconstruction time decreases when more nodes are used and the size of each sub-volume decreases (Fig. 5.6(b)). Overall, many essential structures in the sample that appear in the full reconstruction are also present in the reconstructions at all sub-volume sizes. However, there are diminishing marginal returns on reconstruction time as we decrease the sub-volume size. When N is sufficiently small, other operations in the algorithm start to dominate the reconstruction time, such as memory transfer, sample rotation, and regularization. The increasing severity of artifacts prevented us from reducing the size further.

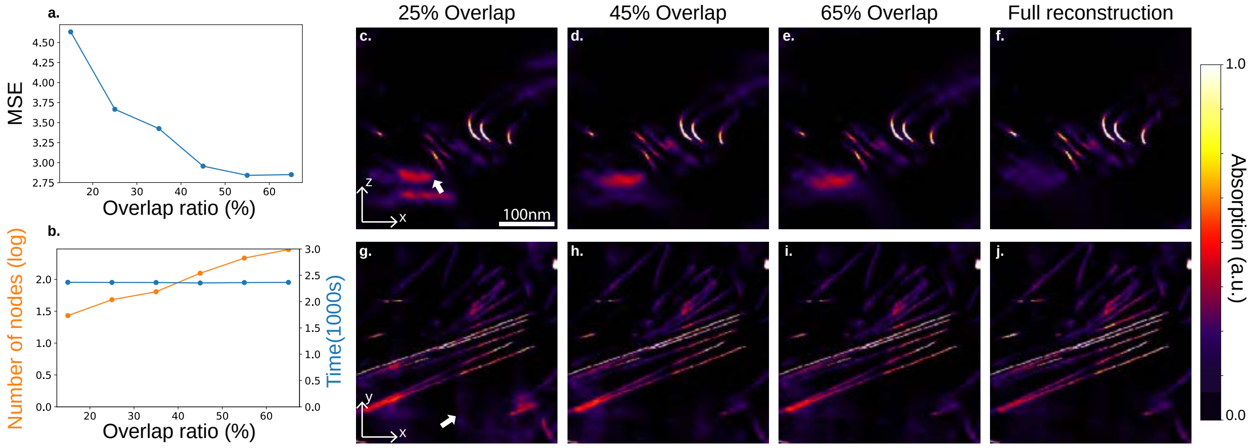

Figure 5.8: Increasing the overlap between adjacent sub-volumes reduces reconstruction errors. (a) Extra reconstruction error (MSE with respect to a single reconstruction) vs. overlap ratio. (b) Number of nodes needed and reconstruction time for each node as overlap ratio varies. (c)-(f) Single $z - x$ slice of reconstructed volume with various overlap ratios. (g)-(j) Single $y - x$ slice of reconstructed volume with various overlap ratios. (f) and (j) are full single reconstructions without any overlaps.

Varying Overlap Ratio

Next, we tested our algorithm's performance for varying overlap ratio, r_o. As mentioned in Section 5.4, overlaps are necessary to take into account edge effects due to diffraction and multiple-scattering. The overlap ratio is varied from 15% to 65% while maintaining the sub-volume size to be a cube with 150^3 voxels. Similar to the previous section, we compared the MSE and compute times when the parameter is changing, with results shown in Fig. 5.8. As predicted, with only 25% overlap (less than the minimum overlap of 30% calculated earlier in Section 5.4), the reconstruction suffers from some artifacts (see white arrows in Fig. 5.8(c) & (g)). With overlap ratio greater or equal to 45%, the MSE starts to converge. It is worth noting that the reconstruction time for individual sub-volume reconstructions are almost identical, because overlap ratio does not play a role in the complexity of individual sub-volume reconstructions; however, the number of nodes needed to cover the entire volume varies. Thus, the total complexity is still increasing as a function of overlap ratio.

Carbon Footprint

Since our methods will use significant compute power for large-volume high-resolution datasets, we report the equivalent carbon dioxide (CO_2) emission of the previous experiments. All numbers are calculated in units of Kilogram (Kg), and as a baseline we assume that each NVIDIA GPU V100 has a carbon footprint of 0.13 Kg/hr. Estimations were conducted using the MachineLearning Impact calculator presented in [72]. The total CO_2 emission is the baseline multiplied by total amount of computation taken across all nodes, as shown in Fig. 5.9. Figure 5.9(a) shows the carbon footprint for experiments that vary sub-volume sizes, and (b) shows the experiments that vary overlap ratio. In Fig. 5.9(a), the carbon dioxide emission converges above 100^3 voxels, because the increase in number of nodes M trades off with reduced time needed to reconstruct smaller volumes. However, the trade-off is no longer true for sizes smaller than 100^3 voxels, as computation overhead starts to dominate. The curve in Fig. 5.9(b) mostly adheres to the predicted result - since the size of sub-volumes remain the same for any overlap ratio, the CO_2 emission is scaling linearly with the number of nodes in order to cover the full volume. Therefore, the overlap ratio should be chosen to be as small as possible, yet satisfying the minimum derived in Section 5.4 to achieve an acceptable reconstruction quality.

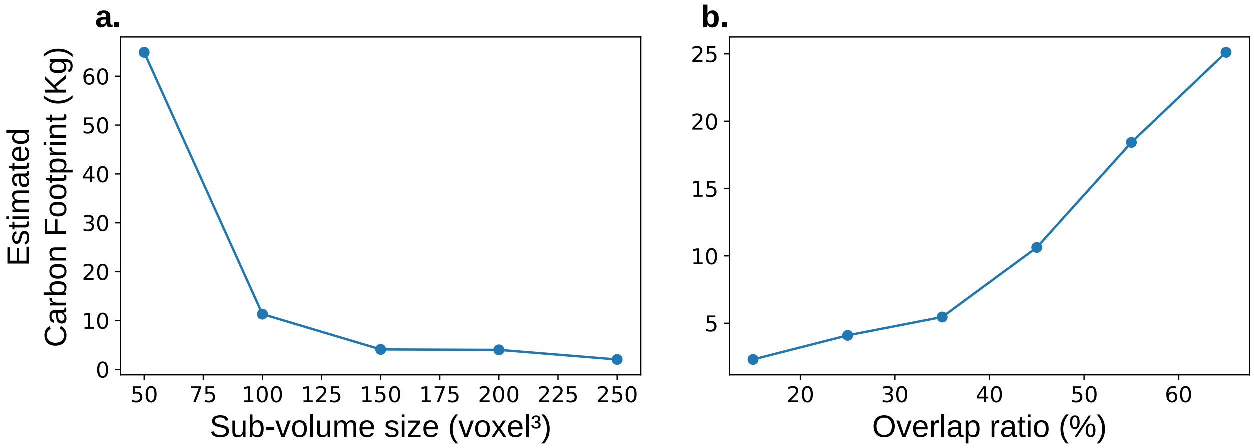

Figure 5.9: Carbon footprint with (a) varying sub-volume sizes and (b) varying overlap ratio. The optimal combination should be made by balancing the two parameters, such as sub-volume size of 150^3 voxels and a 35% overlap ratio.

Full Resolution Reconstruction

We show in Fig. 5.10 a reconstruction of the full lithium-montmorillonite dataset in a volume of $3952 \times 5500 \times 952$ voxels with isotropic resolution of 1.82 Å^3/voxel. This corresponds to a volume of $0.73(x) \times 1.00(y) \times 1.73(z)\mu m^3$. Figure 5.10(a) shows the maximum projection of the volume along the z-axis. All features are reconstructed, with occasional artifacts observed in the background, either due to noise or the stitching algorithm. We zoom in on three regions-of-interest (ROIs) in Fig. 5.10(b-j). Within each, the clay layers have different degrees of curvature and number of neighboring layers. For each sub-volume, two cross-sections at different depths are shown, along with a volume render. As pointed out with

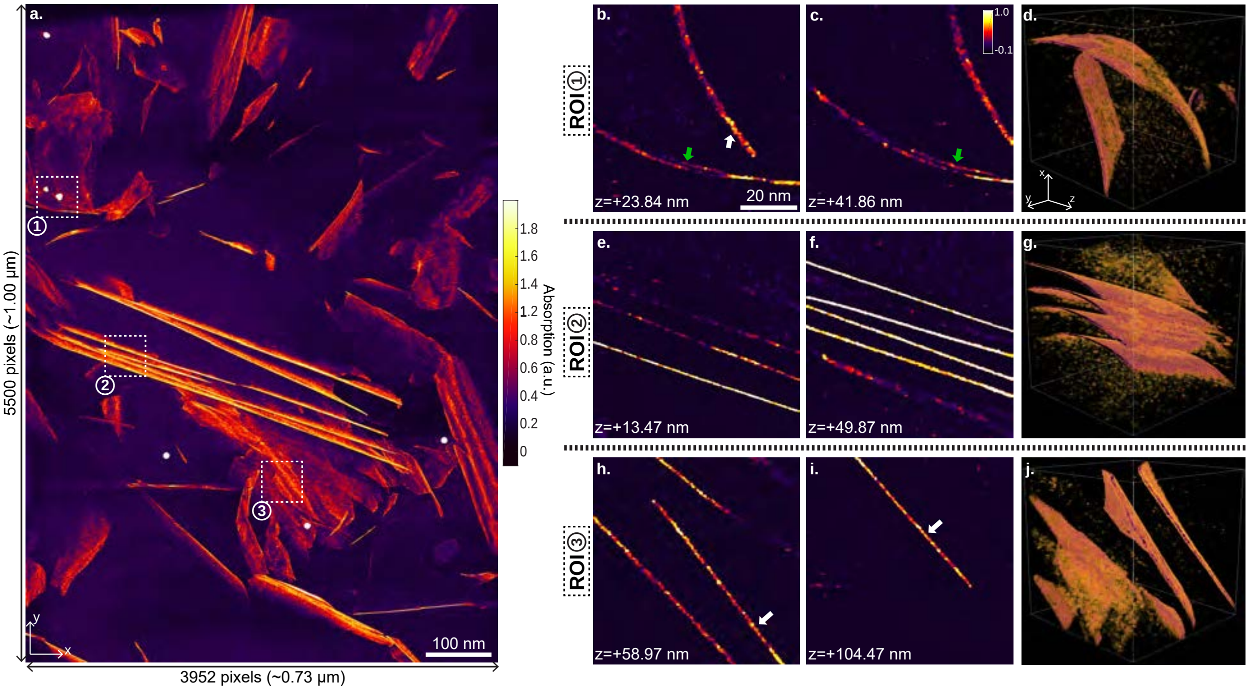

Figure 5.10: Full field-of-view 3D reconstruction of lithium-montmorillonite, with a total of 5500×3952×952 voxels. (a) Maximum depth projection of the absorption channel. Three smaller regions-of-interest (ROIs) are cropped and zoomed in, with sizes of 400^3 voxels. (b)-(c),(e)-(f),and (h)-(i) Show zoom-in lateral slices at different depth for ROI 1, 2, and 3, respectively. (d), (g), and (j) show 3D volume renders for each of the ROIs.

green arrows in Fig. 5.10(b,c), two layers that are closely stacked ($\sim$ 1nm separation) can be observed in ROI 1. If the full volume was reconstructed using the downsampled tilt-series, such observation would not be possible. Also enabled by high-resolution reconstruction, detailed crystal structures that were previously blurred out in [137] are now visible, as indicated by white arrows in Fig.5.10(b,h and i).

Notice that the total depth of the sample is significantly smaller than the other two dimensions, since the sample is fairly flat. Because our method is capable of reconstructing a custom-sized volume, a smaller depth is chosen such that computation resources are not wasted on empty space. The volume is broken down into 702 nodes of cubic volume reconstructions, with an isotropic overlap of 25%. Each sub-volume has a size of 500^3 voxels. The full volume stored in complex 32 bit float has a size of $\sim$154 GB, which is difficult to fit into the RAM of a computer, particularly with the overhead storage requirements of the algorithm. In addition, GPU acceleration is often needed in order for the algorithms to converge in a reasonable amount of time, and volumes with such sizes cannot be fit into the memory of a modern GPU without partitioning.

Discussion

The distributed computing algorithm we proposed allows large tomography datasets to be decomposed into smaller independent sub-problems for faster compute times on clusters. One can choose different values of overlap ratio (r_o) and sub-volume sizes (r_v) according to their tolerance of artifacts, or optimize parameters to minimize the overall carbon footprint while ensuring a reasonable reconstruction quality.

Once the full tilt-series is split into multiple independent projection data, reconstructions are carried out in parallel by any tomography algorithm of choice. In our work, to accurately model the interaction between the electron beam and the clay minerals in the large volume at microscopic level, we adopted the multi-slice algorithm [106] that is capable of accounting for multiple-scattering samples. Our distributed algorithm is similar to that in X-ray CT by Basu and Bresler [7, 8]; however, we use coherent multiple-scattering reconstruction algorithms instead of the filtered back-projection algorithm. We also do not continue to decompose the tomography problem in a hierarchical manner for further parallelization, as in [7, 8] because we find that there is a limit as to how small a sub-volume can be without suffering from severe artifacts. Further acceleration could be achieved by combining the slice-binning idea outlined in [106]. In this method, consecutive axial slices at each tilt angle are summed into a “thicker” slice during the forward propagation. In the update step, the error gradient is equally distributed back to individual slices. Slice-binning allows significant reduction of number of slices required to propagate through the 3D sample, without significant loss of accuracy.

To reduce the reconstruction artifacts, effort can be spent on exploring more advanced algorithms for volume stitching. In our study, we used a simple weighting function for all sub-volumes similar to [20, 22]. More complicated algorithms such as pyramid blending, or 2-band blending, could be explored [44, 32]. However, the major drawback of these methods is that they are content-aware to some extent, and could be altering the volume-of-interest in order to reduce stitching artifacts near the boundary.

5.5 Summary

In this chapter, we described multiple efficient computing strategies for 3D reconstruction from intensity-only TEM tilt-series data. By combining these methods with a multiple-scattering reconstruction algorithm, we have successfully demonstrated the performance boost. We compared various parameters on experimental data and showed performance difference in terms of MSE and reconstruction time, and we reconstructed a large 3D volume of size $(0.73(x) \times 1.00(y) \times 1.73(z)\mu\text{m}^3)$ with resolution of 1.82 Å^3/voxel, the largest cryo-ET reconstruction to our knowledge. With minimal restrictions for the scattering algorithm of choice, and minimal knowledge required of computer architecture, these methods open the door to larger tomographic reconstructions that require heavy computational resources, and provides great flexibility in choosing specific volumes-of-interest to recover.

Chapter 6

Conclusion and Future Directions

In this dissertation, I have explored a new possibility to achieve low-dose and efficient electron tomography for multiple-scattering samples. Conventional electron tomography methods always rely on linear projection assumptions between the sample and the electron probe, hence will either require very thin or weakly scattering samples, which poses great challenges to the sample preparation process, or demands large dosage on the sample, creating irreversible damage. The framework that we proposed allows for reconstruction of thicker and more scattering samples in the field of materials science and biology.

Specifically, in Chapter 2 we laid out the theoretical foundation for the framework. We chose to combine the the tilting of the sample under TEM plane wave illumination as well as the through-focus stack to recover its 3D structure. The tilt series of the sample offers a new 3D perspective of the sample, and is necessary in order to achieve isotropic resolution in x, y, and z. And since the samples do not absorb electrons, when illuminated by a collimated electron beam they show very little amplitude contrast in focus. Therefore, deliberate defocus images are necessary to show phase contrast and quantitatively recover the electric potentials of the samples. In addition to the imaging geometry, both the forward scattering model and inverse process for describing the interaction between electrons and samples were derived. The forward scattering model is part of the image formation processing given a 3D sample, and the inverse process is for iteratively updating the sample estimation. Lastly, in Chapter 2, we briefly described the regularization process in our work as an attempt to enforce our prior information about the sample.

In Chapter 3, our proposed method was validated via simulation at atomic resolution. In the beginning of the chapter our motivation as well as our choice of the phantom were explained, which includes over 100,000 individual atoms configured both as amorphous and crystalline structure. After that, detailed parameters for the simulation were laid out such as voxel size, tilt angles, electron wavelength, and so on. These parameters are practically designed so that they can be achieved in real experiments. Then, a series of scenarios were explored. First, we varied the dosage and mimicked the Poisson nature of the electron arrival process, and we showed that we were able to reconstruct the 3D volume with relatively high accuracy with as low a dosage as 7000 $e^-/\text{Å}^2$. Second, with a fixed amount of total dosage, we tested the trade-off between the number of tilt angles and defocus images. We concluded that at least 2 defocus planes are needed in order to recover quantitatively accurate potentials, and the rest of the electrons ought to be used to have denser tilt angles to improve 3D atomic position estimation accuracy. Third, the performance of our algorithm was tested against different degrees of tilt range. Up to 60° of missing wedge, our method can still recover individual atoms. However, since the axial resolution deteriorates as we increase the amount of missing wedge, the axial atomic position estimation accuracy worsens. In addition to the cases mentioned, we also tested different regularization methods, robustness against atom vacancy in the sample, heavy atoms, and poor experiment conditions when partial coherence is present.

In Chapter 4, we demonstrated an application of our framework in earth sciences. Asymmetric EDL structure along the clay surface normal direction was observed through reconstructing the sample's absorption profile. We first generated a phantom volume with realistic atomic structure of the Montmorillonite clay vitrified in aqueous solution. We then simulated various cases with uneven distribution of ions along the surfaces, giving rise to the asymmetric EDL profile. By forming the tilt-series from the phantom and reconstructing it, we demonstrated the possibility to observe the asymmetry from absorbance profile of the sample. In an experiment, a Montmorillonite clay sample is prepared in cryogenic mode, and a tilt-series is obtained. A sequence of pre-processing steps were performed in order to avoid model mismatch. After pre-processing, the full volume of clay was recovered, from which we segmented out individual layers to confirm what we observed earlier in simulation.

In Chapter 5, we revealed the backbone of our tomography framework – computation. Throughout the dissertation work, many attempts have been made at improving the computation efficiency of our algorithm. First, a slice-binning idea was implemented that avoids propagating the electron wave through all layers in the 3D volume by aggregating consecutive layers. We showed that the reconstruction has similar quality as without slice-binning, while offering as much as an order-of-magnitude boost in computation time. Second, a GPU-enabled python package is implemented to enable fast reconstruction. The package is open-source on Github and has been tested on other tilt-series datasets as well. Third, a distributed computing method was introduced to split the volume into multiple sub-volumes and simultaneously reconstruct them by manipulating the original tilt-series. The method is particularly useful when a compute cluster is available. It is also noteworthy that both diffraction and multiple scattering effects can be preserved in all of the optimizations we attempted.

While our electron tomography framework has been thoroughly tested both in simulation and some experimental datasets, another class of application has yet to be completed – atomic electron tomography (AET). So far, AET has only been done using HAADF STEM [141]. However, not only does HAADF STEM require a large dosage on the sample or strongly scattering atoms to obtain high SNR, it also assumes linear projection and does not account for multiple-scattering effects. As a result, samples that are beam sensitive or multiple-scattering cannot be imaged using this technique, and phase contrast electron tomography provides a natural alternative. Although atomic resolution datasets have been collected, high-accuracy recovery of individual atoms is yet to be achieved. To do that, the following problems need to be explored/solved:

- **Tilt-series alignment without fiducial markers**: Recall that in Chapter 4 gold nano particles were used as registration markers for estimating the tilt error and sample drift between frames. Unfortunately, samples examined at atomic resolution often have comparable size as the gold fiducial markers, so they are no longer available for tilt-series registration. Compared to HAADF STEM, phase contrast electron tomography also often requires more than one defocus image per tilt. Change in contrast at different defocus distances will create extra difficulty for the registration algorithm. Poor tilt-series alignment can cause 'fake' atoms or degraded resolution in the reconstruction. Previously, I explored a joint estimation scheme to solve for both sample and drift simultaneously. However, due to the high dimensionality and ill-posed nature of the problem, being accurate in estimating one can significantly affect the accuracy of the other. If one does not have a confident estimate of the 3D sample, they can hardly estimate an accurate drift. Additional prior information regarding the sample or the system may alleviate this issue.

- **Imaging system aberration correction**: Inconsistent system aberration in electron microscopes due to the change of parameters of magnetic lenses can also degrade the reconstruction quality. The major source of aberration is caused by the difference between the nominal defocus values and the true experimental values. After the sample is tilted to a new angle, it needs to be manually refocused, inconsistent judgement of zero focus can cause a deviation from nominal defocus values. Luckily, for amorphous atomic resolution materials, the defocus values can be estimated after a defocused intensity image is measured – the radial profile of its Fourier transform is characterized by the defocus coherent transfer function (a.k.a. Thon Rings [42]) and can be fitted to estimated the defocus values. The remainder of the system aberrations such as spherical, coma, or astigmatism are more subtle and need finer estimation and correction techniques. Note that this is a more general problem and is not limited to AET. It requires more attention in AET, because higher resolution and atomic position accuracy is desired.

- **Electron probe induced sample change (damage) during experiment**: During the experiments, fast traveling electrons can cause random irreversible effects to the sample, such as knocking atoms off, changing atom positions, or contaminating the sample. All current electron tomography methods (including ours) neglect this effect by assuming a consistent 3D sample throughout the experiment. While it might not be an issue for mid-resolution tomography, it can certainly degrade the reconstruction quality by either creating 'fake' atoms or confusing the alignment algorithms. One idea we have is to consider a time dynamic model to allow sample change during the experiment. Not only do we solve for the 3D sample, but also we estimate the change in between tilts. This model would vastly increase the dimensionality of the problem, and we expect that strong priors are needed to confine the solution space. For instance, the change of sample in between two consecutive acquisitions ought to be sparse.

Bibliography

[1] LJ Allen and MP Oxley. “Phase retrieval from series of images obtained by defocus variation”. In: *Optics communications* 199.1-4 (2001), pp. 65–75.

[2] LJ Allen et al. “Exit wave reconstruction at atomic resolution”. In: *Ultramicroscopy* 100.1-2 (2004), pp. 91–104.

[3] Roberta Angelini et al. “Glass–glass transition during aging of a colloidal clay”. In: *Nature communications* 5.1 (2014), pp. 1–7.

[4] Giacomo Argentero et al. “Unraveling the 3D atomic structure of a suspended graphene/hBN van der Waals heterostructure”. In: *Nano letters* 17.3 (2017), pp. 1409–1416.

[5] James D Atkinson et al. “The importance of feldspar for ice nucleation by mineral dust in mixed-phase clouds”. In: *Nature* 498.7454 (2013), pp. 355–358.

[6] Sara Bals et al. “Three-dimensional atomic imaging of colloidal core–shell nanocrystals”. In: *Nano letters* 11.8 (2011), pp. 3420–3424.

[7] Samit Basu and Yoram Bresler. “O (n/sup 2/log/sub 2/n) filtered backprojection reconstruction algorithm for tomography”. In: *IEEE Transactions on image Processing* 9.10 (2000), pp. 1760–1773.

[8] Samit Basu and Yoram Bresler. “O (N/sup 3/log N) backprojection algorithm for the 3-D Radon transform”. In: *IEEE transactions on medical imaging* 21.2 (2002), pp. 76–88.

[9] PE Batson, Niklas Dellby, and OL Krivanek. “Sub-ångstrom resolution using aberration corrected electron optics”. In: *Nature* 418.6898 (2002), p. 617.

[10] A. Beck and M. Teboulle. “Fast Gradient-Based Algorithms for Constrained Total Variation Image Denoising and Deblurring Problems”. In: *IEEE Transactions on Image Processing* 18.11 (2009), pp. 2419–2434. ISSN: 1057-7149. DOI: 10.1109/TIP.2009.2028250.

[11] MJ Beeching and AEC Spargo. “A method for crystal potential retrieval in HRTEM”. In: *Ultramicroscopy* 52.3-4 (1993), pp. 243–247.

[12] Tanmay AM Bharat et al. “Advances in single-particle electron cryomicroscopy structure determination applied to sub-tomogram averaging”. In: *Structure* 23.9 (2015), pp. 1743–1753.

[13] Richard E Blahut. *Theory of remote image formation*. Cambridge University Press, 2004.

[14] Ian C Bourg and Jonathan B Ajo-Franklin. "Clay, water, and salt: Controls on the permeability of fine-grained sedimentary rocks". In: *Accounts of chemical research* 50.9 (2017), pp. 2067–2074.

[15] Ian C Bourg et al. "Stern layer structure and energetics at mica–water interfaces". In: *The Journal of Physical Chemistry C* 121.17 (2017), pp. 9402–9412.

[16] John AG Briggs. "Structural biology in situ—the potential of subtomogram averaging". In: *Current opinion in structural biology* 23.2 (2013), pp. 261–267.

[17] BM Carpenter, C Marone, and DM Saffer. "Weakness of the San Andreas Fault revealed by samples from the active fault zone". In: *Nature Geoscience* 4.4 (2011), pp. 251–254.

[18] Jeffrey G Catalano. "Weak interfacial water ordering on isostructural hematite and corundum (0 0 1) surfaces". In: *Geochimica et Cosmochimica Acta* 75.8 (2011), pp. 2062–2071.

[19] F-R Chen, D Van Dyck, and C Kisielowski. "In-line three-dimensional holography of nanocrystalline objects at atomic resolution". In: *Nature communications* 7 (2016), p. 10603.

[20] Michael Chen et al. "Multi-layer Born multiple-scattering model for 3D phase microscopy". In: *Optica* 7.5 (2020), pp. 394–403.

[21] Yifan Cheng. "Single-particle cryo-EM—How did it get here and where will it go". In: *Science* 361.6405 (2018), pp. 876–880.

[22] Shwetadwip Chowdhury et al. "High-resolution 3D refractive index microscopy of multiple-scattering samples from intensity images". In: *Optica* 6.9 (2019), pp. 1211–1219.

[23] WMJ Coene et al. "Maximum-likelihood method for focus-variation image reconstruction in high resolution transmission electron microscopy". In: *Ultramicroscopy* 64.1-4 (1996), pp. 109–135.

[24] Sean Michael Collins et al. "Entropic comparison of atomic-resolution electron tomography of crystals and amorphous materials". In: *Physical review letters* 119.16 (2017), p. 166101.

[25] John M Cowley and A F₋ Moodie. "The scattering of electrons by atoms and crystals. I. A new theoretical approach". In: *Acta Crystallographica* 10.10 (1957), pp. 609–619.

[26] Ulrich Dahmen et al. "Background, status and future of the transmission electron aberration-corrected microscope project". In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 367.1903 (2009), pp. 3795–3808.

[27] Amer Deirieh et al. “Particle arrangements in clay slurries: The case against the honeycomb structure”. In: *Applied Clay Science* 152 (2018), pp. 166–172.

[28] RAB Devine and J Arndt. “Si—O bond-length modification in pressure-densified amorphous SiO 2”. In: *Physical Review B* 35.17 (1987), p. 9376.

[29] Yijue Diao and Rosa M Espinosa-Marzal. “Molecular insight into the nanoconfined calcite–solution interface”. In: *Proceedings of the National Academy of Sciences* 113.43 (2016), pp. 12047–12052.

[30] Allison Doerr. “Single-particle cryo-electron microscopy”. In: *Nature methods* 13.1 (2015), p. 23.

[31] Kenneth H Downing and Robert M Glaeser. “Restoration of weak phase-contrast images recorded with a high degree of defocus: the “twin image” problem associated with CTF correction”. In: *Ultramicroscopy* 108.9 (2008), pp. 921–928.

[32] Alexei A Efros and William T Freeman. “Image quilting for texture synthesis and transfer”. In: *Proceedings of the 28th annual conference on Computer graphics and interactive techniques.* 2001, pp. 341–346.

[33] RF Egerton, P Li, and M Malac. “Radiation damage in the TEM and SEM”. In: *Micron* 35.6 (2004), pp. 399–409.

[34] Gerald E Farin and Gerald Farin. *Curves and surfaces for CAGD: a practical guide.* Morgan Kaufmann, 2002.

[35] Marco Favaro et al. “Unravelling the electrochemical double layer by direct probing of the solid/liquid interface”. In: *Nature communications* 7.1 (2016), pp. 1–8.

[36] José-Jesús Fernández, Dan Gordon, and Rachel Gordon. “Efficient parallel implementation of iterative reconstruction algorithms for electron tomography”. In: *Journal of Parallel and Distributed Computing* 68.5 (2008), pp. 626–640.

[37] Richard A Frazin, Alberto M Vásquez, and Farzad Kamalabadi. “Quantitative, three-dimensional analysis of the global corona with multi-spacecraft differential emission measure tomography”. In: *The Astrophysical Journal* 701.1 (2009), p. 547.

[38] H Friedrich, MR McCartney, and PR Buseck. “Comparison of intensity distributions in tomograms from BF TEM, ADF STEM, HAADF STEM, and calculated tilt series”. In: *Ultramicroscopy* 106.1 (2005), pp. 18–27.

[39] Si Gao et al. “Electron ptychographic microscopy for three-dimensional imaging”. In: *Nature communications* 8.1 (2017), p. 163.

[40] Rw W Gerchberg. “Phase determination for image and diffraction plane pictures in the electron microscope”. In: *Optik (Stuttgart)* 34 (1971), p. 275.

[41] Peter Gilbert. “Iterative methods for the three-dimensional reconstruction of an object from projections”. In: *Journal of theoretical biology* 36.1 (1972), pp. 105–117.

[42] Robert M Glaeser. “Invited Review Article: Methods for imaging weak-phase objects in electron microscopy”. In: *Review of Scientific Instruments* 84.11 (2013), p. 312.

[43] Dan Gordon and Rachel Gordon. “Component-averaged row projections: A robust, block-parallel scheme for sparse linear systems”. In: *SIAM Journal on Scientific Computing* 27.3 (2005), pp. 1092–1117.

[44] Nuno Gracias et al. “Fast image blending using watersheds and graph cuts”. In: *Image and Vision Computing* 27.5 (2009), pp. 597–607.

[45] MA Gribelyuk. “Structure retrieval in HREM”. In: *Acta Crystallographica Section A: Foundations of Crystallography* 47.6 (1991), pp. 715–723.

[46] Maximilian Haider et al. “Electron microscopy image enhanced”. In: *Nature* 392.6678 (1998), p. 768.

[47] Johan Hattne et al. “MicroED with the Falcon III direct electron detector”. In: *IUCrJ* 6.5 (2019), pp. 921–926.

[48] Yang He et al. “In situ transmission electron microscopy probing of native oxide and artificial layers on silicon nanoparticles for lithium ion batteries”. In: *Acs Nano* 8.11 (2014), pp. 11816–11823.

[49] Jordon D Hemingway et al. “Mineral protection regulates long-term global preservation of natural organic carbon”. In: *Nature* 570.7760 (2019), pp. 228–231.

[50] Gabor T Herman. *Fundamentals of computerized tomography: image reconstruction from projections*. Springer Science & Business Media, 2009.

[51] Mert Hidayetoglu et al. “Petascale xct: 3d image reconstruction with hierarchical communications on multi-gpu nodes”. In: *arXiv preprint arXiv:2009.07226* (2020).

[52] Mert Hidayetoğlu et al. “Memxct: Memory-centric x-ray ct reconstruction with massive parallelization”. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 2019, pp. 1–56.

[53] Michael F Hochella et al. “Natural, incidental, and engineered nanomaterials and their impacts on the Earth system”. In: *Science* 363.6434 (2019).

[54] MF Hochella Jr and AH Carim. “A reassessment of electron escape depths in silicon and thermally grown silicon dioxide thin films”. In: *Surface science* 197.3 (1988), pp. L260–L268.

[55] Andre Hüpers et al. “Release of mineral-bound water prior to subduction tied to shallow seismogenic slip off Sumatra”. In: *Science* 356.6340 (2017), pp. 841–844.

[56] Matt J Ikari et al. “Spectrum of slip behaviour in Tohoku fault zone samples at plate tectonic slip rates”. In: *Nature Geoscience* 8.11 (2015), pp. 870–874.

[57] Simon Imhof et al. “Cryo electron tomography with Volta phase plate reveals novel structural foundations of the 96-nm axonemal repeat in the pathogen Trypanosoma brucei”. In: *Elife* 8 (2019), e52058.

[58] Jacob N Israelachvili. *Intermolecular and surface forces*. Academic press, 2015.

[59] CL Jia et al. "Determination of the 3D shape of a nanoscale crystal with atomic resolution from a single image". In: *Nature materials* 13.11 (2014), p. 1044.

[60] Xiaoming Jiang, Wouter Van den Broek, and Christoph T Koch. "Inverse dynamical photon scattering (IDPS): an artificial neural network based algorithm for three-dimensional quantitative imaging in optical microscopy". In: *Optics express* 24.7 (2016), pp. 7006–7018.

[61] Zhong Jingshan et al. "Transport of Intensity phase imaging by intensity spectrum fitting of exponentially spaced defocus planes". In: *Opt. Express* 22.9 (2014), pp. 10661–10674. DOI: `10.1364/OE.22.010661`. URL: `http://www.opticsexpress.org/abstract.cfm?URI=oe-22-9-10661`.

[62] Avinash C.. Kak and Malcolm Slaney. *Principles of computerized tomographic imaging*. IEEE press New York, 1988.

[63] Ulugbek S Kamilov et al. "Learning approach to optical tomography". In: *Optica* 2.6 (2015), pp. 517–522.

[64] Dari Kimanius et al. "Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2". In: *Elife* 5 (2016).

[65] Koji Kimoto et al. "Quantitative evaluation of temporal partial coherence using 3D Fourier transforms of through-focus TEM images". In: *Ultramicroscopy* 134 (2013), pp. 86–93.

[66] Earl J Kirkland. *Advanced computing in electron microscopy*. Springer Science & Business Media, 2010.

[67] Earl J Kirkland. "Nonlinear high resolution image processing of conventional transmission electron micrographs: I. Theory". In: *Ultramicroscopy* 9.1-2 (1982), pp. 45–64.

[68] Max Knoll and Ernst Ruska. "Das elektronenmikroskop". In: *Zeitschrift für physik* 78.5 (1932), pp. 318–339.

[69] Shan Shan Kou and Colin JR Sheppard. "Image formation in holographic tomography". In: *Optics letters* 33.20 (2008), pp. 2362–2364.

[70] James R Kremer, David N Mastronarde, and J Richard McIntosh. "Computer visualization of three-dimensional image data using IMOD". In: *Journal of structural biology* 116.1 (1996), pp. 71–76.

[71] F Krumeich et al. "Electron microscopy characterization of silicon dioxide nanotubes". In: *Zeitschrift für anorganische und allgemeine Chemie* 630.7 (2004), pp. 1054–1058.

[72] Alexandre Lacoste et al. "Quantifying the Carbon Emissions of Machine Learning". In: *arXiv preprint arXiv:1910.09700* (2019).

[73] Rowan Leary, Paul A Midgley, and John Meurig Thomas. "Recent advances in the application of electron tomography to materials chemistry". In: *Accounts of chemical research* 45.10 (2012), pp. 1782–1791.

[74] Benjamin A Legg et al. "Visualization of Aluminum Ions at the Mica Water Interface Links Hydrolysis State-to-Surface Potential and Particle Adhesion". In: *Journal of the American Chemical Society* 142.13 (2020), pp. 6093–6102.

[75] M Lentzen. "Reconstruction of the projected electrostatic potential in high-resolution transmission electron microscopy including phenomenological absorption". In: *Ultramicroscopy* 110.5 (2010), pp. 517–526.

[76] M Lentzen and K Urban. "Reconstruction of the projected crystal potential in transmission electron microscopy by means of a maximum-likelihood refinement algorithm". In: *Acta Crystallographica Section A: Foundations of Crystallography* 56.3 (2000), pp. 235–247.

[77] Peng Li and Andrew Maiden. "Multi-slice ptychographic tomography". In: *Scientific reports* 8.1 (2018), p. 2049.

[78] Hsiou-Yuan Liu et al. "SEAGLE: Sparsity-driven image reconstruction under multiple scattering". In: *IEEE Transactions on Computational Imaging* 4.1 (2017), pp. 73–86.

[79] Vladan Lučić, Alexander Rigort, and Wolfgang Baumeister. "Cryo-electron tomography: the challenge of doing structural biology in situ". In: *J Cell Biol* 202.3 (2013), pp. 407–419.

[80] Andrew M Maiden, Martin J Humphry, and JM Rodenburg. "Ptychographic transmission microscopy in three dimensions using a multi-slice approach". In: *JOSA A* 29.8 (2012), pp. 1606–1614.

[81] Daniel Martin-Jimenez et al. "Atomically resolved three-dimensional structures of electrolyte aqueous solutions near a solid surface". In: *Nature communications* 7.1 (2016), pp. 1–7.

[82] G McMullan et al. "Comparison of optimal performance at 300 keV of three direct electron detectors for use in low dose electron microscopy". In: *Ultramicroscopy* 147 (2014), pp. 156–163.

[83] Paul A Midgley and Rafal E Dunin-Borkowski. "Electron tomography and holography in materials science". In: *Nature materials* 8.4 (2009), p. 271.

[84] DL Misell. "An examination of an iterative method for the solution of the phase problem in optics and electron optics: I. Test calculations". In: *Journal of Physics D: Applied Physics* 6.18 (1973), p. 2200.

[85] Tiziana Missana and Andrés Adell. "On the applicability of DLVO theory to the prediction of clay colloids stability". In: *Journal of Colloid and Interface Science* 230.1 (2000), pp. 150–156.

[86] Paul Müller, Mirjam Schürmann, and Jochen Guck. "The theory of diffraction tomography". In: *arXiv preprint arXiv:1507.00466* (2015).

[87] Takanori Nakane et al. "Single-particle cryo-EM at atomic resolution". In: *Nature* 587.7832 (2020), pp. 152–156.

[88] M Niwa et al. "SiO2/Si Interfaces Studied by Scanning Tunneling Microscopy and High Resolution Transmission Electron Microscopy". In: *Journal of the Electrochemical Society* 139.3 (1992), pp. 901–906.

[89] K Norrish. "The swelling of montmorillonite". In: *Discussions of the Faraday society* 18 (1954), pp. 120–134.

[90] Colin Ophus. "Four-dimensional scanning transmission electron microscopy (4D-STEM): From scanning nanodiffraction to ptychography and beyond". In: *Microscopy and Microanalysis* 25.3 (2019), pp. 563–582.

[91] Colin Ophus, Jim Ciston, and Chris T Nelson. "Correcting nonlinear drift distortion of scanning probe and scanning transmission electron microscopies from image pairs with orthogonal scan directions". In: *Ultramicroscopy* 162 (2016), pp. 1–9.

[92] Colin Ophus et al. "Automatic software correction of residual aberrations in reconstructed HRTEM exit waves of crystalline samples". In: *Advanced structural and chemical imaging* 2.1 (2017), p. 15.

[93] Colin Ophus et al. "Efficient linear phase contrast in scanning transmission electron microscopy with matched illumination and detector interferometry". In: *Nature communications* 7 (2016), p. 10719.

[94] Alan W Paeth. "A fast algorithm for general raster rotation". In: *Graphics Interface.* Vol. 86. 5. 1986.

[95] Neal Parikh and Stephen P Boyd. "Proximal Algorithms". In: *Foundations and Trends in optimization* 1.3 (2014), pp. 127–239.

[96] Sung-Ho Park and Garrison Sposito. "Structure of water adsorbed on a mica surface". In: *Physical Review Letters* 89.8 (2002), p. 085501.

[97] Adam Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In: *arXiv preprint arXiv:1912.01703* (2019).

[98] Philipp M Pelz et al. "Phase-contrast imaging of multiply-scattering extended objects at atomic resolution by reconstruction of the scattering matrix". In: *Physical Review Research* 3.2 (2021), p. 023159.

[99] Stephen J Pennycook and Peter D Nellist. *Scanning transmission electron microscopy: imaging and analysis.* Springer Science & Business Media, 2011.

[100] Minh Pham et al. "RESIRE: real space iterative reconstruction engine for Tomography". In: *arXiv preprint arXiv:2004.10445* (2020).

[101] Thanh-An Pham et al. “Adaptive Regularization for Three-dimensional Optical Diffraction Tomography”. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 182–186.

[102] Alan Pryor et al. “GENFIRE: A generalized Fourier iterative reconstruction algorithm for high-resolution 3D imaging”. In: *Scientific reports* 7.1 (2017), p. 10409.

[103] Johann Radon. “1.1 über die bestimmung von funktionen durch ihre integralwerte längs gewisser mannigfaltigkeiten”. In: *Classic papers in modern diagnostic radiology* 5 (2005), p. 21.

[104] Johann Radon. “On the determination of functions from their integral values along certain manifolds”. In: *IEEE transactions on medical imaging* 5.4 (1986), pp. 170–176.

[105] Daniella M Rempe and William E Dietrich. “Direct observations of rock moisture, a hidden component of the hydrologic cycle”. In: *Proceedings of the National Academy of Sciences* 115.11 (2018), pp. 2664–2669.

[106] David Ren et al. “A multiple scattering algorithm for three dimensional phase contrast atomic electron tomography”. In: *Ultramicroscopy* 208 (2020), p. 112860.

[107] David Ren et al. “Total-variation regularized Fourier ptychographic microscopy with multiplexed coded illumination”. In: *Imaging and Applied Optics 2017 (3D, AIO, COSI, IS, MATH, pcAOP)*. Optical Society of America, 2017, p. MM3C.5. DOI: `10.1364/MATH.2017.MM3C.5`. URL: `http://www.osapublishing.org/abstract.cfm?URI=MATH-2017-MM3C.5`.

[108] Maria Ricci, Peter Spijker, and Kislon Voïtchovsky. “Water-induced correlation between single ions imaged at the solid–liquid interface”. In: *Nature communications* 5.1 (2014), pp. 1–8.

[109] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. “Nonlinear Total Variation Based Noise Removal Algorithms”. In: *Phys. D* 60.1-4 (1992), pp. 259–268. ISSN: 0167-2789. DOI: `10.1016/0167-2789(92)90242-F`. URL: `http://dx.doi.org/10.1016/0167-2789(92)90242-F`.

[110] Ernst Ruska. “The development of the electron microscope and of electron microscopy”. In: *Bioscience reports* 7.8 (1987), pp. 607–629.

[111] Barbara Ruzicka and Emanuela Zaccarelli. “A fresh look at the Laponite phase diagram”. In: *Soft Matter* 7.4 (2011), pp. 1268–1286.

[112] Gregory Samelsohn. “Invertible propagator for plane wave illumination of forward-scattering structures”. In: *Applied optics* 56.14 (2017), pp. 4029–4038.

[113] Mary Cooper Scott. *Experimental Atomic Resolution Electron Tomography.* University of California, Los Angeles, 2015.

[114] P G_ Self et al. “Practical computation of amplitudes and phases in electron diffraction”. In: *Ultramicroscopy* 11.1 (1983), pp. 35–52.

[115] Patrik Sellin and Olivier X Leupin. “The use of clay as an engineered barrier in radioactive-waste management–a review”. In: *Clays and Clay Minerals* 61.6 (2013), pp. 477–498.

[116] James L Suter et al. “Large-scale molecular dynamics study of montmorillonite clay: emergence of undulatory fluctuations and determination of material properties”. In: *The Journal of Physical Chemistry C* 111.23 (2007), pp. 8248–8259.

[117] Akihiro Suzuki et al. “High-resolution multislice x-ray ptychography of extended thick objects”. In: *Physical review letters* 112.5 (2014), p. 053903.

[118] Akihiro Suzuki et al. “High-resolution multislice x-ray ptychography of extended thick objects”. In: *Physical review letters* 112.5 (2014), p. 053903.

[119] ME Swanwick et al. “Nanostructured ultrafast silicon-tip optical field-emitter arrays”. In: *Nano letters* 14.9 (2014), pp. 5035–5043.

[120] Kiyofumi Takaba et al. “Protein and organic-molecular crystallography with 300kV electrons on a direct electron detector”. In: *Frontiers in molecular biosciences* 7 (2021), p. 440.

[121] Hajime Tanaka, Jacques Meunier, and Daniel Bonn. “Nonergodic states of charged colloidal suspensions: Repulsive and attractive glasses and gels”. In: *Physical Review E* 69.3 (2004), p. 031404.

[122] Michael Reed Teague. “Deterministic phase retrieval: a Green’s function solution”. In: *JOSA* 73.11 (1983), pp. 1434–1441.

[123] Philippe Thevenaz, Urs E Ruttimann, and Michael Unser. “A pyramid approach to subpixel registration based on intensity”. In: *IEEE transactions on image processing* 7.1 (1998), pp. 27–41.

[124] Lei Tian and Laura Waller. “3D intensity and phase imaging from light field measurements in an LED array microscope”. In: *Optica* 2.2 (2015), pp. 104–111. DOI: 10.1364/OPTICA.2.000104. URL: http://www.osapublishing.org/optica/abstract.cfm?URI=optica-2-2-104.

[125] Christophe Tournassat and Carl I Steefel. “Reactive transport modeling of coupled processes in nanoporous media”. In: *Reviews in Mineralogy and Geochemistry* 85.1 (2019), pp. 75–109.

[126] Thomas R Underwood and Ian C Bourg. “Large-scale molecular dynamics simulation of the dehydration of a suspension of smectite clay nanoparticles”. In: *The Journal of Physical Chemistry C* 124.6 (2020), pp. 3702–3714.

[127] Knut W Urban. “Studying atomic structures by aberration-corrected transmission electron microscopy”. In: *Science* 321.5888 (2008), pp. 506–510.

[128] Sandra Van Aert et al. “Three-dimensional atomic imaging of crystalline nanoparticles”. In: *Nature* 470.7334 (2011), p. 374.

[129] Wouter Van den Broek and Christoph T Koch. “General framework for quantitative three-dimensional reconstruction from arbitrary detection geometries in TEM”. In: *Physical Review B* 87.18 (2013), p. 184108.

[130] Wouter Van den Broek and Christoph T Koch. “Method for retrieval of the three-dimensional object potential by inversion of dynamical electron scattering”. In: *Physical review letters* 109.24 (2012), p. 245502.

[131] Wouter Van den Broek and Christoph T Koch. “Method for retrieval of the three-dimensional object potential by inversion of dynamical electron scattering”. In: *Physical review letters* 109.24 (2012), p. 245502.

[132] Juan-Jesus Velasco-Velez et al. “The structure of interfacial water on gold electrodes studied by x-ray absorption spectroscopy”. In: *Science* 346.6211 (2014), pp. 831–834.

[133] Miloš Vulović et al. “When to use the projection assumption and the weak-phase object approximation in phase contrast cryo-EM”. In: *Ultramicroscopy* 136 (2014), pp. 61–66.

[134] Laura Waller and Lei Tian. “Computational imaging: machine learning for 3D microscopy”. In: *Nature* 523.7561 (2015), p. 416.

[135] Xiao Wang et al. “Consensus equilibrium framework for super-resolution and extreme-scale ct reconstruction”. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis.* 2019, pp. 1–23.

[136] Yu-Chieh Wen et al. “Unveiling microscopic structures of charged water interfaces by surface-specific vibrational spectroscopy”. In: *Physical review letters* 116.1 (2016), p. 016101.

[137] Michael L Whittaker et al. “Dynamic clay microstructures emerge via ion complexation waves”. In: *arXiv preprint arXiv:2012.09295* (2020).

[138] Michael L Whittaker et al. “Ion exchange selectivity in clay is controlled by nanoscale chemical–mechanical coupling”. In: *Proceedings of the National Academy of Sciences* 116.44 (2019), pp. 22052–22057.

[139] Michael L Whittaker et al. “Layer size polydispersity in hydrated montmorillonite creates multiscale porosity networks”. In: *Applied Clay Science* 190 (2020), p. 105548.

[140] Rui Xu et al. “Three-dimensional coordinates of individual atoms in materials revealed by electron tomography”. In: *Nature materials* 14.11 (2015), p. 1099.

[141] Yongsoo Yang et al. “Deciphering chemical order/disorder and material properties at the single-atom level”. In: *Nature* 542.7639 (2017), p. 75.

[142] Li-Hao Yeh et al. “Experimental robustness of Fourier ptychography phase retrieval algorithms”. In: *Opt. Express* 23.26 (2015), pp. 33214–33240. DOI: 10.1364/OE.23.033214. URL: http://www.opticsexpress.org/abstract.cfm?URI=oe-23-26-33214.

[143] Francisco Zaera. “Probing liquid/solid interfaces at the molecular level”. In: *Chemical reviews* 112.5 (2012), pp. 2920–2986.

[144] Piotr Zarzycki et al. “Lateral water structure connects metal oxide nanoparticle faces”. In: *Journal of Materials Research* 34.3 (2019), pp. 456–464.

[145] Chun Zhu et al. “Towards three-dimensional structural determination of amorphous materials at atomic resolution”. In: *Physical Review B* 88.10 (2013), p. 100201.