DIRECT ELECTRON DETECTORS: ARCHITECTURE AND ALGORITHMS

A Thesis
Presented to the Faculty of the Graduate School
of Cornell University
in Partial Fulfillment of the Requirements for the Degree of
M.S.

by
Gabriela Calinao Correa
May 2020
ABSTRACT

Direct electron detectors (DED) are enabling new materials imaging techniques such as four-dimensional scanning transmission electron microscopy (4D-STEM)—and are now able to image biological specimens with higher resolution than x-rays. Limitations imposed by semiconductor manufacturing processes, where consumer electronics dominate device chip surface area, reduce the available space for DED pixels. To push beyond the physical pixel, sub-pixel super-resolution is necessary. I explore the prospects for sub-pixel super-resolution through electron counting as a function of diode depth, pixel pitch, and beam energy. For most energy ranges of interest to electron microscopy, energy is deposited as a string of charge across multiple pixels. I use machine learning to identify the start of the string, determining the true entry point of an electron with greater success than existing electron counting statistics.

To test the effectiveness of different electron counters across beam energies and DED architectures, the electron counter is fed virtual detector readouts. Three electron counters are tested: the maximum intensity (peak) pixel; the center of mass (mean) point; and a convolutional neural network with a rectified linear unit (ReLU). I simulated primary beam energies from 30 to 5,000 keV for silicon and germanium diodes, with pixel pitches from 1 to 500 µm and diode depths from 10 to 1,000 µm. Electron paths are generated through an electron Monte Carlo method with relativistic corrections, then projected into a range of virtual detector pixels. The root mean squared error between the true entry point and the counters guessed entry point is used as a metric of performance. The simulations are performed, assuming a perfect signal with no additional
noise, to test the maximum capability of counter performance.

Super-resolution counting is effective up to 300 keV for a 500 µm diode, and up to 100 keV for shallower diode depth of 50 µm. The machine learning model has great performance with a training dataset of significantly smaller size than a typical 4D-STEM dataset. Similar trends for all counters are observed with respect to beam energy, pixel pitch, and diode depth. The electron counters are generally most dependent on diode depth for performance, followed by beam energy. There are three regions of dependencies on diode depth: a barreling region for backthinned diodes where the electron passes straight through; a peak error region where the electron path is cut off but given sufficient time to wander; and a region where the full point spread function is captured. By combining machine learning with a deep diode, a counting mode via integration is achieved.
This thesis is dedicated to my brother.
ACKNOWLEDGEMENTS

First and foremost, immeasurable thanks are given to my advisor, David Muller. Many thanks to: the ever inspiring Nicole Benedek, for insightful discussion; David Bindel, for invaluable wise advice and mathematical discussion; Darrel Schlom, for encouraging remarks and endorsing my research ideas; and Jin Suntivich, for greatly supporting my continued education while concussed in 2017. Thanks are given to several faculty in computer science including: Anil Damle, for useful discussion in image segmentation and super-resolution; Austin Benson, for useful discussion on tensor representation for 4D-STEM; and Adrian Sampson, for helpful pointers to mounting convolutional neural networks on FPGAs, and for building an inclusive environment.

I greatly acknowledge support by the Alfred P. Sloan Foundation and Department of Energy Computational Science Graduate Fellowship (DOE CSGF), which is provided under grant number DE-FG02-97ER25308. Gatherings of Sloan Fellows at Cornell provided an incredible sense of community I might have not otherwise found at Cornell. Thanks are owed to Jami Joyner and Mark Lewis, for guidance throughout my academic path. The community built by the DOE CSGF is an energetic, inspiring force, which I owe much motivation to. Many thanks to my mentor, Judith Hill, for encapsulating this energy and guiding me along my tumultuous path. I’d like to thank amazing friends and mentors met at Lawrence Berkeley National Lab, including Peter Ercius, Colin Ophus, Judy Cha, and Mary Scott. Many thanks to all who taught me of detectors, including Ian Johnson, Azriel Goldschmidt, Peter Denes, and Mark Tate.

This work used resources provided by the National Science Foundation Platform for the Accelerated Realization, Analysis, and Discovery of Interface Materials (PARADIM) under Cooperative Agreement No. NSF-DMR-1539918,
and resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231. This work made use of the Cornell Center for Materials Research Shared Facilities which are supported through the NSF MRSEC program (DMR-1719875). The FEI Titan Themis 300 was acquired through NSF-MRI-1429155, with additional support from Cornell University, the Weill Institute and the Kavli Institute at Cornell. Additional support for the FIB/SEM cryo-stage and transfer system was provided by the Kavli Institute at Cornell and the Energy Materials Center at Cornell, DOE EFRC BES (DE-SC0001086).
TABLE OF CONTENTS

Biographical Sketch .................................................. iii
Dedication ................................................................. iv
Acknowledgements ...................................................... v
Table of Contents ...................................................... vii
List of Figures ........................................................... viii

1 Introduction ............................................................ 1
  1.1 Imaging .............................................................. 1
  1.2 Phase space electron imaging ................................. 2
    1.2.1 Tensor representations .................................. 4

2 Diode Design for Direct Electron Detectors .................. 8
  2.1 Introduction ...................................................... 8
    2.1.1 Direct Electron Detectors (DED) ....................... 8
  2.2 The electron wanders .......................................... 11
    2.2.1 Lattices and Electrical circuitry ................... 13
    2.2.2 Polya’s Theorem ......................................... 18
    2.2.3 Random walk in 3D .................................... 20
  2.3 Numerical Experiments ........................................ 22
    2.3.1 Diode Material .......................................... 24
  2.4 Virtual electron detection .................................. 24
  2.5 Results and Discussion ...................................... 29

3 Sub-pixel Super-resolution Electron Counting ............... 34
  3.1 Electron Counting ............................................. 35
    3.1.1 A simple electron counter .......................... 36
    3.1.2 Can a machine count? ................................. 37
    3.1.3 Image Segmentation ..................................... 46
  3.2 Outlook .......................................................... 48

4 Conclusions and Outlook ......................................... 50
  4.1 A Domain-Specific Architecture for Electron Microscopes . 50
  4.2 Conclusion ...................................................... 52
# LIST OF FIGURES

1.1 A demonstration of imaging technologies at various scales, including (a) a black hole\(^{16}\), (b) cronuts, and (c) skyrmions\(^{42}\) ........................................ 2

1.2 Microscope resolution over the years. Electron microscopy surpassed light microscopy in the middle of the 20th century. The prevalence of computing gave way to aberration-corrected microscopes in the 21st century. Direct electron detectors are paving the way towards resolution advancement, paired tightly with image analysis algorithms. ........................................ 3

2.1 Three different direct electron detectors currently available for use. The (a) Timepix3 created by CERN\(^{18}\) (b) MAPS detector at Lawrence Berkeley National Laboratory\(^{8}\) and (c) EMPAD at Cornell\(^{22}\) ........................................ 9

2.2 Direct electron detector architectures. A pixel array detector (PAD) is shown on the left, with a monolithic active pixel sensor (MAPS) shown on the right. Both detectors have a diode layer, each of different thickness. The diode layer of the PAD is dominated by a depletion zone above the PN diode, with doping on the lower layer adjacent to the solder bump, which is typically about 80 µm. A thin ohmic contact via deposited aluminum and implant doping is on the top layer of the PAD. Readout electronics for the PAD are located underneath the solder bumps. For MAPS readout electronics are located on top of the diode layer. 10

2.3 Histogram of pixel intensity for many electrons incident upon a PAD detector with 500 µm diode depth. From left to right the peaks include the zero-loss peak, a peak for one electron, and peaks for incrementally higher electron counts. This energetic information assists in differentiating coincident electrons. 12

2.4 Graphical lattice representations ........................................ 14

2.5 One dimensional closed lattice ........................................ 15

2.6 Two dimensional lattice, bounded by a sphere \(S^{(3)}\) ........................................ 17

2.7 Simple random walk in three dimensions ........................................ 20

2.8 Monte Carlo simulations of electron paths (only 1000 shown) at varying beam energies. With higher energy, the electron has greater capability to travel far from where it started. The top row contains a projected side view of the electron paths incident on the surface, while the bottom row is a plan view projection. Red dots indicate electron entry points. 23

2.9 Error with respect to energy comparing silicon performance (black line) with germanium (red line). Germanium is shown to be a superior diode material through our simulations 25
2.10 A single electron path with its overlaid detector readout, at 300 keV in a 500 µm deep diode with 50 µm pitch. The summation of all pixels is the energy of the incident electron, for sufficient diode depth. Different counting metrics for this particular path are labeled. .................................................. 26

2.11 A sampling of electron path histograms with varying pixel pitch. Most interestingly, the peak and mean pixels are a significant distance from the true entry point. Each row has the scale bar value labeled. .................................................. 27

2.12 A single electron path, cut off at three different diode depths. On the far left, a three dimensional bar plot is shown to display intensity differences in each pixel. The following plots are the same information projected onto a two dimensional representation at a 500, 300, and 50 µm diode depth respectively. ................. 28

2.13 Error with respect to energy for three different detector geometries shown in Figure 2.1. Pixel pitch is shown as a thin horizontal line, in each plot labeled by a diode depth. Super-resolution regimes are shown, as the regions where the error is less than the pixel pitch. The peak pixel metric is plotted as a solid line, while the center of mass is plotted as a dotted line. In general, center of mass has better performance than the peak pixel. .................. 30

2.14 Root mean squared error for peak (red) and mean (black) pixel as pixel width and depth varies, given perfect image segmentation and no readout noise. ............................................. 31

2.15 Performance dependencies on pixel pitch, diode depth, and incident electron energy. Normalized root mean squared error for the mean point of a detected path, given imaging conditions, is shown to illustrate trends. These slices of the data cue show key dependency on depth, followed by energy. ...................... 32

3.1 2D histogram \(h(x, y)\) of an electron’s path through the detector \(p(x, y, z)\), shown in white. The electron’s entry point at the origin is indicated by a red square. Centers of mass according to different pixel \(z\) depths are indicated by triangles: magenta for 500 µm, yellow for 300 µm, and cyan for 50 µm. ................. 40

3.2 Test and training mean error and standard deviation across different training sample sizes. Test data size was fixed at 500 samples. Performance already reaches subpixel accuracy at 500 training samples, but has inconsistent performance at larger data sets. ................................................................. 41

3.3 Training loss performance over epochs. Simulation was done for 300 kV electrons using 5 µm bins in an 80x80 grid. Loss is still overall decreasing in this particular instance, but other runs see divergence where loss increases dramatically at later epochs ... 42
3.4 2D histogram \( h(x, y) \) of an electron’s path through the detector and visualization of its 3 convolutional layers. Twelve (out of 32) channels are shown here. The electron’s entry point at the origin is indicated by a white square. Red square indicates prediction pixel from the neural network. This histogram is one of the test cases for the detector geometry of Medixpix of a 300 keV electron.

3.5 Initial results for root mean squared error (RMSE) plot comparing MIP, COM, and our neural network. Plots for the EMPAD are shown in red, Medpix in blue, D4MAPS in black, and a new architecture in yellow. Datasets for the ConvNets are not the same as the data used to calculate MIP and COM, making this an indirect comparison. Given previous results, double the number of data samples are needed to improve consistency for the ConvNet. Trends indicate that the ConvNet provides better resolution than MIP and COM outside of the super-resolution regime.

3.6 Simulations using 5 µm bins at 50, 300, and 500 µm diode depths. Error of the neural network versus COM and MIP are plotted for each depth. As depth increases, the COM and MIP gains in error due to longer electron propagation whereas the neural network stays fairly consistent. At low and high energies, COM and MIP are still good since the electron path is usually only one pixel.

3.7 Error with respect to energy for three different detector geometries. Pixel pitch is shown as a silver horizontal line, in each plot labeled by a diode depth. Super-resolution regimes are shown, as the regions where the error is less than the pixel pitch. The training and test performance of the machine-learning-based electron counter is shown, in comparison to the peak pixel calculated simultaneously.
A donut hole would be in a donut’s hole, but we must look a little closer. And when we do we see the donut hole has a hole in its center. It is not a donut hole, but a smaller donut with its own hole—and our donut is not a hole at all!”

–Benoît Blanc, Knives Out
CHAPTER 1
INTRODUCTION

1.1 Imaging

Materials discovery and design thrives amidst a broad selection of characterization techniques. These techniques, linking theory and synthesis, allow for the intense interrogation of materials behavior from sub-atomic levels to astronomical scales. For many years imaging largely lay in the experimental domain. With the advent of new detectors, high-performance computing, big data, machine learning, and visualization techniques—imaging technologies are undergoing a massive transformation. The cognification of imaging technologies is transforming tedious tasks into an automated commodity with intelligent instruments.

The original detectors in science, human eyes, have a long history: Biologists hired artists to sit at optical microscopes, sketching what they saw; Astronomers stared at the skies and charted the stars. Film quickly overtook these methods, imaging with high dynamic range and precision. As semiconductor technologies emerged, analog and digital data detection was realized. With digital data, quantitative interrogation of materials via computing emerged.

Illumination, lenses, detection: the three primary characteristics of many imaging systems. From skyrmions, to donuts, to black holes—today we have the technology to see in immense detail the world around us as shown in Figure 1.1. Quantitatively characterizing these details allow for the understanding and design of systems. At each length scale, a variety of sensors and algorithms
are regularly employed to acquire images.

![Figure 1.1: A demonstration of imaging technologies at various scales, including (a) a black hole, (b) cronuts, and (c) skyrmions.](image)

This thesis focuses on pushing the technological regime of sub-angstrom imaging with electron microscopy. Here, I focus on enabling materials by design through direct electron detectors (DED). The availability of fast, reliable, high-resolution detectors, and computing, has revolutionized imaging in electron microscopy. Several decades of development in cryogenic electron microscopy (cryoEM) achieved the Nobel Prize. DED finally enabled one to see biological specimens with higher resolution than x-rays.

### 1.2 Phase space electron imaging

Pushing the limits of visibility of the small gave rise to imaging the phase space of a material. Many methods measure either momentum space or real space.

*also referred to as diffraction space, Fourier space, and reciprocal space*
at a given acquisition. This is largely due to previous limitations in regards to computing resources and data acquisition speed.

Figure 1.2: Microscope resolution over the years. Electron microscopy surpassed light microscopy in the middle of the 20th century. The prevalence of computing gave way to aberration-corrected microscopes in the 21st century. Direct electron detectors are paving the way towards resolution advancement, paired tightly with image analysis algorithms.

Phase space electron imaging (PSEI) collects a full map of information in both domains. The sample of interest is probed at scan points $s(x, y)$ in real space. At each scan point a full diffraction pattern $P(p_x, p_y)$ is collected. This
forms an intensity tensor of $I(s_x, s_y, p_x, p_y)$.

The four-dimensional (4D) nature of this data was initially gathered in a scanning transmission electron microscope (STEM). As such this technique is also referred to as 4D-STEM in the literature, yet this term was previously assigned to other techniques. This technique does not necessitate STEM be used, yet for consistency with existing literature we shall continue referring to the technique as 4D-STEM.

DED enable the fast automated acquisition of diffraction data; through this the 4D-STEM datasets are found [1]. Improving the quality and usability of diffraction data expands the applicability of 4D-STEM, and is the purpose of this thesis. The sheer quantity of information gathered today is made possible by computed control of microscopes.

### 1.2.1 Tensor representations

Drawing analogy to traditional techniques, a 4D-STEM dataset may be reduced to images akin to those gathered by traditional techniques. The sample in real space $S(s_x, s_y)$ and diffraction patterns in momentum space $P(p_x, p_y)$ of a 4D-STEM dataset contains equivalent information.

Generically we have a transformation $T$ mapping information gathered in $X$ to another space $Y$, the image.

\[ T : X \rightarrow Y \]
Often, this transformation is simply the identity operation. That is, the information gathered directly forms the image saved. This is the case for bright field (BF) and annular dark field (ADF) detectors, along with selected area diffraction (SAD). For these cases $X, Y \in \mathbb{R}^2$.

These imaging techniques may be extended to tomographic methods, which in addition to the $S(s_x, s_y)$ scan, acquires information along a third dimension $\theta$ by tilting the sample between scans. This information then maps to $Y \in \mathbb{R}^3$, a 3D volume to visually inspect.

Selecting a subset of 4D-STEM data directly maps to these spaces as well. For ADF and BF images, reducing $P(p_x, p_y)$ matrix to a single value accomplishes this. The mapping $T$ in this case is a summation $\sum \rho$. In the simplest case this is $\forall \rho \in P(p_x, p_y)$ where each element is included in the summation. To form better images akin to ADF and BF we instead take $\rho$ to be a subset of $P(p_x, p_y)$ chosen within a circular or annular mask.

To extract a SAD pattern from a 4D-STEM intensity tensor $I(p_x, p_y, s_x, s_y)$ a single diffraction pattern $P(p_x, p_y)$ is selected from a position in the sample. This is essentially a digital version of a SAD experiment, and assists in analyzing data. An interface to select an area of the sample through a virtual ADF detector, and see its associated diffraction pattern is available.
T_{ADF} : I(s_x, s_y) \rightarrow \mathbb{R}^2  

(1.1)

T_{SAD} : I(p_x, p_y) \rightarrow \mathbb{R}^2  

(1.2)

T_{4D-STEM} : I(p_x, p_y, s_x, s_y) \rightarrow \mathbb{Z}^2  

(1.3)

T_{tomo} : I(s_x, s_y, \theta) \rightarrow \mathbb{R}^3  

(1.4)

T_{phare} : I(p_x, p_y, s_x, s_y, \theta, \phi) \rightarrow \mathbb{Z}^3  

(1.5)

We use this formalism to develop an understanding of detector dynamics, analyze algorithms, and build domain specific high performance computing architectures for the massive firehoses of data DED are evolving to be.

Materials innovation has led the way of detection technologies. Intimately understanding how materials interact with the energetic particles being detected is key to pushing resolution of electron microscopy. Chapter 2 investigates and expands upon our understanding of materials for detector technologies. To date, direct electron detectors have paved the ways to imaging a wide variety of materials. Still, these techniques have their limitations, and may not be applied to all materials.

Current DED are dose limited by existing counting statistics. Chapter 1 develops advanced counting algorithms to attain peak resolution. Reducing the electron point spread function to a singular point requires a more refined statistical approach for common imaging conditions. Extending imaging conditions to capture the full phase space of a material more urgently requires accurate counting techniques, as data dimensionality doubles and analysis algorithms are increasingly complex.

The electrons traverse a long voyage, beginning far from where this thesis
begins. Power generated unsustainably traverses lossy power cords towards a microscope. Can what the electron finds help them find a better beginning, through renewable resources? Transmitted through a crystal source, electrons begin their voyage through the microscope modulated by magnetic optics. Impacting the sample, a wealth of knowledge is gained. Tired from their long adventure, the electrons finally land on a detector.

This is where our epilogue begins, the final step of the electron’s journey. What happens to the electrons, once detected? Do they get lost on their way home? Upon first impact on the detector diode, Chapter 2 tells a story of how some electrons are lost, and most are found; more importantly, highlighting how we can save all the electrons. Chapter 3 shows how we can make every, single, electron count—increasing precision and accuracy while compressing data by orders of magnitude: distilling the data. And where the data goes is the domain of Chapter 4, where some electrons become photons for a short while as they traverse computing architectures.

Materials Imaging has reached incredible heights. As François Chollet once said,

As a general rule, if you need to improve a system by more than 3 orders of magnitude along some dimension, you’re likely to have to rebuild it on top of a completely different technology stack altogether

This thesis works towards a completely different technology stack for electron microscopy.
CHAPTER 2
DIODE DESIGN FOR DIRECT ELECTRON DETECTORS

2.1 Introduction

How do we see? Detectors are the eyes of our microscopes. From film to sensitive electronics, detectors enable the dynamic acquisition of data. Leading detector designs are combining imaging algorithms with domain specific computing architectures to intake the increasingly informative data. Yet, how are the electrons seen in the first place—and can we do better? Detectors have their limitations. The non-linearity of film, paired with abysmally slow acquisition of developing frames for data analysis, make it non-ideal for emerging applications. Imaging plates\cite{20,21} and charged coupled device (CCD) cameras suffer blurring and other noise problems due to photon propagation.

Direct electron detectors (DED) enable data acquisition with the speed of CCDs, at the dynamic range of film. For the success of 4D-STEM methodologies, converging on a fast, dynamic, and low-noise detector is essential. In this chapter I discuss optimal diode designs for DED. To begin with I discuss existing detector architectures and diode designs. This sets the stage for future designs and possible modifications to existing methodologies.

2.1.1 Direct Electron Detectors (DED)

Two primary semiconductor devices are currently used for DED architectures in electron microscopy: monolithic active pixel sensors (MAPS) and pixel ar-
ray detectors (PAD). MAPS have a complementary metal oxide semiconductor (CMOS) architectures making them very amenable to fabrication, thanks to the similarities with current computer architectures.\footnote{With its beginnings in synchrotron light sources, MAPS are since developed to have better DQE and MTF through diode backthinning. MAPS detectors such as a 4D detector at Berkeley Lab\footnote{and the Gatan K2 have a diode depth around 50 $\mu$m. Images of currently available detectors are shown in Figure 2.1.}}

![Figure 2.1: Three different direct electron detectors currently available for use. The (a) Timepix3 created by CERN\cite{18}, (b) MAPS detector at Lawrence Berkeley National Laboratory\cite{8} and (c) EMPAD at Cornell.\cite{22}]

PAD have a back-bonded technology illustrated in Figure 2.2 with 80 $\mu$m solder bumps. Several designs are currently available, including EMPAD\cite{22} and CERNs Medipix\cite{14} / Timepix\cite{18} / Velopix\cite{23}. These detectors typically have a diode depth of 300 to 700 $\mu$m,\footnote{limited by the available thickness of pure crystalline semiconductor wafers. Deeper diodes yield higher dynamic range, and tolerance to high beam currents. The EMPAD for example has a dynamic range of a million to one electrons at 200 keV,\footnote{ad a diode depth of 500 $\mu$m.}}

Both PAD and MAPS detector architectures have diode layers to capture the electrons incident on their surfaces, as shown in Figure 2.2. While readout
Figure 2.2: Direct electron detector architectures. A pixel array detector (PAD) is shown on the left, with a monolithic active pixel sensor (MAPS) shown on the right. Both detectors have a diode layer, each of different thickness. The diode layer of the PAD is dominated by a depletion zone above the PN diode, with doping on the lower layer adjacent to the solder bump, which is typically about 80 µm. A thin ohmic contact via deposited aluminum and implant doping is on the top layer of the PAD. Readout electronics for the PAD are located underneath the solder bumps. For MAPS readout electronics are located on top of the diode layer.

Regardless of detector architecture, the diode remains a key component in signal quality. Pixel width also has a significant role, yet point spread function, not pixel size, determines detector resolution. Capturing the point spread function accurately is the diode’s key objective.

Shallow diodes minimize multiple scattering effects. In the ideal shallow case, an electron enters the detector diode, and generates electron hole pairs that reach readout electronics. By this, there is minimal measurement error in where the electron first hit the detector. Scattering is often seen as a source of
noise, yet must not be neglected irrelevant as it is also a source of information. A completely arbitrary amount of energetic information is deposited into shallow pixels, as the electron escapes out the back before dissipating all of its energy. As a result, integration mode in MAPS performs poorly in comparison to counting modes.\footnote{13}

Scattered electrons inelastically deposit their energy within the diode, following a Landau distribution.\footnote{11} For higher electron counts, this distribution follows the form of a Lyapunov function, as shown in figure \ref{fig:landau_distribution}. The diverging standard deviation of a Landau distribution does not tend towards Gaussian, as the classical central limit theorem is insufficient. The tail of intensity distribution from electron events is quite large. This is fundamentally why counting modes work better than charge integration.\footnote{12}

Deeper diodes ensure that the full electron energy is captured, enabling a way to count multiple impact events at much higher voltages. PAD designs thrive on the concept of not cutting out any scattering—including lateral and backscattering. A minimal quantity of charge backscatters out of the detector, as in any planar geometry. Completely captured charges allow for full signal integration. Charging issues also become less of a problem in deep diode geometries.\footnote{13}

\subsection{The electron wanders}

By Polya’s theorem the electron is guaranteed to intersect its own path in the diode if given enough time to wander.\footnote{19} This guarantee is inhibited by the constant energy loss due to inelastic scattering, and the diode depth. Shallow diode
Figure 2.3: Histogram of pixel intensity for many electrons incident upon a PAD detector with 500 µm diode depth. From left to right the peaks include the zero-loss peak, a peak for one electron, and peaks for incrementally higher electron counts. This energetic information assists in differentiating coincident electrons.

Depths cut the path short, before the electron has time to go astray. For sufficiently deep diodes, all inelastic scattering is captured giving useful information on electron energy loss throughout the detector.

The validity of certain statistical assumptions is highly dependent on the type of data gathered. Note how MAPS and PAD have deviated from the middle ground of say, around a 200 µm thick diode. This may be traced back to the electrons tendency to wander within the diode. To understand this wander, we
turn to the theory behind drunken random walks. The contents of this section are based off readings in Random Walks and Electrical Networks, by Peter G. Doyle and J. Laurie Snell.6

Particles traversing materials experience random motion if unguided by an external force. Many other non-physical processes also experience random phenomena. Distilling this random motion into a mathematical description is an ancient problem, still extended today and in this work. In this section we show (a) electrical systems are a good analogy for random walks, and (b) Polya’s theorem is powerful.

2.2.1 Lattices and Electrical circuitry

A $d$-dimensional lattice $\mathbb{Z}^d$ has vertices $\{x_i\}_{i=1}^d \in \mathbb{R}^d$ spaced at integer coordinates. Each vertex is connected to its 2D nearest neighbors by an undirected line segment: the edges of our graph. These edges run parallel to their shared dimension, and orthogonal to other dimensions in the same space. By a more general definition, coordinates may be spaced by a lattice constant $a$. Lattices for $d = 1, 2, 3$ are shown in Figure 2.4.

Lattices have many applications across the sciences, as they establish a grid defining space. For example, Bravais lattices have translational symmetry by definition, and are used to describe crystal structures. For our purposes, the lattice and its edges serve as a pathway for a drunken particle to randomly wander.

In electrical analogy, we may construct a formalism where each edge is
equivalent to a 1 Ω resistor. We apply a 1 V potential across the network of resistor edges, and perform basic circuit analysis to come to the same conclusions as probabilistic models arrive at.

**1D closed case**

Take a closed lattice, with dimensionality $d = 1$. Vertices expand along $x = 0, 1, 2, \ldots, N$. Let us consider a drunken particle Fred wandering between home at $x = N$, and the bar at $x = 0$. If Fred ends up at the bar or at home, he stays there. Letting $N = 5$ we have the illustration below, with its electrical equivalent.

Starting from any vertex $x_i$, the probability of our little drunk particle choosing a particular edge to amble is $1/2d$, defining a simple random walk. For our 1D example, this is simply $1/2$. Each edge in the graph is a block on a street. At each block Fred has the opportunity to continue forwards with probability $1/2$, or turn back with probability $1/2$. Starting from any vertex $x$, this amounts to
probability $p(x) = x/5$ of Fred reaching home before the bar. This amounts to a more general definition for $N$,

1. $p(0) = 0$
2. $p(N) = 1$
3. $p(x) = \frac{p(x-1) + p(x+1)}{2}$

where $0, N$ are our boundary coordinates, and $x$ represents the interior coordinates $x = 1, 2, \ldots, N - 1$.

Here we prove that this is identical to the voltage of each resistor in Figure 2.5. Given our voltages indicated as $v(0) = 0$ and $v(N) = 1$ on our boundary vertices, we may use simple circuit analysis to obtain the voltages at the interior points.

The voltage across a resistor is related to the current flow and resistance by
Ohms Law, $V = IR$. Here we designate the current from vertex $m$ to $n$ as $I = i_{mn}$ and the voltage $V = v(m) - v(n)$. Current flowing into a 1D vertex must be equivalent to the current flowing out, by Kirchhoffs Laws. So, for a current flowing from $m$ to $n$ to $o$ we have the following relation.

\[
0 = i_{mn} - i_{no}
\]

\[
0 = \frac{v(m) - v(n)}{R} - \frac{v(n) - v(o)}{R}
\]

\[
0 = \frac{v(x - 1) - v(x)}{R} - \frac{v(x) - v(x + 1)}{R}
\]

\[
v(x) = \frac{v(x - 1) + v(x + 1)}{2}
\]

where we allow $x = n$ in our final form. Here we have an equivalent to our previous conditions for interior coordinates, confirming that voltage and probability at any vertex is identical.

**2D closed case**

To defined a closed lattice in higher dimensions, we turn to a spherical representation. Allow $r \in \mathbb{Z}$ to be an integer valued radius of our $d$-dimensional sphere, $S^r$. The paths within the boundaries of the sphere $G^r$ embedded in our lattice $\mathbb{Z}^d$ exclude vertices further than $r$ from the origin. Here distances are travelled purely on edges, making a sphere not necessarily round as distances in Euclidean space would prescribe. Figure 2.6(a) shows this sphere is much like a square, while Figure 2.6(b) shows its electrical equivalent.
Escape probability

The electrical formulation of these problems bring forward a very effective measure, the effective resistance $R_{EFF}$. This allows us to consider the resistance across multiple edges, and surmise the system’s properties. Following Ohm’s law, we designate the effective resistance between points $a$ and $b$ as the voltage between the two, divided by the corresponding current. An electrical engineering circuits course may cover this property in extensive depth, yet here we follow a simplified model.

Considering a 1 V potential applied between points $a$ and $b$, we have

$$R_{EFF} = \frac{V_{ab}}{i_{ab}} = 1/C_{EFF}$$

where $C_{EFF} = i_{ab}/1$ V is the effective conductance. We may directly relate this to the escape probability $p_{escape}$, the probability of our particle Fred reaching point
b before returning to point a where it began.

Now that we have shown that electrical formalisms are analogous to graph models, we shall put them aside to explore classical proofs of random walks on lattices.

2.2.2 Polya’s Theorem

To move beyond a two dimensional lattice with \( d = 2 \), we introduce the concepts of recurrence and transience. Considering the infinite unbounded lattice \( \mathbb{Z}^d \) of dimension \( d \) we call a walk along the lattice recurrent if it returns to its home starting point as it wanders. If there exists a chance that the particle may never return home, the walk is transient. The escape probability \( p_{\text{escape}} \) is tightly related to the dimensionality of our lattice. Polya’s theorem states,

**Theorem.** A simple random walk is recurrent for \( d = 1, 2 \), and transient for \( d > 2 \).

\[
\begin{align*}
\text{RECURRENT if } & p_{\text{escape}} = 0 \\
\text{TRANSIENT if } & p_{\text{escape}} > 0
\end{align*}
\]

Why is this? It is remarkable to think that wandering the 1D or 2D blocks of a city is guaranteed to bring you back to your starting point. We assign this probability of returning to the starting point as \( u = p_{\text{return}} \). How many times may we return to the origin throughout our random walk? The expected number of times \( k \) at the origin, including the original, is 
\[
m = \sum_{k=1}^{\infty} ku^{k-1}(1 - u) = \frac{1}{1 - u}.
\]
Recurrence is achieved when \( u = 1 - p_{\text{escape}} = 1 \), sending \( m = \infty \). Transience, is correspondingly \( u < 1 \) and \( m < \infty \). Now we may use \( m \) and \( u \) as a determination factor much like \( p_{\text{escape}} \).
When will we return? Designating the probability of being at the starting point on the $n^{th}$ step as $u_n$ it follows that we start with $u_0 = 1$. Here we have integer valued steps $n \geq \infty$. Summing these probabilities we return to our definition on the expected number of times at the origin, $m = \sum_{n=0}^{\infty} u_n$ giving us a way to calculate $m$ based on probabilities at individual steps.

**1D infinite case**

For the infinite 1D case we maintain that the probability of going forward on a block or returning backwards is $1/2$. If we start at a point, the particle absolutely must return on the same path travelled as our random walk is confined to one dimension. This brings us to the conclusion that only even numbers of steps $2n$ bring us home, with probability $1/2^{2n}$. We can choose $n$ times to go return from the $2n$ possible times, giving us the probability $u_{2n} = \left(\frac{2n}{n}\right) \frac{1}{2^{2n}}$. Summing this over all steps $n$ we approach infinity, confirming that a 1D simple random walk is recurrent.

**Independent 1D random walks**

Interestingly, a simple random walk in two dimensions is equivalent to two independent one dimensional walks. Therefore, $u_{2n}$ for the 2D case is equivalent to the 1D case squared. This is not the case for $d > 2$. This ties to a significant relationship,

**Lemma.** If two random walks are about the same, they are either both transient or both recurrent.
2.2.3 Random walk in 3D

Instead of mathematically defining a random walk in three dimensions, we turn to an intuitive physical system: diffusion in a crystal lattice. The aforementioned Bravais lattice systems are used to describe crystals, allowing translational symmetry.

Confining the definition of a Bravais lattice to our definition of a $\mathbb{Z}^d$ lattice, we use a cubic crystal structure. This adapts our edges to have the same properties as the $d = 3$ case. By exact calculations, we may determine the probability of return to the origin $u$. For $d = 1, 2$ we had $u = 1$ due to the recurrent nature by Polya’s theorem.

Figure 2.7: Simple random walk in three dimensions.
Giving our lattice cartesian coordinates \(a, b, c\) we may indicate the probability \(p(a, b, c; n)\) of a random walker starting at the origin, is \((a, b, c)\) after \(n\) steps. In this case we may think of the walker as a vacancy in a 1-atom basis simple cubic crystal.

At the start, we have \(p(0, 0, 0; 0) = 1\). Thereafter,

\[
p(a, b, c; n) = \frac{1}{6} p(a - 1, b, c; n - 1) + \frac{1}{6} p(a + 1, b, c; n - 1) + \frac{1}{6} p(a, b - 1, c; n - 1) + \frac{1}{6} p(a, b + 1, c; n - 1) + \frac{1}{6} p(a, b, c - 1; n - 1) + \frac{1}{6} p(a, b, c + 1; n - 1).
\]

this derives down to

\[
p(a, b, c; n) = \frac{1}{(2\pi)^3} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \left( \frac{\cos x + \cos y + \cos z}{3} \right)^3 \cos (xa) \cos (yb) \cos (zc) \; dx \; dy \; dz
\]

which we then sum over \(n\) to obtain our expected number of returns,

\[
m = \frac{3}{(2\pi)^3} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \left( \frac{1}{3 - (\cos x + \cos y + \cos z)} \right)^3 \; dx \; dy \; dz
\]

By substituting the above with gamma functions we eventually arrive at

\[
m = 1.516386... = 1/(1 - u)
\]

\[
u = 0.340573... < 1
\]
This means, that we have roughly a 34% chance of returning to the same starting point upon wandering a 3D infinite lattice $Z^3$. For higher dimensionalities this percentage decays. It becomes much less likely our transient particle will find its way home after a random walk.

The formalism of random walks covered here are simply a beginning to understanding complex systems. Diffusion in complex crystal systems is a first step, allowing for the computational characterization of many materials and transport properties. Our random walk may also be damped: that is, the particle looses energy or speed as it wanders—eventually coming to a stop. Due to inelastic scattering electrons certainly do not have an infinite random walk throughout crystal lattices. This mathematical formalism of an ideal case helps frame electron scattering with its relativistic corrections.

### 2.3 Numerical Experiments

As the electron impacts the surface at a position $(x_o, y_o, 0)$ it begins its random walk throughout the diode material. At each point $i$ in its trajectory the position $(x_i, y_i, z_i)$, energy $E_i$, and direction $d_i$ is tracked. As it travels a distance relative to its mean free path, the electron moves unperturbed. At the end of this traveling step the electron scatters inelastically, depositing an energy $\partial E$ at that position. Eventually the electron looses enough energy that it can no longer move on; at the end of its journey when $E < 0.5\text{eV}$ its final position is saved.

The lengths of many paths accumulate into a point spread function, with characteristics unique at each energy. The point spread function captured by the detector has the highest impact on detector resolution. A spread of energies
was applied each diode material. Starting the journey with a higher energy, generally gives the electron a longer path. Full electron paths are generated, then cut off as they hit different diode depths. The cross-sectioned \((y, z)\) and top-down \((x, y)\) views are shown for a thousand electron incident at the origin indicated by a red dot in Figure 2.8.

Simulations of electron paths, to determine the extents of performance across a wide parameter range of diode architectures, are executed in this study. An electron Monte Carlo method by David Joy with relativistic corrections implemented by Peter Ercius generates electron paths. I implemented further modifications for electron paths to be rapidly simulated in detector diodes. These are used to test possible DED designs, and train electron counting algorithms discussed in the next chapter.

![Monte Carlo simulations of electron paths](image)

Figure 2.8: Monte Carlo simulations of electron paths (only 1000 shown) at varying beam energies. With higher energy, the electron has greater capability to travel far from where it started. The top row contains a projected side view of the electron paths incident on the surface, while the bottom row is a plan view projection. Red dots indicate electron entry points.
2.3.1 Diode Material

Materials used in the diode layer are the first concern when initiating simulations. Where is the electron wandering? Going back to our analogy of a drunken random walk: an electron wandering a material, is akin to a drunkard wandering a terrain. The material, or terrain, the electron wanders in is of high importance in considering how information is captured. Will the electron destroy everything in its path? Is the terrain so resiliently rough that the electron never reaches readout electronics? These are items to consider when choosing a DED diode material.

The choice of diode material marks a detector's lifetime in its microscope milieu. A requirement of radiation hardness ensures electronic components resist damage from the bright beam. Electrons regularly irradiating a detector pose many problems for lifetime and performance. Many materials are very well characterized for this.\cite{17} Current DED available use silicon in the diode layer, due to the accessibility of high purity wafers for fabrication.\cite{10} Beyond radiation hardness, material choice has a high impact on detector performance as measured by DQE and MTF. In this study we investigate silicon and germanium. We find that germanium has a higher performance as a diode material for DED, as shown in Figure\ref{fig:performance}.

2.4 Virtual electron detection

To test the effectiveness of different electron counters across beam energies and DED architectures, we feed the electron counter virtual detector readouts. Three
Figure 2.9: Error with respect to energy comparing silicon performance (black line) with germanium (red line). Germanium is shown to be a superior diode material through our simulations.

electron counters are tested: the maximum intensity (peak) pixel; the center of mass (mean) point; and a convolutional neural network with a rectified linear unit (ReLU). We simulated primary beam energies from 30 to 5,000 keV for silicon and germanium diodes, with pixel pitches from 1 to 500 \( \mu m \) and diode depths from 10 to 1,000 \( \mu m \). The root mean squared error between the true entry point and the counters guessed entry point is used as a metric of performance. The simulations are performed, assuming a perfect signal with no additional noise, to test the maximum capability of counter performance.

To form test and training datasets for all counters, paths are generated with independently and identically distributed random origins. Simulated paths are
Figure 2.10: A single electron path with its overlaid detector readout, at 300 keV in a 500 µm deep diode with 50 µm pitch. The summation of all pixels is the energy of the incident electron, for sufficient diode depth. Different counting metrics for this particular path are labeled.

cut off at designated diode depths, then binned into a sparse matrix weighted by the energy lost at each coordinate. Each bin is a pixel of corresponding pitch. Indices of starting point are stored as the y-label, while detector readouts are stored as the x-label. This generates a dataset with randomized starting points and random paths, but the same image size.

The 3D electron paths projected onto a 2D detector readout are the conceptual equivalent of histograms with a bin size equivalent to the pixel size. Each \((x, y, z)\) point along the electron’s trajectory is put into its corresponding histogram bin with a weight \(\partial E\), the energy deposited at that point. The total sum of bars in a path’s histogram is equivalent to the energy at which the electron impacted the detector with. Precise pixel and sub-pixel entry points are randomized input parameters. Each electron path is projected onto the equivalent of a detector pixel readout, for a spread of pixel widths between 1 and 150 µm.
Figure 2.11: A sampling of electron path histograms with varying pixel pitch. Most interestingly, the peak and mean pixels are a significant distance from the true entry point. Each row has the scale bar value labeled.

This projection onto 2D follows Polya’s theorem in the sense of frequent path self-intersection. Additionally, laterally scattering paths impose much error, as the electron also has a tendency to deposit the majority of its energy at the end of its trajectory. The overall distribution of energies deposited adjacent to the electron’s entry point is seen in the point spread function shown in Figure 2.8. The plan and cross sectional view of our point spread function elucidates the behavior of different diode geometries. A narrow distribution at the neck of the electron’s entry point does indeed reduce lateral and backscattering in the
diode, and is an ideal geometry for low dose imaging, were few electrons are likely to impact at the same point.

Figure 2.12: A single electron path, cut off at three different diode depths. On the far left, a three dimensional bar plot is shown to display intensity differences in each pixel. The following plots are the same information projected onto a two dimensional representation at a 500, 300, and 50 µm diode depth respectively.

This spread is perhaps most easily characterized by the point spread function (PSF). For relatively low electron energies, the PSF follows a Gaussian distribution. The energy distribution at energies larger than 60 keV becomes inhomogeneous, deviating from a Gaussian point spread function. This is furthermore evident as one considers single electron paths. In Figure 2.11 we have a random sampling of paths at a depth of 500 µm at 300 keV where the electron paths end, only when its energy is fully dissipated by inelastic scattering.
2.5 Results and Discussion

Histograms of paths at depths of 50, 300, and 500 µm are shown in Figure 2.12. Here we see the difference in peak accuracy, which is more quantitatively illustrated in Figure 2.14 where RMSE vs. depth is shown. Our findings indicate that deep diodes with short pixel widths are the ideal geometry for DED. The diode must be sufficiently deep to capture the full electron path. Shorter pixel strides simply ensures a higher resolution image of the path to work with, yet comes at the cost of expensive manufacturing and excessive amounts of data.

At high beam energy, performance is dependent on pixel pitch. This is due to the electron barreling straight through the diode, and nearly equivalent to having a shallow diode. One may picture this as a lower quality image of the electron path. Depths on the other hand, have an interesting trend of RMSE capping out at around a depth of 400 µm. As shown in the plots below, electron energy is the dominating feature affecting RMSE. This is closely followed by pixel width, and then by diode depth.

There dependence on pixel width, is expected as this is the modulus by which our electron path is cut off. Essentially the smallest pixel pitch possible is best, yet this is limited by semiconductor fabrication techniques. For CMOS based detectors smaller discrete pixels are attainable due to very advanced fabrication techniques aligning with current commercial technologies. PAD based designs have the advantage of a continuous detector layer, yet the size of solder bumps remains a limitation.

Imaging at lower energies, depends heavily on diode depth. Deeper diodes yield a high dynamic range, and tolerance to high beam currents. Shallow, or
Figure 2.13: Error with respect to energy for three different detector geometries shown in Figure 2.1. Pixel pitch is shown as a thin horizontal line, in each plot labeled by a diode depth. Super-resolution regimes are shown, as the regions where the error is less than the pixel pitch. The peak pixel metric is plotted as a solid line, while the center of mass is plotted as a dotted line. In general, center of mass has better performance than the peak pixel.
Figure 2.14: Root mean squared error for peak (red) and mean (black) pixel as pixel width and depth varies, given perfect image segmentation and no readout noise.

backthinned, diodes reduce coincident events by cutting electron paths short. Coincident events may also be detangled in a deeper diode—as the full energy of the incident electron is captured. Pixel pitch is limited by DED architecture—yet the point spread function, not pixel size, determines detector resolution.

Super-resolution regimes are subsequently affected by these differences. With depths a cutoff at 50, 300, and 500 µm, a maximum super-resolution beam energy of 100, 200, and 300 keV is respectively observed in Figure 2.13. The RMSE of detectors certainly drops for shallow diode depths, as the electron is given no chance to wander far. A depth of 500 µm captures the full electrons wander, yet this inherently implies that COM and MIP RMSE is worse.

Peak pixel electron counting has the highest error, followed by center of mass and machine learning, further discussed in Chapter 3. All geometries follow the same initial log-log relationship for geometries below their super-resolution regime. Back-thinning is an effective method given low doses of electrons, and a high frame rate detector. Exact electron counting gives an explicitly linear
detector readout. There are two ways to achieve this: through low dose < 50\mu m diode, or through energy accumulation in a < 400\mu m diode.

![Image of performance dependencies on pixel pitch, diode depth, and incident electron energy.](image)

Figure 2.15: Performance dependencies on pixel pitch, diode depth, and incident electron energy. Normalized root mean squared error for the mean point of a detected path, given imaging conditions, is shown to illustrate trends. These slices of the data cue show key dependency on depth, followed by energy.

Similar trends for all counters are observed with respect to beam energy, pixel pitch, and diode depth, seen in Figure 2.15. The electron counters are generally most dependent on diode depth for performance, followed by beam energy. There are three regions of dependencies on diode depth: a barreling region for backthinned diodes where the electron passes straight through; a peak error region where the electron path is cut off but given sufficient time to wander; and a region where the full point spread function is captured.

Rapid electron counting affects the quality and usability of data. By sparsifying data with sub-pixel super-resolution, data is refined into a higher quality. By compressing detected frames by orders of magnitude, data becomes more usable. Currently, data rates are increasingly inhibitive to analysis. Automating live electron counting as a default detector feature is a computational intense ordeal, yet one that makes NP-complete algorithms solving the problems of to-
mography and phase retrieval much, much easier.

Further electron counting developments must tackle the task of image segmentation, in addition to super-resolution. Image segmentation proves increasingly difficult with higher doses, as multiple coincident electron events are common. Faster frame rates assist in reducing coincident electron events. Quantifying incident electron energy leaves promise for detangling coincident events. Deeper diodes capture the full point spread function, measuring the complete energy of each electron. Shallow diodes capture a completely arbitrary amount of energy per electron. By incorporating the energy information provided by a deeper diode, precise differentiation between the number of electrons incident on a region is possible.
CHAPTER 3
SUB-PIXEL SUPER-RESOLUTION ELECTRON COUNTING

To test the effectiveness of different electron counters across beam energies and DED architectures, we feed the electron counter virtual detector readouts. Three electron counters are tested: the maximum intensity (peak) pixel; the center of mass (mean) point; and a convolutional neural network with a rectified linear unit (ReLU). To form test and training datasets for all counters, paths are generated with independently and identically distributed random origins.

Simulated paths are cut off at designated diode depths, then binned into a sparse matrix weighted by the energy lost at each coordinate. Each bin is a pixel of corresponding pitch. This generates a dataset with randomized starting points and random paths, but the same image size. Indices of electron entry points are stored as the y-label, while detector readouts are stored as the x-label. The root mean squared error between the true entry point and the guessed entry point is used as a metric of performance.

Low energy incident electrons have Gaussian point spread functions. For higher beam energies common in electron microscopy the point spread function deviates from Gaussian. A reduced point spread function is attainable with the full energy deposited into a pixel, up to a dose level relative to the pixel’s tolerated energy. For example, the EMPAD II tolerates 2 nA of beam current per pixel at 300 keV, which is quite sufficient for counting big numbers of electrons. Not only does this sparsify data, it reduces the point spread function into a single point—optimizing resolution to its peak.

The accuracy of electron counting methods is evaluated in this chapter. New
machine learning based methods are explored as well. To get a precise count of electrons first I dive into the theory behind electron journeys across pixels. For this end, an in depth discussion of random walks is provided in the first section. The following section applies random walks to training data and test data, for a variety of statistical models.

3.1 Electron Counting

Collecting each electron to hit the detector with high accuracy is a task affecting every bit in terabytes of data. As an electron hits the detector it travels through a random path—often hitting multiple pixels. The pixel of highest intensity is commonly not where the electron came in, yet this is a baseline assumption by many existing counting techniques. Given the massive amounts of data, it is impractical to have a human count every electron. Instead, detectors are hooked up to high performance computing systems ranging from workstations to supercomputers with data rates up to 400 Gbps. Our goal here is to develop a rapid electron counting algorithm for these systems.

Knowing the electron’s entry point is an extremely effective way to compress data by orders of magnitude. Detectors store a full image frame at every collection interval; this can amount to up to 200 TB of data per hour. The only relevant data is the origin pixel. If we know the origin pixel, we can throw out the rest of the data. This compresses data by orders of magnitude. Predicting the origin pixel would serve to significantly compress the amount of data stored by reducing noise data from electron propagation.

The pixel of interest is only where the electron entered because that only
reflects the physics of the sample and not the detector itself. One naive way to predict where the electron entered is to simply see where the electron left the most energy: the peak pixel. Another possible method is to measure the center of mass of the electron path, the mean pixel, weighted by the energy it leaves at each pixel. Neither are completely accurate ways to determine the origin pixel.

3.1.1 A simple electron counter

For a MAPS detector installed at Lawrence Berkeley National Laboratory, I implemented a simple electron counter. Here I review this implementation to outline the general task at hand. Electron counting aims to minimize the entirety of an electron path’s shadow in the detector diode, to a single point corresponding to where it entered.

A simple way to determine where the electron entered is by thresholding the data and finding the peak pixel within a cluster. This assumes no coincident electrons, which is a perfectly valid assumption under low dose and high frame rates. Data was simulated through PRISM onto a virtual MAPS detector. Experimental data was acquired on the detector itself. This counting method is routinely used on both.

Given the massive datasets, a single 4D-STEM image was split into chunks of information consisting of a header, then image blocks. Each header contained metadata pertaining to the subsequent blocks. Images are appended after each other in each block, yet not sequentially. This is done to expedite data offloading speed, as images do not have to wait for their predecessor before leaving the detector hardware themselves. Post-processing re-orders images.
Electron counting is embarrassingly parallel as each diffraction $P$ in the full four-dimensional 4D-STEM image $I$ may be independently processed. After an image is loaded, each pixel is thresholded. The threshold $t = \mu(P) + \eta \sigma(P)$ may be set by taking a random sampling of images and choosing $\eta$; for our purposes $\eta$ was set to 8. The mean $\mu(P)$ and standard deviation $\sigma(P)$ may be precomputed to have a threshold predetermined.

Applying this threshold compresses data by two orders of magnitude, while improving image quality. This method does not reduce an electron entry point to a single point, or even a single pixel, yet still manages to accomplish the task of electron counting. Additionally this method is extremely fast, only limited by data transfer rates. Later iterations of this simple thresholding method includes denoising and local peak finding.

### 3.1.2 Can a machine count?

As an initial proof-of-concept, a simple three-layered convolutional neural network (convnet) was constructed to determine electron entry points based on a single electron path detected. To form training and test datasets, several paths are generated with random origins $(x_0, y_0)$ and made into a histogram $h(x, y)$. For all paths, we will use the same binning. Additionally, the index of where the starting point $(x_0, y_0)$ is binned is stored as the $y$-label. This generates a dataset with various starting points and random paths, but with the same image size.

We use simulated detector data and teach a machine to count electrons accurately. Pre-existing Monte Carlo simulations of these detectors generate electron paths. With this, we make massive training and test datasets. We take simulated
electron path data in a detector with known points of origin. This is then fed into our neural network, along with the true label of the electrons incoming point. Here we implement machine learning to correctly count electrons entering detectors.

For the machine learning implementation, we use the Keras python package along with the default Tensorflow backend. Keras is a deep learning package with relatively easy implementation. Most importantly, it boasts minimum delay in getting results, enabling faster feedback and improvement. Keras also has several example implementations on the CIFAR-10 dataset, showing its capability for image processing and recognition, making it a great fit for our project.

For our neural network, we use a 2D convolutional layer, followed by a rectified linear unit activation (ReLU), followed by two more 2D convolutional layers. To pull out the final index label, we then add a flatten layer and a dense layer with a two dimensional output. The final index predicted is then compared to the true index with a mean square error loss. For our optimizer, we use an Adadelta optimizer. This optimizer was chosen for its adaptive learning rate, which requires minimum adjustment on our part.

Originally, we measured our performance by simply measuring the pixel distance between predicted and true labels. Prior experiments measuring mean and peak pixels used root mean square error, so we adapted that metric when comparing to previous results. A more sophisticated performance metric may take into account the possibility of choosing a pixel with an intensity of zero, where the electron was never present. While useful for training, this poses difficulties in comparing directly to existing methods.
Paths $p(x,y,z)$ of electrons travelling a silicon detector are generated via Monte Carlo simulations implemented in python scripts. These paths may be cut to correspond with various detector geometries. The histograms $h(x,y)$ generated by these paths correspond to an image frame read off by any one of these detectors.

Resulting simulations of electron paths may be understood as a random walk of a drunk electron spilling energy as it meanders. After travelling a certain distance, the electron scatters at a random angle. Energy is continuously lost at a fixed rate along its path. Anecdotally, the further the electron goes, the drunker it gets, and the more it stumbles.

At the start, the electron tends to travel longer distances in a straight path as it has higher energy. After losing energy, the electron does not propagate very far, and may illuminate a pixel repeatedly. As the electron randomly walks, it frequently intersects its own path. This causes some pixels to activate multiple times.

Figure 3.2 shows our preliminary results measuring the absolute pixel distance error as we increase the training data size and compare to performance on test data with 500 samples. Each sample was a 39x39 histogram with 10 $\mu$m sized bins. The origin was chosen at random from -40 to +40 $\mu$m, an 8x8 grid of possible starting points. The training was done for 60 epochs.

We see that the neural network is capable of reaching subpixel accuracy within only 500 training samples. Performance is inconsistent both within a certain data size as shown by the large standard deviation as well as within changing data sizes. Losses measured across epochs also show this inconsistency.
where losses measured on the training data do not consistently decrease and sometimes diverge dramatically. One example of loss across epochs is shown in Figure 3.3. However, initial results are promising in that they show the feasibility of detecting the starting point through machine learning.
Figure 3.2: Test and training mean error and standard deviation across different training sample sizes. Test data size was fixed at 500 samples. Performance already reaches subpixel accuracy at 500 training samples, but has inconsistent performance at larger data sets.

Visualizations of the neural network shown in Figure 3.4 shows that the neural network appears to be learning the shape of the electron path. One possible interpretation is that the neural network is essentially memorizing possible shapes of electron paths with their associated origin to perform its predictions. Of concern are the outputs of completely zero arrays. This usually leads to a (0, 0) prediction, and we notice that the neural network can occasionally be stuck at this guess across epochs. In order to address this, we chose to load only the weights from the best performing epoch when performing predictions.
To perform a more realistic simulation, we generated data according to real existing detector geometries. We used three detector geometries: the EMPAD with 150 $\mu$m width, 500 $\mu$m depth; the Medipix with 50 $\mu$m width, 300 $\mu$m depth, and the MAPS with 10 $\mu$m width, 50 $\mu$m depth. Paths are simulated for these geometries at energies of 30, 60, 80, 120, 200, 300, 500, 1000, 2000, 3000, 4000, and 5000 keV. Our training data set has 5000 samples and the test set has 1000 samples. Root mean square pixel error is measured and scaled with the pixel size to get the real distance error.
Figure 3.4: 2D histogram $h(x,y)$ of an electron’s path through the detector and visualization of its 3 convolutional layers. Twelve (out of 32) channels are shown here. The electron’s entry point at the origin is indicated by a white square. Red square indicates prediction pixel from the neural network. This histogram is one of the test cases for the detector geometry of Medixpix of a 300 keV electron.

The first set of results are shown on Figure 3.5, later developed into information shown in Figure 3.7. Here the simulations with 500 µm depth geometry are outperformed by prior simulations, which were done with 5 µm bin sizes instead of mimicking the detector geometry. Pixel width has a very large effect on performance, as shown in Chapter 2, and is the dominating cause of error in this comparison. The neural network is generally accurate to one or two pixels, which when scaled by the large pixel size of 150 µm leads to a large error. Geometries with smaller pixel sizes like the Medipix and especially the MAPS were very competitive with prior results.

Since we saw that the micron distance error of the neural network improved with smaller pixels, we performed another simulations but using an 80x80 histogram with a fixed 5 µm bin size and a -25 to 25 µm starting range for $(x_0,y_0)$. The original pixel depths for the three detectors are still used: 50, 300, 500 µm. The same energies from the previous simulation were also used. This used a training data size of 5000 and tested with a data size of 1000. This was also run through 60 epochs and the RMSE performance was compared to measuring the
Figure 3.5: Initial results for root mean squared error (RMSE) plot comparing MIP, COM, and our neural network. Plots for the EMPAD are shown in red, Medpix in blue, D4MAPS in black, and a new architecture in yellow. Datasets for the ConvNets are not the same as the data used to calculate MIP and COM, making this an indirect comparison. Given previous results, double the number of data samples are needed to improve consistency for the ConvNet. Trends indicate that the ConvNet provides better resolution than MIP and COM outside of the super-resolution regime.

starting pixel using MIP and COM. Instead of pulling from previous literature, we directly measured the MIP and COM from the test set. The results are shown in Figure 3.6.

At 50 µm depth, the neural network only outperforms around 100-300 keV. At larger depths, it outperforms COM and MIP at a wider range of energies. In general, the neural network outperforms COM and MIP when the electron path is long. When the path is only one or two pixels, the COM and MIP have very low errors comparable or better than the neural network. Short paths occur when either the depth is very short and terminates the electron path, or the energy is too low or too high. Error of the neural network is fairly consistent
Figure 3.6: Simulations using 5 µm bins at 50, 300, and 500 µm diode depths. Error of the neural network versus COM and MIP are plotted for each depth. As depth increases, the COM and MIP gains in error due to longer electron propagation whereas the neural network stays fairly consistent. At low and high energies, COM and MIP are still good since the electron path is usually only one pixel.

regardless of path length, so when the COM and MIP error goes up due to longer electron path, the neural network outperforms them.

By adding additional training, the performance of this convnet increased, shown in Figure 3.7. Still, the training datasets are much smaller than a typical 4D-STEM dataset. To further improve performance, a combined metric using a loss function that penalizes for choosing a pixel not in the path may be of great benefit. These results indicate that machine learning is a valid approach for electron counting, and may achieve extreme super-resolution with further development.
Figure 3.7: Error with respect to energy for three different detector geometries. Pixel pitch is shown as a silver horizontal line, in each plot labeled by a diode depth. Super-resolution regimes are shown, as the regions where the error is less than the pixel pitch. The training and test performance of the machine-learning-based electron counter is shown, in comparison to the peak pixel calculated simultaneously.

3.1.3 Image Segmentation

The methodology of generating electron paths randomly as in Chapter 2, is extended to simulate the full scale detector as opposed to subregions. Random
electron entry points \((x_o, y_o)\) are generated with an imposed normal distribution, such that electrons are more likely to enter at the middle of the detector. From each random entry point, a full electron path is simulated and pixelated by binning the energies at each scattering point into a sparse matrix \(K\). This is then added to the full momentum space \(P\) detected virtually.

Each \(P\) is fed in as the training data containing a random number of detected electrons from 1 to \(N\). For the first version, a label associated with each \(P\) was simply the number of electron paths in the image. The number of training images processed is also quite small for a 4D-STEM dataset, at 10,000 images with at most 10 electrons in each. Accuracy saturates at just above 55% in this situation. Before pursuing a more realistic dataset, we look to adjust the labels and loss function. The labels consisting of the number of detected electrons, must be updated to contain a list of electron entry points.

By simplifying models pre-trained on ImageNet, we accomplish image segmentation and counting classification simultaneously. In the first iteration of this model extremely insufficient labels and a simple loss function yield moderate success, even on small training datasets of a thousand. The label used is the number of electrons incident on \(P\), with a loss function of RMSE. This model accomplishes an accuracy better than random guessing. Further expansion on this model by providing accurate labels of exact electron entry points, may be a far better teacher for this convnet. Modifying the loss function of RMSE to exclude zero-valued pixels is a method to improve both machine learning models discussed in this chapter.
3.2 Outlook

In this section we reviewed the theory of a random walk and how it is used to train two convolutional neural networks with the purpose of counting electrons incident on a detector. While these electron counters are still a long way from routinely replacing existing counting metrics, they demonstrate the promise of a different numerical approach. Electron counting via machine learning allows for the potential to incorporate integration mode, accounting for coincident electron events. By developing this technique, we extend the capability of direct electron detectors towards a perfect resolution.

Rapid electron counting affects the quality and usability of data. By sparsifying data with sub-pixel super-resolution, data may be refined into a higher quality. By compressing detected frames by orders of magnitude, data becomes more usable. Currently, data rates are increasingly inhibitive to analysis. Automating live electron counting as a default detector feature is a computational intense ordeal, yet one that makes NP-complete algorithms solving the problems of tomography and phase retrieval much, much easier.

Further electron counting developments must tackle the task of image segmentation, in addition to super-resolution. Image segmentation proves increasingly difficult with higher doses, as multiple coincident electron events are common. Faster frame rates assist in reducing coincident electron events. Quantifying incident electron energy leaves promise for detangling coincident events. Deeper diodes capture the full point spread function, measuring the complete energy of each electron. Shallow diodes capture a completely arbitrary amount of energy per electron. By incorporating the energy information provided by a
deeper diode, precise differentiation between the number of electrons incident on a region is possible.
Electron microscopy is exuding exponentially more data, edging into a computational regime. Aberration correction and electron counting may be realized through direct electron detection and edge computing. Digitized data brings about complex computational imaging problems of high dimensionality. Paired with sample discovery and preparation, there is much to discover ahead.

In particular, DED—which provide a plethora of information in materials systems—are approaching 200 TB/hour. This is made possible by use of a massive embedded system, connecting a direct electron detectors to high performance computing systems. This technology extends to any big data detector today, yet for microscopy purposes a domain specific architecture may be constructed.

4.1 A Domain-Specific Architecture for Electron Microscopes

Domain-specific computing architectures for electron microscopy are only recently emerging. Here we delve into two designs developed throughout the course of this thesis. The first is at Lawrence Berkeley National Laboratory (LBNL), and the second at Cornell University. The two are designed for different direct electron detector architectures, leading to different assumptions of data rates and sparsity.

At LBNL, a monolithic active pixel sensor (MAPS) is connected to NERSC’s newest supercomputer Cori. This is scheduled to update with NERSC 9 system once it goes online. Direct electron detector data collected on the electron
microscope is pre-processed on FPGA boards then sent across a 400 Gbps fiber optic network to NERSCs software defined network, landing on Cori compute nodes. To prioritize bandwidth and data integrity, data on compute nodes are immediately sent to Coris Burst Buffer.

After processing, a portion of the data is immediately returned to the user to guide their experiment in real-time. This simple step would take several months, as opposed to seconds, without the use of high performance computing (HPC). The raw data remaining on Cori is subsequently used for advanced computational imaging. Without HPC the detector may approach a frame rate of 400 fps; with HPC the detector is projected to attain 100,000 fps.

At Cornell, the EMPAD is connected to HPC at Johns Hopkins through the Platform for Accelerated Realization, Analysis, & Discovery of Interface Materials (PARADIM). Data collected on the EMPAD is automatically sent across 10 Gbps ethernet connected to a fiber backbone routing towards Johns Hopkins. As data is sent, a cryptographic hash is generated at both ends to ensure data integrity. A DOI number is automatically generated per dataset as well. Computational resources are accessible through SciServer, with analysis scripts in Jupyter notebooks.

EMPAD frame rates are not nearly as high as the MAPS detector at Berkeley Lab, due to differing detector architecture and application. At Berkeley Lab the detectors are of much higher throughput to attain extremely sparse data for simplified counting modes, applicable to low-dose imaging conditions. The EMPAD’s high dynamic range permits it to image at high doses and use integration mode effectively. Counting modes on the EMPAD are being designed to take into account the energy deposited per event as well, forming in a sense an
integrated counter.

4.2 Conclusion

The tools developed in this thesis are only a beginning to many techniques, by allowing acquisition of usable data of high quality. Usability is simply measured by the ability to efficiently analyze data, whether it be by compressing data volumes or routinely using high performance systems. Quality of data is the amount of valuable information contained in each bit.

Once high quality data is attained imaging modes may be expanded to more algorithmically intense applications. My interest in particular is achieving ptychographic tomography for a wide variety of samples. The four-dimensional 4D-STEM volumes become five- or six-dimensional depending on the acquisition techniques. This amounts to solving multiple NP-complete problems on teravoxels to petavoxels of data. These daunting volumes may be reduced through effective electron counting and reducing the necessary amount of data collected.

By solving this first stepping stone, our problem is now tractable. Collecting less data, of higher quality, is achieved through optimal detector design. Electron counting reduces our massive datasets to a highly compressed forms, reducing our amount of data by several orders of magnitude while improving resolution. These developments further 4D-STEM as a whole, in addition to contributing to ptychotomography.


[22] Mark W. Tate, Prafull Purohit, Darol Chamberlain, Kayla X. Nguyen, Robert Hovden, Celesta S. Chang, Pratiti Deb, Emrah Turgut, John T. Heron, Darrell G. Schlom, Daniel C. Ralph, Gregory D. Fuchs, Katherine S. Shanks, Hugh T. Philipp, David A. Muller, and Sol M. Gruner. High Dy-
