NORTHWESTERN UNIVERSITY

Vision-Based Automation for Accelerated Structural Interpretation in Atomic-Resolution Microscopy

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Materials Science and Engineering

By

Eric Stephen Schwenker

EVANSTON, ILLINOIS

December 2021

ABSTRACT

Vision-Based Automation for Accelerated Structural Interpretation in
Atomic-Resolution Microscopy

Eric Stephen Schwenker

At its core, the purpose of microscopy is to make objects and their underlying structures visible under high magnification. With the remarkable progress of electron microscopy, the sub-micron “high” magnification of light microscopy has been completely refashioned to encompass subatomic length scales. Unfortunately, higher-magnification does little to negate existing interpretability challenges present in images of crystal imperfections– which comprise some of the most scientifically intriguing and technologically relevant materials images. And any sort of direct interpretation advantage that this localized imaging affords, is quickly overwhelmed by the sheer volume of images that must be processed to extrapolate effects to the bulk of the crystal. Fortunately, computer vision has emerged as a tool for automated analysis and interpretation of images from large volumes of complex and/or noisy visual inputs – conditions nearly synonymous with images collected at atomic resolution. This thesis is focused on the development of vision-based automation pipelines with applications specific to structural interpretation of electron microscopy

images. Moreover, it is shown that vision-based automation can be used to harmonize the power and scaling of computation, with the quantitative insights now accessible via experimental imaging, physics-based modeling and simulation, and even peer-reviewed scientific literature.

We begin by exploring methods for quantifying image similarity in atomic-resolution microscopy. Image similarity is an essential consideration in the general interpretation efforts, as the intensity signals from the experimental micrograph often must be compared to simulation to better understand acquisition parameters or validate proposed structures. Second, we focus on the problem of determining 3D atomic structure from experimental STEM and STM images and develop custom automation tools to find candidate structures that are both energetically feasible and produce images consistent with what is observed experimentally. Finally, we highlight the development of a pipeline for constructing self-annotated microscopy datasets from scientific literature. It is our vision that the pipelines developed here will help enable meaningful automation in the structural interpretation of atomic-resolution microscopy images, both as a mechanism for suggesting plausible structures that match experimental observations, and as a first step in translating recorded scientific knowledge from existing images to future images unseen.

Acknowledgements

Here is an attempt to succinctly acknowledge those who have shaped this experience for me. I am immensely grateful to all of you.

First, thank you to my advisors. **Chris**, our discussions on figured bass during undergraduate advising left quite the impression. Thank you for your guidance during my PhD, and for encouraging my outside musical interests. **Maria**, you're not only someone I consider a role model for what it means to be a diligent, curious, and engaged scholar, but also someone I consider a friend. Thank you for this, and for giving me the confidence and flexibility to explore materials science in conjunction with my multifarious interests.

Next, thank you to my colleagues at the **Center for Nanoscale Materials** and in the **Wolverton Group**. The multidisciplinary collaborations that resulted from my ties to both organizations were invaluable to me. In particular, **Fatih**, you were there for day-to-day technical discussions, nudges towards interesting ideas, and general encouragement. Thank you for taking me under your wing in the early years of my PhD. **Weixin** and **Trevor**, our work on the MaterialEyes project was fun, not only because the topic was exciting, but I really feel like we made a great team! Thank you for your hard work.

Last, but not least, is my family. All of you deserve perhaps the most significant thank you. **Mema**, **Papa**, **Grandma**, and **Grandpa**, you were all here to see me start this journey, and in so many ways, spending time with you through the years planted seeds in my head that grew into this desire to pursue a PhD. I wish you could have seen me finish. **Scott**, you've always pushed me to exude confidence and never sell myself short. This advice has transformed me in all facets of life, thank you. **Mom** and **Dad**, I'll never be able to repay you for all the sacrifices you've made as parents to make the best life for me. Thank you for supporting the "forever student" in me, and for always making it a priority to visit. **Whitney**, leaving your home, family, and job to start over in a new city so I could pursue a PhD is a true act of selfless love. Additionally, your resilience, unrelenting support of me, and ability to turn long runs into a fun activity, are the biggest reasons I will look back with a certain fondness on what have undoubtedly been some of the most challenging years of my life. Simply put, I can't thank you enough for sticking with me and keeping me going - I love you so much. And finally, **Evelyn**, you are a source of true joy in my life, and I am just so glad you're here. The happiness you see on my face as I am finishing, is the same happiness I remember seeing on my dad's face when I found out he was done with his PhD all those years ago. I still remember this.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

The knowledge of three-dimensional structure and evolution of matter at the atomic scale represents one of the most fundamentally important challenges to modern science and technology. The positions and motions of atoms in a material form the basis for all further study of its constituent properties and macroscopic behavior, and the pursuit of this understanding has been the driving force for decades of research in materials characterization at the atomic level. Along the way, microscopy has provided a powerful qualitative means for visualizing and interpreting structures, and now, with the incorporation of advanced electron optical components [10–13], scanning transmission electron microscopy (STEM) has emerged as accurate quantitative visualization tools for atomic structure [14]. Indeed, this has transformed atomic-resolution imaging into a nearly routine technique for structural analysis; however, the breadth, complexity and volume of data generated in the course of imaging has expanded significantly, leaving potentially valuable visual information from experiments vastly underutilized. And while computer vision (CV) and deep learning (DL) have had an unprecedented impact on automation, decision support, and smart search systems in complex, high-volume image applications [15], the overarching atomic-scale structure problem from experimental imaging remains elusive.

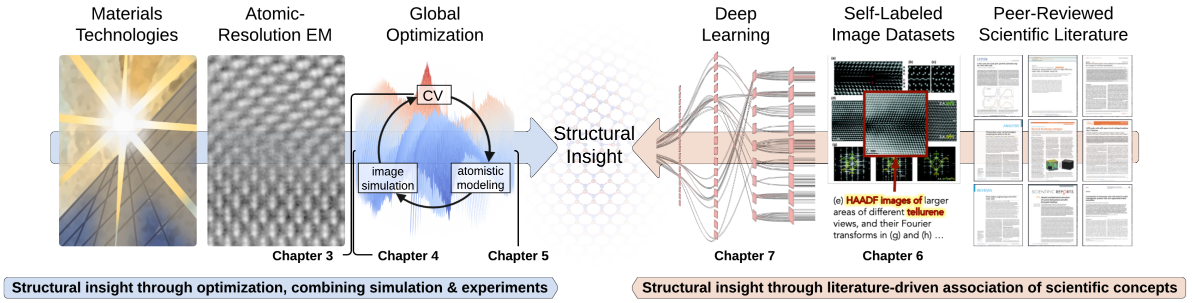

Materials Technologies
Atomic-Resolution EM
Global Optimization
CV
image simulation
atomistic modeling
Structural Insight
Deep Learning
Self-Labeled Image Datasets
Peer-Reviewed Scientific Literature
(e) HAADF images of larger areas of different tellurene views, and their Fourier transforms in (g) and (h) ...
Chapter 3
Chapter 4
Chapter 5
Chapter 7
Chapter 6
Structural insight through optimization, combining simulation & experiments
Structural insight through literature-driven association of scientific concepts

Figure 1.1. The grand vision for structural interpretation pipelines in atomic-resolution imaging. A left-to-center traversal illustrates a viable path to structure prediction, starting from atomic-resolution images. A right-to-center traversal highlights a workflow for information extraction from scientific images scraped from journal articles. At the center, the general term "structural insight" is used to encompass both 3D structure prediction in addition to textual descriptions referencing important structural characteristics of the material system at hand. The chapters of this thesis are aligned with this schematic along the bottom margin, to provide greater context for how each study fits into this grand vision.

In recent years, global optimization algorithms [16–20] have been deployed to search high-dimensional energy landscapes for plausible structures; however, the material structure in question is often complex and non-periodic at the atomic scale, so minimum energy solutions will not always corroborate experimental findings. To obtain plausible structure solutions for complex materials using global optimization, a new paradigm of analysis is needed that integrates experimental characterization results with atomistic simulations [21]. Because microscopy is primarily an image-based characterization method, the success of this integration relies largely on methods from computer vision to quantify similarity between image simulations of the proposed structures, and experimental targets. This global optimization approach, driven by vision-based comparisons of simulation and experimental imaging data, is illustrated schematically in a left-to-center read of Figure 1.1, and is one of the fundamental topics explored in this thesis.

While computer vision addresses similarity in microscopy from the perspective of visual appearance, visual appearance alone does not guarantee the identity of the underlying 3D structure; more context is often necessary. As an anecdotal example, consider shadow art [22], where 2D shadows cast by an assortments of everyday objects (often piles of trash) can be configured to resemble the silhouette of any famous city skyline [23]. In structure determination scenarios, atomistic simulation, which was the initial technique mentioned, can serve as the additional context. But as alternative, consider that perhaps image context could be enriched with the inclusion of characterization data and descriptions from literature, as a vast majority of reported scientific knowledge based on experimental findings is presented in the form of a paper, patent, or thesis, *etc.* Unfortunately, human efforts to extract this useful data across all relevant methods in which scientific information is disseminated cannot possibly scale with current outputs, thus motivating the need for automated information extraction and association by machines. Automated data extraction has been explored in the chemistry and materials science spaces, notably with the ChemDataExtractor (CDE) [24], and a machine learning (ML) tools to extract synthesis parameters for oxide materials from literature [25]. Both the CDE and ML tools use Natural Language Processing (NLP) as the backbone for automatically distinguishing and resolving groups of words as concepts. Moreover, literature is often the first place researchers turn to test if their ideas can be corroborated, contradicted, or even entirely novel – so it makes sense to incorporate this knowledge in the automation loop.

In a mature form, literature-based knowledge would provide general information to enhance the understanding of the structure in a general microscopy image. In this sense, it is perhaps best described as explanatory image captioning, which is a current activate area of research in computer vision [26, 27]. In this work, we develop a pipeline to enable large-scale image aggregation with literature-informed descriptions, using image captions to propose informative labels. With annotated datasets constructed from scientific literature, users are well-positioned to train neural networks for classification and recognition tasks specific to microscopy – tasks often otherwise inhibited by a lack of training data. This literature-driven association of properties/concepts back to proposed structures from experimental microscopy images is illustrated schematically in a right-to-center read of Figure 1.1 and is another significant topic in this thesis.

It should be noted that the paths in Figure 1.1), outlining the main visions for this work, complement each other. For example, on the left in Figure 1.1, global optimization-driven structure search, constrained by energetic minimization and the similarity between an experimental atomic-resolution microscopy image and a simulated image of a proposed structure, can help identify possible atomic structures for captured image. Examples of this structure solution pipeline are outlined in Chapter 4 and 5, with related image similarity comparison scenarios addressed Chapter 3. Background information pertinent to integrating computation with experiments is included in Chapter 2. Chapter 6 follows from the right side of Figure 1.1, and details an approach to construct self-annotated image datasets from literature. These self-annotated datasets could be used to train deep learning (DL) models, the primary direction for future considerations discussed in Chapter 7. Unlocking DL for these materials images would mean the ability to provide rich domain context or even practical explanation to novel experimental images. All things considered, elucidating materials structure and structure-properties relations (*i.e.*, "structural insight") lies at the heart of both paths.

CHAPTER 2

Background and Methods

This chapter presents several topics that are fundamental to the study of vision-based automation pipelines in this work. Section 2.1 covers a brief history of simulation in electron microscopy, as well as summarize the mathematical foundation for the computation of a simulated electron microscopy image. Section 2.2 reviews some of the optimization approaches used for determining 3D atomic structures, with an emphasis on how image simulation can be integrated into the overall workflow. Section 2.3 is an overview of the field of computer vision with mention of applications to general microscopy. Finally, section 2.4 introduces natural language processing and the various ways it can be leveraged to analyze text in captions of scientific journal articles.

2.1. Electron Microscopy Image Simulation

As a traditional materials characterization technique, electron microscopy often refers to the classic broad-beam system developed by Ruska and Knoll in the early 1930's [28, 29]. Since its conception, numerous improvements to both hardware and software components have elevated electron microscopy from a qualitative imaging device, to a quantitative characterization tool capable of providing chemical, structural, and electronic information for materials with extraordinary precision [14]. The workhorse of quantitative electron microscopy is the aberration-corrected scanning transmission electron microscope (STEM). Aberration correction creates a highly-focused, sub-nanometer electron probe, small enough to resolve inter-atomic spacings in most crystal structures. STEM operates by scanning this probe across a specimen, and then assembling the scattered electron signal serially into an image. At large collection angles, elastic "Rutherford" scattering dominates. In this high-angle annular dark-field (HAADF) regime, scattering is approximately proportional to both atomic number, Z, and number of atoms in an atomic column. Despite the simplicity of this description, real systems often show significant deviations from a "Z-contrast" interpretation, and further consideration must be given to the finer effects confounded in the image contrast [30]. Many of these are specimen-related factors (*e.g.*, sample thickness, lattice strain, defects, interface terminations, *etc.*) that change the probe channeling conditions and introduce diffraction contrast effects. Detector geometry plays a role as well [31]. Several studies have addressed limitations of Z-contrast interpretation [32–34]. In these instances, inferring structure from a small collection of sample images represents an inverse problem with no unique solution. For progress to be made on the inference end, the interpretation of image contrast requires comparison against physics-based "forward modeling" simulations. In the following, I summarize the mathematical framework underlying STEM image simulation.

2.1.1. STEM Image Simulation

There are a few widespread concepts throughout STEM image simulation, and the variety of techniques that do exist, largely evolve from differences in approximation, or in the allocation of computing resources. Fundamentally, imaging electrons can be modeled by superimposing several planewaves to form a spherically convergent probe wavefunction, ψ, and simulation involves the transmission of this probe through the specimen. The result is a modulated probe at the specimen's exit surface, defined in context of the specimen or slice j, as:

$$\psi^{j+1}(\mathbf{r}) = \psi^{j}(\mathbf{r})t^{j}(\mathbf{r}) \tag{2.1}$$

where $\mathbf{r}$ is a 2D vector of the probe position in the image plane, $\psi^{j}(\mathbf{r})$ and $\psi^{j+1}(\mathbf{r})$ are the probe wavefunctions at the incident and exit surface, respectively, and $t(\mathbf{r})$ is the specimen transmission function. The transmission function captures the interaction between the electrostatic potential of the specimen and the charge on the imaging electrons, as a single scattering event, and is thus defined as:

$$t^{j}(\mathbf{r}) = \exp\left(i\sigma v_{z}^{(j)}(\mathbf{r})\right) \tag{2.2}$$

where σ is a probe-specimen interaction parameter, and $v_{z}^{(j)}(\mathbf{r})$ is the potential of the slice j volume for a position $\mathbf{r}$. This specimen potential is often modeled as a linear superposition of potentials of each atom m in the slice j:

$$v_{z}^{(j)}(\mathbf{r}) = \sum_{m=1}^{N} v_{zm}\delta(\mathbf{r} - \mathbf{r_{m}}) \tag{2.3}$$

where $v_{zm}(\mathbf{r})$ is the projected atomic potential, which is frequently tabulated and stored prior to the calculation. This linear superposition of atom potentials assumes large separations between atoms – which is not the case in a bonded solid. However, the effects of bonding are usually small for high-angle scattering, which is one of the assumptions for image simulation in this work. Finally, image formation (for a specific probe position) is then a matter of taking the modulated wave function in diffraction space,

$$\Psi^{(j+1)}(\mathbf{k}) = \mathrm{FT}\{\psi^{(j+1)}(\mathbf{r})\} \tag{2.4}$$

and integrating the absolute square over a specified detector area

$$I(\mathbf{r}) = \int_{k\epsilon\mathbf{D}} \left|\Psi^{(j+1)}(\mathbf{k},\mathbf{r})\right|^2 d^2\mathbf{k} \tag{2.5}$$

where FT{ } is the Fourier transform, k, is a 2D coordinate vector in Fourier space encoding detector angle, and **D** represents the space of all detector angles considered. The result of this integration is the signal corresponding to a specific probe position. This is repeated until the probe has sampled the entire specimen and a full image is recorded. The value of the specimen/slice thickness has obvious implications related to the probe's transmission. In the most naïve scenario, the thickness of the specimen is ignored, which means that quantitative accuracy is sacrificed for practical (3D) specimens. Accordingly, this approach is referred to as the method of simulation for thin samples, and though this is considerably limiting, there are instances where the image is subjectively the same as one obtained from a much more expensive calculation [35]. The image simulation techniques in this work (outlined below) make use of and/or extend these concepts to achieve improved simulation accuracy or speed. The ultimate goal is to achieve an improvement in both, but this is often the trade-off.

Convolution. The convolution method refers to a simple incoherent imaging model based on the principles of Fourier optics, which defines an image as a convolution between a point spread function (PSF) and an object function. In general terms, a PSF describes the 2D intensity distribution of a point source image, and the object function is a transmission function that carries information about the object. In the context of microscopy simulation, the convolution method is an approximate linear imaging model, assuming incoherent imaging, and is given by

$$I(\mathbf{r}) = t^*(\mathbf{r}, \mathbf{Z}) \circledast |\psi(\mathbf{r})|^2 \tag{2.6}$$

where $\mathbf{r}$ is the 2D probe position, and $\psi(\mathbf{r})$ is the probe wavefunction (PSF) as before. The object function, $t^*(\mathbf{r}, \mathbf{Z})$, represents approximate probabilities of scattering to large detector angles (an approximate transmission function), and can be defined with partial cross sections [35], or as a linear superposition of atomic numbers [36]

$$t^*(\mathbf{r}, \mathbf{Z}) = \sum_{m=1}^{N} Z_m^{1.7} \delta(\mathbf{r} - \mathbf{r_m}) \tag{2.7}$$

where Z_m is the atomic number of the mth atom, and the $Z^{1.7}$ dependence (a deviation from the Z^2 dependence predicted by Rutherford scattering) better captures the effects of core electron screening, which is observed in peak single atom signals in HAADF STEM [35]. This Z dependence scattering is the foundation of Z-contrast imaging, and generally, the signal appears to vary between $Z^{1.5}$ (lower Z) to $Z^{1.7}$ (typical Z) with carbon ($Z = 6$) serving as an approximate division between the two regimes. This convolution formulation is convenient because with both input signals transformed into the frequency domain, the calculation of the image involves only multiplication. This provides a tremendous speed advantage over other techniques, and though convolution is procedurally different than the method for thin samples, it captures many of the same pertinent features [35]. Keep in mind that neither convolution nor the method for thin samples are quantitatively accurate for specimens exceeding a few nanometers in thickness; however, because of their speed and the fact they can still provide valuable qualitative structural insight, they are often some of the most practical simulation options.

Multislice. The multislice method of Cowley and Moodie [37] is widely used in STEM image simulation to accurately account for both specimen thickness (up to a few thousand angstroms), and plural scattering events. As the name implies, this method involves first dividing the specimen into thin slices along the electron beam direction. And then, just as in the method for thin samples, projected atomic potentials are used to define a transmission function, and the probe wavefunction for a single position is transmitted through a thickness. From here, a multiplication in Fourier space propagates the transmitted wave to the next potential slice, and the sequence of transmission and propagation continues until the wavefunction has made it all the way through the specimen. The propagation step uses the Fresnel propagation operator:

$$p(\mathbf{k}) = \exp\left(i\pi\lambda|\mathbf{k}|^2 t\right) \tag{2.8}$$

where λ is the relativistic electron wavelength and t is the thickness of the slice. The sequential transmission and propagation of the probe wavefunction can be written in a compact form using multiple Fourier transforms as:

$$\psi^{(j+1)}(\mathbf{r}) = \mathrm{FT}^{-1}\{p(\mathbf{k})\mathrm{FT}\{\psi^{(j)}(\mathbf{r})t^{(j)}(\mathbf{r})\}\} \tag{2.9}$$

where $\mathrm{FT}^{-1}\{\ \}$ is the inverse Fourier transform. Once the wavefunction has made it through, the same integration over the collection angles in the diffraction plane are used to extract pixel intensity value for the pixel in question. The multislice algorithm is computationally efficient for simulation of conventional plane-wave TEM images of ordered structures, but not for STEM, where the number of probes can easily exceed tens if not hundreds of thousands for realistic sample sizes. To make matters worse, multislice is often used to validate experimental results where there is a good deal of uncertainty surrounding the precise values for simulation parameters and atomic coordinate positions, so often more than a single simulation is needed. Furthermore, if defects are involved, the cell size for simulation must be artificially expanded to essentially isolate the defects from possible wrap-around errors [35]. Fortunately, the multislice approach itself is embarrassingly parallel in the sense that there is no dependency between each probe as it is propagated through the sample. Recently, to address issues of speed and scaling, hardware-based approaches such hybrid CPU+GPU and multi-GPU approaches to solving the general electron scattering problem [38–40] have begun to exploit various opportunities for data parallelization. In this work, a more efficient formulation of the electron scattering problem, the plane wave reciprocal space interpolated scattering matrix algorithm (PRISM), is used as a method to speed up the simulation process and maintain levels of accuracy comparable or equivalent to multislice.

PRISM. The main innovation of the plane wave reciprocal space interpolated scattering matrix algorithm (PRISM) method [41] is in the handling of the probe wave function. Instead of propagating a focused probe through the sample at all positions sequentially, a subset of the plane waves is used to construct the converged probes, which are defined for all positions across the sample at once. Then, using a Fourier interpolation factor to select a subset of plane waves, these partial waves are propagated through the sample with multislice. It is assumed that plane waves forming a complete plane wave basis set, can be multiplied by their associated complex coefficients (after modulation) and summed to generate the electron probe. There are markedly fewer plane waves to propagate through the samples than individual probes, and this is the primary source of speed up. Beyond this, PRISM shares slicing and computation of the projected potential steps with multislice. In this work, PRISM is used for simulating HAADF STEM datasets.

2.1.2. STM Image Simulation

The development of accurate methods for STEM simulation were largely motivated by the fact that experimental STEM images could be used to probe atomic-scale structural details. Likewise, researchers also realized the potential of yet another powerful form of microscopy for imaging finer electronic structure details at the surface of a sample: scanning tunneling microscopy (STM). Probing electronic structure is possible on the basis of the quantum mechanical principle of electron tunneling. In practice, a small bias applied between a probe tip and a surface induces electron tunneling across the gap, and variations in the tunneling current (constant height mode), or in the height to produce the same tunneling current (constant current mode), are measured as the tip is rastered across the sample. This tunneling current is important because it is proportional to the local density of states [42] and when viewed as an image, provides a glimpse of the atomic and electronic structure of the material surface. However, interpretation of these images without context from simulation, can lead to erroneous assumptions about the atomic structure of the material.

The foundation of the STM image simulation is a density functional theory (DFT) calculation. DFT calculates electronic wavefunctions for an optimized geometry, which yields a number of key properties such as total energy, band structure, and most relevant to STM calculations, the partial charge densities. The partial charge density data is recorded in the PARCHG file - using VASP (Vienna Ab-initio Simulation Package [43]). With this file as the primary input, a set of four parameters define a simulated image: (1) electron density value (corresponds to a constant current value set before acquisition of the experimental image), (2) its associated tolerance, and an (3) upper and (4) lower bound on the vertical distance above and below the surface, respectively. The STM image is then formed using the partial charge density data within the window defined by the electron density value and the tolerance, where image intensity corresponds to the z-coordinate at a given xy-grid point. This approach is consistent with the description given in the seminal work by Tersoff and Hamann [42]. Because a DFT calculation is involved in its own right, this image simulation approach that depends on the partial charge information from a DFT calculation is not well-suited for situations where little is known about the structure itself. With sufficient knowledge of the underlying structure though, modest iteration within a parameter space influencing image appearance is feasible, and is a topic explored in greater depth in this work.

2.2. Global Optimization for the Nanostructure Problem

Historically, X-ray diffraction (XRD) has been the mainstay for building atomic models of crystalline structures. With XRD, measured diffraction patterns are indexed relative to known patterns to help deduce crystal structures, phases, orientations, *etc.* However, XRD does not provide the resolution needed to observe local structural differences or defect geometries on an atom-by-atom basis. Moreover, even popular characterization techniques like pair distribution function (PDF) and X-ray absorption spectroscopy that are sensitive to local structure changes, do not provide requisite resolution to truly probe local structure. Refer to [44] and [45] for further details on XRD and PDF, respectively.

These shortcomings have since paved the way for direct atomic-resolution imaging methods that utilize STEM, electron diffraction imaging, atom probe tomography, *etc.*; yet in many cases, the experimental capacity to take measurements has exceeded the ability to determine atomic structure from measurements in the first place. In other words, this inability to obtain information about the positions of the atoms directly from observation is less a challenge of instrument resolution and precision, and more a challenge inherent to the ill-posed, inverse problem that is the "nanostructure problem" [46]. Rather, the "nanostructure problem" is the problem of determining 3D atomic structures from nanoscale observations, and because of the lack of analytical expressions, massive computing costs for simulation, complexity of the objective, *etc.*, unambiguously relating nanoscale observations to 3D structures is highly nontrivial. It is the nanostructure problem which underlies or exacerbates many of the fundamental materials understanding barriers addressed in this work. In the following, I provide an overview of global optimization and related concepts.

2.2.1. Global Optimization

Global optimization is found across a wide variety of quantitative disciplines as a way to find the “best available” solution given an objective defined over a complex input domain [47]. In the context of materials physics, global optimization is often used in conjunction with atomistic or first-principles calculations, to match structural and energetic quantities for force fields and nanoclusters [48–51]. For determining 3D atomic structures (*i.e.*, nanostructure problem), global optimization strategies tend to fall into a few different categories based on how the structures are represented, updated, and assessed. These strategies include, but are not limited to, minima-hopping [16], metadynamics [17], particle swarm [18], basinhopping [19], evolutionary algorithms [20], *etc.* In this work, rather than single objective (typically, total energy) global optimization, we are concerned with multi-objective optimization to determine structures which are not the globally lowest energy solutions but are nonetheless experimentally relevant.

An overall high-level framework for global optimization applied to the nanostructure problem is outlined in the flowchart in Figure 2.1. The specific strategy used in this work is a custom, multiobjective variant of the basinhopping (BH) approach [19] that we call “grand canonical” basinhopping. The multiobjective fitness measurements assess the quality of a proposed solution from both a simulated-to-experimental matching perspective, as well as an energetic-feasibility perspective. Both the grand canonical assumption, as well as the components of the multiobjective fitness measurements are explained in greater detail in the following sections and are implemented in context of the FANTASTX (Fully Automated Nanoscale To Atomistic Structures from Theory and eXperiments) software package [52]. Starting with a single structure, or a population of candidate structures,

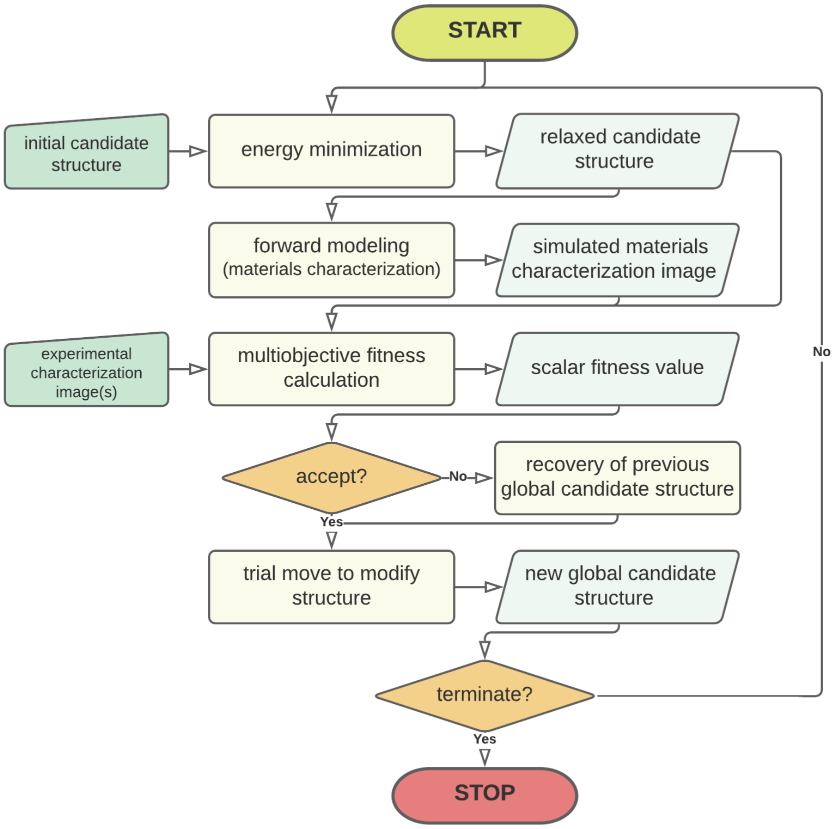

START
initial candidate structure
energy minimization
relaxed candidate structure
forward modeling
(materials characterization)
simulated materials characterization image
experimental characterization image(s)
multiobjective fitness calculation
scalar fitness value
accept?
No
recovery of previous global candidate structure
Yes
trial move to modify structure
new global candidate structure
terminate?
No
Yes
STOP

Figure 2.1. The global optimization strategy applied to the nanostructure problem. The user inputs are the initial candidate structure and the experimental characterization images. The candidate structure is modified incrementally and subject to fitness evaluation for each time-step. Fitness evaluation is multiobjective, involving (1) an image-based “match-to-experiment”, and (2) energy calculations from structural relaxation. The termination criterion is typically fitness and/or iteration-based.

global optimization proceeds as the simulated probes of nanostructure (*e.g.*, simulated STEM images of the candidate structures in many cases), are compared with the experimental targets to determine the fitness of the proposed structure(s). If current solutions do not register as high fitness, they are often rejected, and the previous global candidate structure is recovered.

2.2.2. Grand Canonical Basinhopping

Basinhopping [19] is a Monte Carlo-like method where solutions evolve by taking random steps (random perturbations of the current structure), that are accepted with probability based on the Metropolis-Hastings criterion [53]:

$$P_{accept} = \min\{1, e^{-(F_{new} - F_{curr})/T}\} \tag{2.10}$$

where F_{curr} and F_{new} are the finesses of the current and new structures, respectively, and T, is the system temperature at the current step. An energetic minimization step follows a perturbation to ensure that the perturbation "hop" reaches a region around the minimum of the potential energy surface "basin". The energetic minimization step is usually a DFT calculation, or an atomistic simulation based on empirical potentials [54, 55]. The size of the system (number of particles), and the availability of an interatomic potential will often dictate which is used. Standard perturbations for atomic structures in basinhopping involve random steps, which are random fluctuations of the atomic position. In order to address the uncertainty surrounding the stoichiometry of the system, this work considers atom addition and subtraction operations under constant chemical potential μ (*i.e.,* a grand canonical ensemble). As part of the Metropolis-Hasting criterion (Eqn. 2.10), the system temperature quantifies how often a less fit solution is accepted over a more fit one, and is the sort of mechanism that allows a structure to "hop" out of local minima. If the system temperature is zero, BH is greedy in that only solutions that improve fitness will be accepted, which is sufficient if the structure is making a descent to a global minimum. However, in order to assure that the structure can get there, solutions that are less desirable must be accepted and used along the way. This idea has strong parallels to solution diversity in genetic search [56], with the intuitive justification being that a completely homogeneous population of solutions are not capable of generating novelty in solution space.

2.2.3. Multiobjective Fitness Definition

Outside of a handful of design decisions pertaining specifically to the optimization strategy – most of these involve the update and termination criterion – the fundamental decision that is ubiquitous to all optimization is the definition of an objective, or what we refer to as “fitness” in the solution space. In ideal optimization situations, there is a single fitness quantity that can be measured from a solution, and that fitness is indicative of or strongly correlated with the true objective. Moreover, the solution space is convex, and optimization proceeds as a strict minimization or maximization of that fitness quantity. In realistic problems , optimal decisions require a trade-off between conflicting objectives, and by nature, the solution space is often non-convex and/or non-smooth. In context of the nanostructure problem addressed in this work, we identify total energy and image-based match-to-experiment as the two conflicting objectives to optimize. A grain boundary image is perhaps the quintessential example of a conflict between these two objectives when its structure is the target of global optimization. This is because a grain boundary is a metastable system. A grain boundary image represents a snapshot of a transient structure (albeit on a large time scale), and from a materials science perspective, this happens to be the state under which many of these materials are used. However, if minimizing the energy is the sole objective during global optimization, the structure will never appear as a solution during optimization. From an optimization perspective, this means that energy is obviously important from a physical standpoint, but only to the extent that the structure resembles what can be observed during imaging. This is the crux of defining this multiobjective fitness.

Two popular strategies for evaluating candidate solutions based on a multiobjective fitness are the weighted-sum and the Pareto-sampling approach. The weighted-sum approach is a classical strategy that involves assigning a relative importance (weight) to each objective, and then combining the objectives into a single, scalar cost function. The weighting can be set *a priori*, with sufficient domain knowledge and understanding of variable scaling, but in many cases, the significance of the objectives cannot be determined until the trade-offs between them are clearly understood. Moreover, penalty constraints can be added in the form of a regularization term, or as commonly employed in physics-based modeling – a subroutine to emulate the constraints of a physical model (*e.g.*, implement checks for bond lengths, coordination, *etc.*). This work utilizes a weighted-sum approach to define fitness for a candidate structure as:

$$Z_{total} = \alpha E_{total} + \beta d_{similarity} \tag{2.11}$$

where Z_{total} is the fitness (objective) score, E_{total} is total energy of the structure from the energy minimization step, $d_{similarity}$ is the image similarity distance, and α, β are the associated weights that are set based on the relative importance of the terms. A major downside to the weighted-sum approach is the possibility of a true optimal solution not being found/accepted, either because the fitness function excludes important aspects of the problem, or because of an inappropriate setting of the weights [57]. A Pareto-sampling approach (multiple single objective pareto sampling, MSOPS [58]) addresses the shortcomings associated with inappropriate weight settings by enabling all trade-offs among combinations of multiple objectives to be evaluated (considering the multiple objectives directly), and then sampling among the subset of solutions that cannot be improved in one objective without sacrificing performance in at least one other. These solutions that cannot be improved in a single objective without sacrificing others represent the space of best possible tradeoffs and are referred to individually as Pareto solutions, or collectively as a Pareto front. Pareto optimality originated from the concept of non-inferiority in context of economics [59]. For an overview of Pareto approaches in context of the broader field of multiobjective optimization, [59] is a good resource.

2.3. Computer Vision in Electron Microscopy

Computer vision algorithms encompass a variety of tools that are characterized by their ability to use low-level image processing primitives (*e.g.*, edges, colors, *etc.*) to observe images in a way that promotes reconstruction of physical properties, motion tracking of individual objects, and in some cases even scene understanding on a semantic level. For example, consider an algorithm that uses visual inputs to alert a 911 operator of a person in distress, based on a series of connected visual concepts originating from the localization of human joints in an image of a person lying unconscious on the ground in a crowd (*e.g.*, joints are connected via limbs, which together form a body, which can assume a pose, which in some cases is indicative of a physical ailment). This sort of action commands true understanding of the situation and is predicated on successful hierarchical processing and fusion of visual information. And though deriving an understanding from images and video is something that most humans can do in an almost effortless fashion, such tasks can pose an immense challenge to machines. For a good overview of computer vision as a discipline, refer to the lecture notes from Ying Wu's advanced computer vision course [1].

At the foundation of the computer vision (CV) hierarchy is image processing. Figure 2.2 provides a convenient visual illustration of the CV hierarchy. Image processing is not vision by itself, but rather, the preprocessing that ensures an image is in a format suitable for vision. With the hope of developing electron microscopy into a truly quantitative characterization technique, early adapters of image processing in the field of microscopy recognized the importance of delineating between image processing that produces a more

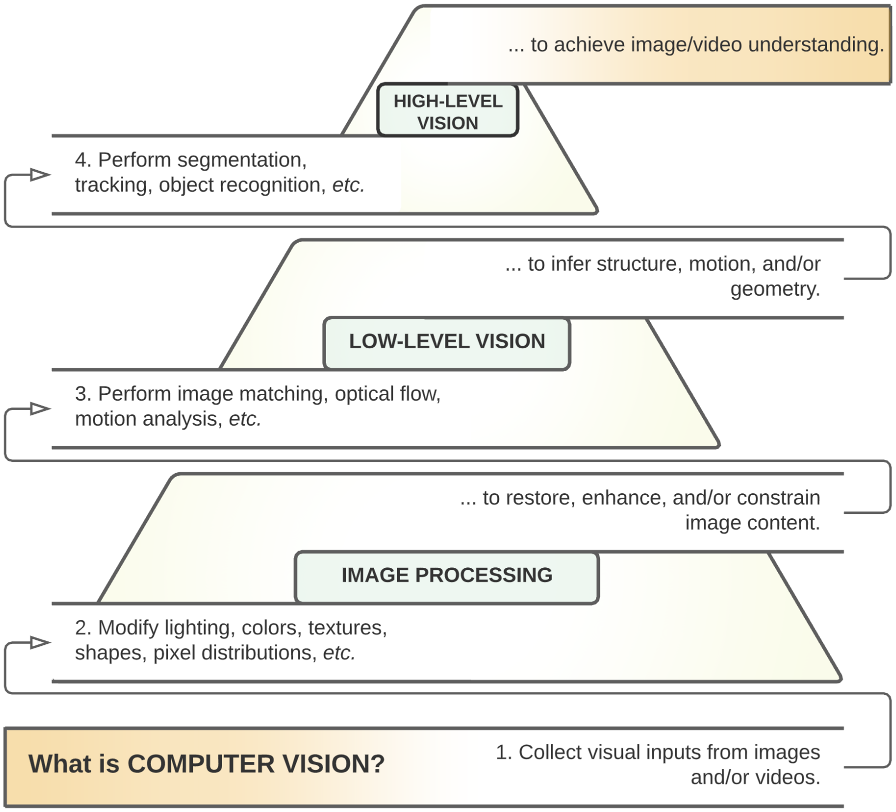

... to achieve image/video understanding.
HIGH-LEVEL VISION
4. Perform segmentation, tracking, object recognition, *etc.*
... to infer structure, motion, and/or geometry.
LOW-LEVEL VISION
3. Perform image matching, optical flow, motion analysis, *etc.*
... to restore, enhance, and/or constrain image content.
IMAGE PROCESSING
2. Modify lighting, colors, textures, shapes, pixel distributions, *etc.*
What is COMPUTER VISION?
1. Collect visual inputs from images and/or videos.

Figure 2.2. Hierarchical flow of processing steps that define a standard CV problem. At the base of the hierarchy is image processing, the first step after data collection. Image processing modifies characteristics of the image at a pixel level for the purposes of highlighting specific image content. Low-level vision builds on the processed image to begin extracting primarily structural elements on the image, and finally, high-level vision uses the inferred structure/geometry to perform recognition or segmentation with the ultimate goal of achieving image-based understanding. Adapted from [1]

pleasing visual image and that which respects true, unaltered image content [60]. Processing to improve aspects of the image by applying linear operations that affect color, contrast, and/or brightness is referred to as an enhancement. Common examples of this include black level subtraction (subtracts a uniform signal background from the object) and differential amplification (amplifies object signal relative to the background) [60]. As an alternative to enhancement, consider a class of optimization-based approaches that attempt to return an image to some ideal, but faithful form, by inverting degradation. Processing of this sort is referred to as restorative and perhaps the most common example for electron microscopy is the Wiener filter. A traditional Wiener filter produces an optimal estimate of the full uncorrupted microscopy image in the least squares sense, assuming signal and noise are uncorrelated, and is used extensively for image denoising – particularly in context of the powerful block-matching and 3D filtering (BM3D) denoising [61]. For more information and mathematical rigor, refer to [62], as Wiener filters represent a fundamental noise reduction approach with many variations.

After image processing, the first low-level vision tasks can be performed. Here, "low-level" implies that images are viewed or modified directly on the level of their pixel values (*i.e.*, pixel-wise). Following Figure 2.2, the focus here is on image matching and optical flow. Image matching, or what is referred to in this work as the task of quantifying image similarity, was addressed, somewhat indirectly, for electron microscopy in the pioneering works of Schiske [63] and later Gerchberg and Saxton [64], with the realization that a wave function's amplitude and phase at the exit surface of an object – which provides the valuable structure information for the object - could be reconstructed using iterative comparisons with intensity measurements. Assessment of the quality for each iteration involved comparison of the amplitudes associated with the generation of an image candidate, with the amplitudes of the reference image using (mean) squared error (MSE), which is pixel-wise image matching in its most basic form. However, as others have since recognized, MSE is not always ideal for vision tasks because it quantifies errors between the signals uniformly, which often does not map well to perceived visual quality [65, 66].

More sophisticated methods for quantifying image similarity attempt to weight different aspects of the error signal according to their visibility, or define similarity as a comparison of image information, or features rather than just all individual pixels. The other form of low-level vision notably found in electron microscopy involves techniques related to “optical flow” [67, 68]. Optical flow is motion estimation from local spatio-temporal variations of pixel intensities in an image sequence and is the basis of dense particle or object tracking [69, 70]. For example, quantification of motion can be used to implement spatial corrections between adjacent images to account for beam induced motions in cryo-EM [71] (*i.e.*, image registration), or for image segmentation purposes (*e.g.*, background flow is often different from object flow in the foreground). Image similarity, registration, and segmentation are covered in greater detail in subsequent sections.

As vision approaches “high-level” (Figure 2.2), the techniques and processes involved tend to require some degree of image understanding, which builds on the higher-level ways in which pixels are combined and interpreted. In materials science, several early studies interested in high-level vision in the form of automatic microstructure recognition or image parameter extraction began exploring tools like visual bag of words, texture and shape statistics, and/or pre-trained convolutional neural networks (CNNs) to first construct feature descriptors [72–74] for each image in the dataset. These approaches often showed promising results over the specific classes of materials represented in the imaging; however, in most cases, they were best suited for datasets of a modest size, and were heavily overshadowed by some of the developments in deep learning that used deep convolutional neural nets in an end-to-end fashion, not just for feature extraction, but to obtain human-level accuracy on high-level imaging and video recognition tasks [75].

Unsurprisingly, there was a marked increase in materials science studies using deep CNNs as a means to locate atomic species and classify the type of defects [76, 77], analyze CBED patterns [78], and automate particle extraction from raw cryo-EM micrographs [78]. Recently, in an effort to create high quality micrographs of beam sensitive materials, Ede and Beanland [79] train a CNN, in the context of a generative adversarial network (GAN), to complete realistic scanning transmission electron micrographs from partial scan data. In the most rudimentary form, filling in image pixel values over certain intervals is a type of image processing called image *inpainting*, which is in essence interpolation (*i.e.*, estimate a pixel value at a location, using a weighted averages of pixels' neighborhoods). However, in the context of a GAN framework, in which two neural networks improve their respective performance task via competition, the *inpainting* and the learning associated with generating an image that is both visually and physically realistic is more nuanced and is thus considered a high-level vision task. Overall, deep learning and computer vision are burgeoning areas of research within the materials science community as their potential value to the field is becoming widely recognized. The following sections include dedicated discussions on measuring image similarity, image registration, and deep learning with CNNs to provide greater detail to some of the specific CV tools and techniques utilized in this work. I would suggest for the reader to consult [80] and [81] for further discussion on how CV is used to improve both microscopy data and materials information in electron microscopy, as well as in biological and medical data, respectively.

2.3.1. Image Matching

In computer vision, image matching is a fundamental low-level vision tasks, requiring an objective that is high for images that are similar, and low for images that are not. Throughout this work, image matching (the specific task of finding the single most appropriate image to an image designated as a *target*), is referred to generally as quantifying image similarity. In all contexts, the *target* is the image that is fixed or is considered the image being matched to in various comparison scenarios.

The majority of the image similarity measurements in this work utilize low-level pixel-wise comparisons, (*i.e.*, only pixel-wise intensity information is considered). Contrast this to a class of techniques which first extract image features and perform matching as a vectorized comparison of features. The Earth Movers Distance (EMD) approach, described below, functions as a higher-level vision task and is the only feature-based method considered here. For more information on feature-based matching, refer to [82, 83]. The following introduces various image similarity measurements explored in this work.

Mean Square Error. As a result of its computational simplicity and desirable optimization characteristics (*i.e.*, it is quadratic and convex), pixel-wise mean square error (MSE) is a popular similarity measurement. MSE is expressed as:

$$MSE(X,Y) = \frac{1}{n_p} \sum_{\mathbf{p}} [X(p) - Y(p)]^2 \tag{2.12}$$

where $X(p)$ and $Y(p)$ are the image signals for all pixel values $\mathbf{p}$, and n_p is the number of total pixels in the image. Here, the choice of which image serves as the *target* is arbitrary from a measurement perspective since MSE is symmetric (*i.e.*, the order of operations does not matter). Image processing researchers and practitioners, particularly those in the field of image quality assessment, sometimes scrutinize the use of MSE (and the related PSNR) in quantifying visual similarity because it is typically not the best indicator of visual distortion [65] – but admittedly, this is highly application dependent [84]. A popular MSE variant, the root MSE (RMSE), is nothing more than the square root of the MSE value. RMSE has units that are directly interpretable meaning the RMSE has the same unit as the unit on the pixel values, which could perhaps be beneficial if the values of the pixels in the image are quantitatively expressive.

Structural Similarity Index. In an effort to incorporate image matching objectives that quantify observable characteristics of the human visual system (HVS), the Structural Similarity Index (SSIM) [85] is included in the following analyses. SSIM quantifies the retention of signal structure between images (something the HVS is highly adapted to extract), and in this context, is an average of the weighted multiplicative combination of terms that represent luminance, contrast, and structure of a test block with respect to its fixed reference. The expression for the mean SSIM (referred to here as just SSIM) is:

$$\begin{aligned} SSIM(X,Y) = \frac{1}{n_b} \sum_{\mathbf{b}} & \left[\frac{2\mu_X(b)\mu_Y(b) + C_1}{\mu_X(b)^2 + \mu_Y(b)^2 + C_1} \right]^{\alpha} \\ & \left[\frac{2\sigma_X(b)\sigma_Y(b) + C_2}{\sigma_X(b)^2 + \sigma_Y(b)^2 + C_2} \right]^{\beta} \left[\frac{\sigma_{XY}(b) + C_3}{\sigma_X(b)\sigma_Y(b) + C_3} \right]^{\gamma} \end{aligned} \tag{2.13}$$

where μ_X and μ_Y and σ_X and σ_Y are sample means and standard deviations for all pixel values within a given pair of blocks $\mathbf{b}$, σ_{XY} is the sample cross correlation between the blocks, C_1, C_2, C_3 are small positive constants to avoid numerical instabilities, α, β, and γ are weights given to the luminance, contrast, and structure terms, respectively, and n_b is the number of total blocks pairs in the image. The actual "structure" that is part of the namesake, is represented by a loss of linear correlation (the third term weighted by γ). Because SSIM technically serves as a full reference metric (*i.e.* it is formulated to compare a given input against a pristine *target*), the X indexed values are considered to reference the *target* image. Although from a measurement value standpoint, the *target* designation is arbitrary because SSIM (like MSE) is symmetric.

Visual Information Fidelity. Visual information fidelity (VIF) [85, 86] incorporates a statistical model of the HVS into the measurement of image similarity. Figure 2.3 provides a high-level overview of the information-theoretic approach underlying VIF. At the foundation of the VIF measurement, is the calculation of two information quantities: (1) mutual information between the *target* (*i.e.* the *actual* reference image) and output of the *target* passed through the HVS channel (this is considered reference information content), and (2) mutual information between the *target* and the output of the distorted image passed through the HVS channel (this is considered distorted information content and it is assumed that the distorted image can be modeled as the *target* passed through a distortion channel). In this work, VIFP, a computationally simpler multi-scale pixel-based implementation of VIF is used, but the concept illustrated in Figure 2.3 remains

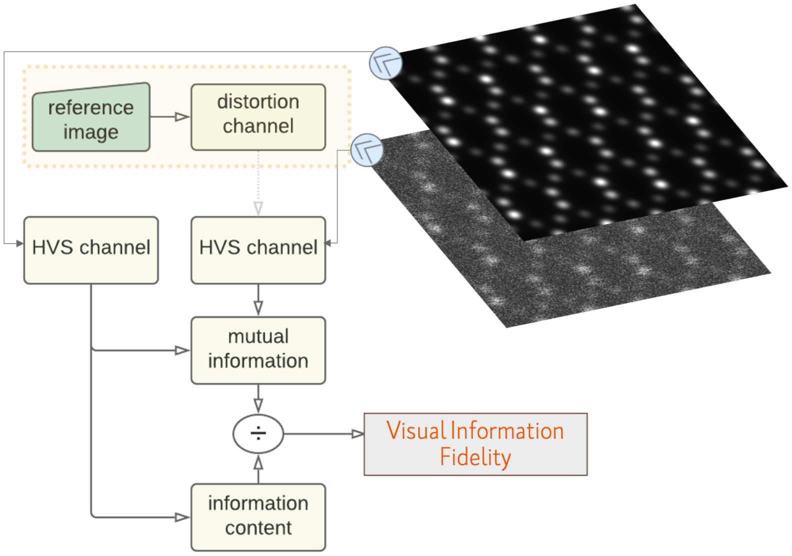


Figure 2.3. System diagram for visual information fidelity (VIF). In theory, the distorted image is modeled as the reference image subject to some distortion. In practice, a given distorted image is fed directly into the HVS model and its relation to the reference via the distortion operation is assumed. The final measurement is the ratio between two information measures: the reference information content (*i.e.*, the mutual information between the actual and perceived *target* images), and the mutual information between the actual *target* and distorted images. Adapted from [2]

the same. VIFP can be expressed as :

$$VIFP(X,Y) = \frac{\sum_{\mathbf{s}} \sum_{\mathbf{b}} I(x(s,b); f_{HVS}(y(s,b))}{\sum_{\mathbf{s}} \sum_{\mathbf{b}} I(x(s,b); f_{HVS}(x(s,b))} \qquad (2.14)$$

where x and y are arrays of pixels values inside a block, b, at a particular scale s, $I(\bullet;\circ)$ is the mutual information between the inputs, and f_{HVS} is the HVS channel applied to an input. A "0" value for VIFP means that all information is lost to distortion. A value of "1" means that the reference content is identical to the distortion. Values greater than "1" imply that the contrast of the distortion channel is enhanced over the reference.

Earth Movers Distance. The Earth Mover's Distance (EMD) can be used to measure the distance between images that are summarized by vector representations of their features [87]. As such, EMD is considered a feature-based matching approach, and assumes that the minimum work required to move the given image's collection of features (*i.e.*, piles of "earth") across some distance into the collection of features associated with the *target* (*i.e.*, holes), is correlated with visual similarity between the images. The amount of "earth" to move and the capacity of the available holes is variable, which means that EMD allows for comparisons between sets of features that are different in size. The work minimization that is the backbone of the EMD calculation is based on the classic transportation problem from linear programming [88]. Let S_X and S_Y be the collection of weighted feature vectors the images and, $\mathbf{D} = [d_{ij}]$ be the ground distance matrix in feature space, and $\mathbf{F} = [f_{ij}]$ be the flow matrix (i.e. specifying how much of one feature is transported to another). With this, the objective is to find a flow that minimizes the overall work:

$$WORK(S_X, S_Y, \mathbf{F}) = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{i,j} f_{i,j} \tag{2.15}$$

subject to the normal constraints of a transportation problem [88]. Once the optimal flow is determined as the solution to the transportation problem, EMD is defined as the resulting work normalized by the total flow that occurs as part of the process:

$$EMD(S_X, S_Y) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} d_{i,j} f_{i,j}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j}} \tag{2.16}$$

For microscopy images, there are many feasible ways to vectorize and weight the collections of features, S_X and S_Y. This is explored further in Chapter 3.

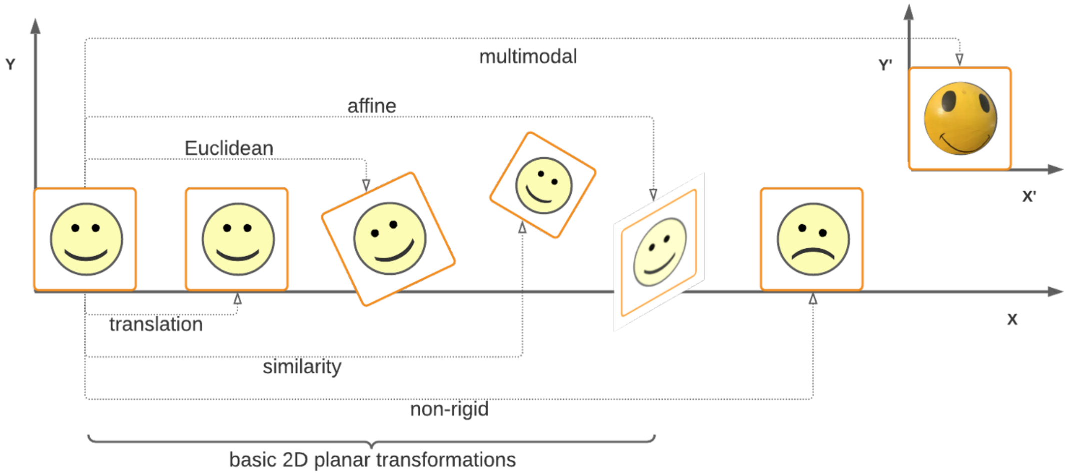


Figure 2.4. Basic 2D planar transformations (linear transformations applied with matrix multiplications to vertices/pixels) alongside a non-rigid (allows local warping) and multimodal transformation. Adapted from [3].

2.3.2. Image Registration

The goal of image registration is to establish a one-to-one pixel-wise correspondence between two or more adjacent images. References to image registration in this work imply the computation of 2D planar transformations - illustrated in Figure 2.4. Search strategies to obtain the values parameterizing the ideal spatial correction necessary to bring the two images into coincidence are often formulated as an optimization problem whose objective makes use of an image similarity measurement. Mismatch of modality (*i.e.* registering a pristine simulation with a real experimental image) is a significant problem in the context of simulation-to-experimental image comparisons, which often involves rectifying sizable differences in contrast, noise levels, structural uncertainty *etc.* In Figure 2.4, a multimodal registration involves defining an appropriate transformation from the 2D planar “drawn” smiley, to a realistic 3D photographed smiley.

In general, the image registration methods implemented in this work are based on a coarse-to-fine search strategy. The “coarse” search begins after the selection or simulation of a candidate image patch (referred to as a reference window) and uses downsampling and quantization (reduces the size and depth of pixel value in the images) to make the initial comparisons relatively inexpensive. During this step, the *target* image is held fixed, and a variant of cross-correlation is used to find the appropriate translation of the reference window. In the literature, this approach is often called template matching [89]. A cross-correlation computation involves moving a portion of the reference window over a *target* and computing the sum of products at each location. The cross-correlation will be highest at positions where the reference window is best aligned (most similar) to the *target*. If rotational differences between the images exist, then an additional iterative step that considers rotations of the reference window and the associated RMSE between pixel values is used to determine the optimal rotation of the reference window. This transformation, a combination of translation and rotation, is often referred to as a Euclidean transformation, because Euclidean distances are preserved (see Figure 2.4 for and example). The fine search step then proceeds as an iterative optimization of a “similarity measure” on the original (or upsampled) versions of the images. Depending on the upsampling and interpolation techniques, subpixel translations as well as minor rotational offsets can be computed.

Whenever the correspondence between the structures in the images cannot achieved using a basic 2D planar transform (Figure 2.4), the next strategy to explore often involves some form of localized stretching. These strategies belong to the realm of non-rigid image registration and are beyond the scope of registration techniques considered in this work. To some extent, local modifications to the structure used in simulation of the reference window images is one way to capture localized deformations, however, in the broad literature on the topic, stretching operations are applied to the image directly, and this sort of bottom-up modification to the imaged object is not as common. Non-rigid registration techniques are perhaps most commonly utilized in the field of medical imaging, and in particular, are well-studied in context of brain for capturing patterns of functional activity, resolving smaller anatomical differences across structures, *etc.* [89]. For a more comprehensive overview of non-rigid registration techniques, refer to [90].

2.3.3. Deep Learning

Deep learning (DL) is a type of machine learning (*i.e.*, a technique that "enables computer systems to improve with experience and data" [91]) that uses deep, nested hierarchies to facilitate the learning of complex representations. One of the workhorse DL architectures is the convolutional neural network (CNN). In the most general sense, a CNN is a dedicated sequence of linear and non-linear mathematical operations (layers) that transform an input volume into an output volume of arbitrary spatial size. A raw 2D image is perhaps the most common initial input for a CNN (the seminal work [75] is based on 2D imaging), and the intervening data, referred to as feature maps, evolve sequentially through each layer to reveal a progression of higher-level abstractions of the initial input.

Generally, linear operations are applied to feature maps in a sliding window manner. These operations, referred to as convolutions, multiply a portion of the feature map by a set of critical weights. To introduce non-linearity and dimensionality reduction into the system, the output feature map of the convolution is then typically fed to an activation layer, followed by some form of pooling. The activation layer uses a function (*i.e.*, a rectified linear unit (ReLU)), which applies a non-linearity that forces negative numbers to zero, and the pooling layer downsamples the features in each window to a single value. The final spatial size of the CNN output is dictated by the task. For example, in binary image classification, the initial color image of size $w \times h \times 3$ is transformed to a single column vector of size 2×1, which in some form represents the probability of each class. CNNs are trained in the sense that, to the extent that the final values do not match the anticipated (as measured by some objective), gradient methods are used to update the weights of the network. Weight updates are repeated until the network has learned the full transformation so that performance on the training set is adequate, yet general enough for future unseen cases.

Unfortunately, one of the challenges associated with training a CNN for microscopy or other general scientific applications, is obtaining a large labeled image dataset. To put things in context, one of the most popular public benchmark datasets for object recognition with natural images, ImageNet [92], contains over 14 million labeled images (each containing often several labeled objects). For scientific images, proper image annotation is often a highly nuanced task for experts only. To the extent that the task can be posed in layman's terms, there are resources available for crowdsourced data labeling. Of particular interest to this work, is crowdsourced data labeling through Amazon's Mechanical Turk. With this platform, we can post images for annotation and recruit a sizable workforce to help in the construction of a dataset for CNN training.

Object Recognition with YOLOv3. The goal of object recognition is to extend the image-level classifications to specific objects within the image. This requires each object first be identified and then referred to in some way by its location and class. Bounding boxes are rectilinear polygon annotations - usually rectangles – that indicate that presence of an object (usually of a specific class) in an image. As such, an image that passes through a CNN trained for object recognition will be transformed into a collection of bounding boxes, which can then be further processed for the downstream task at hand. The YOLO (You Only Look Once) [93] object detection algorithm is a popular method for fast object detection that boasts great localization and classification accuracy on a variety of public benchmarks. The object detection tasks in this work build on top of YOLOv3 [94]. Specifically, in Chapter 6 we detail the use of a YOLOv3-based object detection pipeline, which is constructed to locate and classify all of the individual image components of figures scraped from scientific journal articles.

Image Segmentation with Adversarial Networks. In pixel-wise segmentation, the goal is to assign each pixel of an input image to a specific class, and the output is of the same spatial size as the input. This is despite several transformations within the network which contract and expand the spatial size of the intervening data. The CV literature on segmentation is rich, as this is one of the oldest and most popular problems in the field. As such, I suggest interested readers begin with the overview of the traditional segmentation approaches in Richard Szelski's textbook [3], and then consult the literature for more recent advances. Here, I discuss image segmentation from the perspective of adversarial nets, which since their formal introduction in 2014 [15] have garnered much attention in the field of DL. Designing an adversarial net framework involves training a generative model (usually a CNN), by forcing it to compete against a discriminative adversary (a second CNN) which is also being trained, but in direct conflict with the generator. The discriminator wants to be able to tell if the results from the generator are real or fake. If the generator successfully fools the discriminator, then the distribution from which real data arises is learned. This framework for CNNs to compete in an adversarial fashion is referred to as a generative adversarial network (GAN) [15]. In its standard form, the generative model of the GAN starts generating samples based on random noise. To incorporate additional domain input in the design of both the generator and discriminator, it is common to construct a "conditional" GAN. Here, conditioning on additional information allows for more control over the data generation process, and in this form, the network is referred to as a conditional GAN (i.e. cGAN) [95].

With the cGAN framework, it is possible to pose the image segmentation problem as one of image-to-image translation. The adversarial image-to-image translation variant mentioned in Chapter 7 as part of a future direction to pursue, is a cGAN based on the popular pix2pix setup [96], which involves learning a custom mapping from input image to output image in an end-to-end fashion (output is same size and resolution as input). This is accomplished with a UNet [97] generator, taking an observed input image, X, alongside some and random noise vector z, and learning the mapping to Y, the segmented image (*i.e.*, $G : \{X, z\} \rightarrow Y$). Trained in an adversarial way, the generator learns to create valid image segmentations on unseen images that are as convincingly "real" as the ground truth segmentations provided in the training set. After training is complete, the generator functions as a custom filter that can be used to segment further images that are similar to the class of images presented during training. The particular segmentation problem suggested in Chapter 7 involves the separation of image annotations (*i.e.*, scale bars, subfigure labels, geometric shapes to highlight important details, *etc.*) from microscopy image content. This is a special case of the general microscopy image segmentation problem (separate and classify all image content) that is essential if images from literature are to be used in context of further image processing or CV pipelines.

2.4. Natural Language Processing for Scientific Literature

With the explosion of web-based communications, vast amounts of information are conveyed in the form of digital text. Even before the importance of digital text was realized, linguists had been interested in how structural models of language [98], language universals [99], and syntactic text parsing [100] (just to name a few) could be used to help machines analyze and understand human language and intent. The eventual maturation of these concepts in context of the availability of critical computational resources, created the foundation for modern natural language processing (NLP).

In its most rudimentary form, NLP uses computerized rule-based approaches to make sense of concepts in unstructured text. In practice, this involves the use of dictionary lookups, handcrafted word features, and hand-coded language rules. When the language is highly constrained, and the domain of applicability is narrow, rule-based approaches can be quite effective [101–103]. In contrast, when topic and content are relatively unconstrained, (*e.g.*, for general web-based communications such as articles, chat windows, *etc.*), the overwhelming size and the inherent ambiguity of natural language renders rule-based NLP inadequate as a standalone solution. As a result, the priorities in the field became reoriented towards simple statistical approaches over complex rule-based analysis [104]. In particular, machine learning methods based on probabilities were gaining traction, and despite some notable skepticism towards probabilistic language models [105], this makes sense: utilizing the overabundance of real language data, statistical models can learn the most discriminative representations for languages and the most common formalisms [104]. This statistical approach had a staying interest up until the early 2010's when it was gradually replaced by neural machine translation (see [106] for a seminal example), which used large neural networks (NNs) to read a sentence in one language and output the correct translation directly in another language. NNs have achieved impressive accuracy across a variety of NLP tasks including automatic speech recognition [107], visual question answering [108], sentiment analysis [109], *etc.* The popular Word2Vec [110] uses NNs to project large sparse vector representations of words into a lower-dimensional space that preserves semantic relationships. This has powerful implications for assigning textual descriptions to collections of words or phrases that are further explored it this work.

In the realm of physical science, scientific literature has been the primary target domain for the development of NLP tools. This is largely a byproduct of both abundance, and the fact that these journal articles use a systematic and highly-formalized scientific language to communicate ideas. Some of these first tools for NLP in this domain are designed to tackle "Chemical Entity Recognition", which involves the identification of molecular structure information from both structure tables and prose in the "Experiments" or "Synthesis" sections of scientific articles [24, 111, 112]. These basic tools established exclusive grammars related to chemistry documents and have been refined to handle arbitrary chemical and biological entities in scientific and technical document such as patents [113, 114], as well as extended to the field of materials science for creating autogenerated materials property databases [115]. Eventually, these specialized rule-driven material-informatics parsers were usurped by machine learning-based methods applied to more general unstructured text [116, 117]. The ability to robustly process textual data from literature opens the door for the creation of literature-trained machine learning pipeline [117, 118], which is an underlying motivation for this work. Here I outline NLP tasks that are used in the construction of one for these pipelines.

2.4.1. Rule-Base Caption Assignment

Figures are the standard means for displaying visually meaningful data in scientific articles. Moreover, figures are almost always accompanied by a caption – which can be very rich with descriptive information pertaining to specific elements of the figure. In fact, a reader can often get a good impression of the full article with only a careful inspection of the figures and captions. However, for a computer, the task of recognizing how the words of the caption text map to each individual image or collection of images in the full figure (*e.g.* caption assignment), is ambiguous at best. Here, we detail elements of the approach used in this work for caption assignment by means of rule-based sentence parsing. Much of this approach involves the create of custom extensions for standard NLP tools provided by the spaCy NLP library for Python [119].

In particular, this approach is based heavily on the use of custom regular expressions. Regular expressions are sequences of characters that describe patterns of text to extract during search. Regular expressions are used at the beginning of caption assignment to identify instances where a specific component of the figure (*i.e.*, a subfigure) is being referenced. These instances are referred to as *subfigure tokens* and the pattern of an open parenthesis, followed by a collection of internal alphabetic or numerical characters, proceeded by some closing of the opening punctuation, is one of several pattern options considered when defining *subfigure tokens*. Consider the caption text after subfigure tokenization, as outlined in Table 2.1. In the example in Table 2.1, the bolded *subfigure tokens* are recognized within the caption, and are classified to determine whether further enumeration is necessary to figure out the number of subfigures implied. For example, the **(a) and (b)** token belongs to the `parenthesis_02_and_alpha_02`, where `parenthesis`

Table 2.1. Example subfigure tokenization of a caption text revealing the token classification and the corresponding implied subfigures.

Caption text (raw)

(a) and (b) TEM images of 1.93 wt% Ru–WSe2. **(c)** HRTEM image of 1.93 wt% Ru–WSe2. **(d, e)** The enlarged area denoted in **(c)** corresponds to the HRTEM images of WSe2. **(f)** HAADF-STEM image of 1.93 wt% Ru–WSe2. **(g - i)** The EDS mapping of Ru, W, and Se, respectively.

Text adapted from: *Inorg. Chem. Front.,* 2019, **6**, 1382-1387

Subfigure tokens:

Token	Token class	Subfigures implied
(a) and (b)	`parenthesis_02_and_alpha_02`	[a , b]
(c)	`parenthesis_02_none_alpha_01`	[c]
(d, e)	`parenthesis_02_comma_alpha_02`	[d , e]
(c)	`parenthesis_02_none_alpha_01`	[c]
(f)	`parenthesis_02_none_alpha_01`	[f]
(g - i)	`parenthesis_02_dash_alpha_03`	[g , h , i]

refers to the opening punctuation, `02` signifies punctuation both before and after the internal character, `and` is the internal delimiter between characters inside the punctuation, `alpha` is the internal character type, and the final `02` denotes the number of explicit characters that are present in the token. This token, with `and` as the internal delimiter, only references two subfigures. Tokens that do not have an internal delimiter, or have commas separating each of the characters are interpreted similarly where each subfigure is given in the token explicitly. This is in contrast to cases where the `dash` is the internal delimiter, signifying that subfigures appearing in the interval are implied and should be included when text is assigned to the given token. The **(g - i)** token at the bottom of Table 2.1 is a good example of this, as it actually references subfigures **g**, **h**, and **i**.

Table 2.2. Example of a custom sentence-level regex applied to a POS-tagged caption, and the resulting text strings assigned to each subfigure.

Caption text (custom POS tagging)
`(‘(a) and (b)’, ‘CAP’),(‘TEM images’, ‘NC’),(‘of’, ‘IN’),(‘1.93 wt% Ru{WSe2’, ‘NC’), ...  ,(‘(d, e)’, ‘CAP’),(‘The enlarged area’, ‘NC’),(‘denoted’, ‘IR’),(‘in’, ‘IN’),(‘(c)’, ‘CAP’),(‘corresponds’, ‘IR’),(‘to’, ‘IN’),(‘the HRTEM images’, ‘NC’),...`
Pattern #1:
`("CAP", "!", "NC", "IN", "*", ".")` *What does this mean?* Start with a caption delimiter and do not record any text until a noun chunk is found that contains a proposition immediately after. From there, include all text until a full stop is detected
Text assigned to (a and b) from Pattern #1
(a) TEM images of 1.93 wt% Ru-WSe2 (b) TEM images of 1.93 wt% Ru-WSe2

From here, *subfigure tokens* become the anchor points around which caption text is distributed. To distribute the caption text, the sentence is again tokenized, but this time by assigning the proper part-of-speech (POS) to each word. This is a common preprocessing step in NLP pipelines referred to as POS tagging. For this custom implementation of POS tagging, *subfigure tokens* (`CAP`) are identified alongside standard POS tags, such as singular noun (`NN`), verb (`VB`), preposition (`IN`), *etc.* In addition to the custom tagging of *subfigure tokens*, noun phrases are further consolidated into noun chunks (`NC`), and a tag for internal referencing is added, which marks an instance where a *subfigure tokens* is mentioned within the scope of another. With the caption now a sequence of customized

POS tokens, each sequence is matched against a dictionary of common sentence patterns to find all the text associated with the discussion of a given subfigure. In addition to looking for sequences of POS tokens, these patterns utilize include all “*”, and exclude until “!” wildcards, and a general maximum count on the number of *subfigure tokens* in a given sentence as a means to achieve both flexibility and specificity when assigning text to a subfigure. Consider the POS-tagged caption text in Table 2.2. Pattern #1 is provided as an example of an encoded sentence structure that is common in figure captions. The interpretation of the encoded pattern itself is provided in Table 2.2. The text surrounding the **(a) and (b)** token fits this pattern, and as a result, is parsed accordingly.

2.4.2. Statistical Topic Modeling

Caption assignment is based heavily on POS tagging, however, by itself, POS tagging is not sufficient for understanding what is actually being described by the text. One way to start understanding the relationship between individual words and the main idea of a body of text is through a technique called topic modeling. Topic modeling is a statistical methodology that clusters documents into groups, such that documents within a given group are related by common themes, or “topics”, and would be considered similar to a human observer. The field itself can be traced to a few seminal works that approach the general problem of trying to automatically organize, summarize, and understand large electronic archives [120–123]. In this work, topic modeling is used to further enhance the descriptive power of an image’s assigned caption text, by providing a more general context for the presence of the caption text. For example, a caption might only discuss how a material functions as a “highly-promising anode material because it is energy dense”.

The context of this being in reference to “Li-ion batteries” is an example of the type of description that might evolve for this image if topic modeling is considered.

Latent Dirichlet Allocation (LDA) [123], the primary method considered for topic modeling in this work, involves parameter optimization associated with two important representations: (1) a “topic”, which is a multinomial probability distribution over a collection of words, and (2) a “document”, which is modeled as a collection of topics. LDA uses a maximum likelihood approach to provide estimates of two parameters that address both “topic” and “document” probabilities, respectively. These parameters happen to be distributed according to a Dirichlet distribution, part of the technique’s namesake, which rather than sampling from the space of all real numbers, it is sampling over a probability simplex. By fitting these parameters to the document set, the topic distributions and are learned from data. One of the main challenges of this technique from a generalization standpoint, is that LDA demands that the number of topics be specified *a priori*. The number of topics is often only determined after utilizing a combination of human assessment and statistical tools. Jelodar *et al.* [124] provide a useful overview highlighting the multitude of ways in which LDA has impacted sematic data mining for information retrieval, social media analysis, recommendation systems, *etc.*

In this work, LDA is applied to the abstract texts of scientific journal articles using gensim [124], a popular open-source NLP library designed specifically for topic modeling. Abstracts are useful for topic modeling because the information they provide is highly contextualized, and each abstract tends to cover a singular topic. Models developed in this study are trained on a corpus of nearly 15,000 abstract texts related to nanomaterials, with both short abstracts and under/overused words filtered out to avoid reinforcing inaccurate word associations in the language model. This is not an uncommon preprocessing step for NLP. As further preprocessing, all of the documents in the corpus are then transformed into bag of words (BoW) vectors, which are vectors that represent the number of times a word within the predefined vocabulary appears in an image. After this, the BoW vectors are fed into the LDA model, along with indication of the number of topics. Once a model is trained, it can be used to infer topic distributions on unseen text. If a name is given to a particular topic (either manually or by some automated assignment approach) and an image is associated with the text of the document, then this can provide additional contextual annotation to images. The process used to automatically assign a single word or short phrase to describe each topic is discussed in the next section.

2.4.3. Neural Network Word Embeddings

LDA topic models are constructed from large amounts of raw, unannotated text that have been vectorized using a BoW approach. The BoW approach provides perhaps the simplest form of numerical text representation: a vector of word counts. One notable drawback to this technique is that word order is ignored, which can have a significant influence on the meaning or expectation of the sentence. Interestingly, the development of what is now a popular word embedding technique, “Word2Vec” [125], showed that the word vectors obtained from training a simple neural network to predict single missing words from surrounding context, or vis-versa, had what seemed like almost magical powers: semantics were captured in mathematical manipulations of the vectorized words. The seminal example is that king – man + woman gives a vector answer that is very close to the vector representation for the word queen [126]. The “Word2Vec” style embeddings have even been extended to the materials science domain. Following the process outlined in "Word2Vec" [125], Tshitoyan *et al.* [127] show that the king/queen-style arithmetic example has an analogous counterpart in the materials science domain: ferromagnetic − NiFe + IrMn antiferromagnetic.

The specific use case of "Word2Vec" in this study involves a computation of the cosine similarity between a collection of embedded words to all the embedded words contained in the predefined corpus. In particular, the LDA model infers topic distributions on documents, where each topic distribution is a linear combination of keywords. These keywords form a collection of words that are averaged and compared to all words in the corpus. The top-N most similar words, based on "Word2Vec" embeddings, provide informative labels to describe the LDA topic for the text. This is powerful because when the text can be explicitly assigned to a given image, the semantics implied by not only the paper, but the several thousands of papers in the field that are part of the corpus, help label this image. The downside to this is that there is a tendency to miss novelty as the words returned will bias contexts where the specific word is usually found, as opposed to the actual specificity described in the paper. For example, nanoparticles are commonly used in papers describing possible cancer treatments. If this represents the most prototypical use case for that particular word in the corpus, then all is fine. If the paper is talking about nanoparticle processing for optical applications, it is straightforward to see how this all can become misleading. That is why it is extremely important to view both LDA and "Word2Vec" as mechanisms for providing informed context to automated labeling approaches, but also to be aware of their scope and limitations.

CHAPTER 3

Quantifying Image Similarity for Automation in STEM

The task of quantifying image similarity between simulation and experiment is an essential component of automated microscopy interpretation and analysis. This work uses an extensive dataset of simulated atomic-resolution microscopy images to evaluate the effectiveness of common preprocessing techniques alongside both direct-pixel and feature-based image similarity measurements. The dataset, titled atomagined, is open source (https://doi.org/10.18126/szeq-yde5) and contains > 200,000 simulated images with added synthetic noise to emulate distortion conditions common to the STEM imaging mode. The choice of preprocessing techniques and image similarity measurements are important prerequisites for automation in microscopy interpretation that is commensurate with current hardware, simulation, and acquisition process improvements.

3.1. Introduction

With the arrival of aberration-corrected electron optics, transmission electron microscopy (TEM) including scanning transmission electron microscopy (STEM) have become indispensable tools for nanoscale materials analysis. TEM/STEM images provide an unprecedented view of both the static and dynamic structure of matter with single-atom sensitivity [12, 128–131]. However, in many cases, the image itself is only the starting point into quantitative structural insight. It is often necessary to consider comparisons with simulated images to determine how experimental contrast should be interpreted [132], and an enhanced capacity to collect and store atomic-resolution images has only exacerbated the severity of the bottleneck faced when the comparisons are manual.

Initial efforts to establish automation in microscopy analysis have focused on the development of image features to capture the shape, location, and overall appearance of pertinent aspects of the image [133, 134]. DeCost *et al.* [72] extended this feature development approach by applying a "bag of visual words" paradigm [135], to classify microstructure image data. Somnath *et al.* used a combination of singular value decomposition (SVD) denoising with pattern matching to identify individual atomic coordinate positions and defects [136]. Recently, convolutional neural networks (CNNs) [137], trained to identify local chemical and structural states in unprocessed STEM images [76, 138], have been combined with various object detectors to create automated defect recognition tools [77]. While these studies have all advanced the status quo for feature recognition and image classification in atomic-resolution microscopy, they have not explicitly addressed the role that various preprocessing, image similarity measurements, and distortion types play in the task of measuring image similarity.

In general, there are two main approaches to measuring image similarity: (1) direct-pixel and (2) feature-based approaches. Direct-pixel approaches quantify similarity pixel-by-pixel. For feature-based approaches, the image is processed as a set of feature vectors, and comparisons are performed in feature space. With recent improvements in the identification of atomic coordinate positions [139], feature-based approaches that make use of coordinate and intensity information are of particular interest as a natural extension to existing methods of automated analysis which capture and process these positions [77].

Here, we evaluate several direct-pixel similarity measurements, alongside a feature-based approach, on an extensive dataset of simulated high-angle annular dark-field (HAADF) STEM images. This dataset decouples (image) acquisition-related noise components from local structure changes, making it useful for testing the susceptibility of preprocessing methods and similarity measurements to conditions frequently realized in experimental imaging. Matching accuracy, ranking accuracy, and feature-based representations are discussed in relation to the development of automated search and comparison tools.

3.2. Methods

3.2.1. Image Similarity Measurements

The goal of measuring image similarity is to be able to quantify the visual likeness between a pair of images. Throughout this study, image similarity is reported as a distance. This means that small distance values indicate the highest-degree of similarity, with zero distance indicating a perfect match. The upper-bound on the distance (when it exists) depends on preprocessing and often on image size. In the simplest form, pixel values are considered explicitly (pixel-based approach) and the measure of likeness is a calculation of the mean of the squared differences (*i.e.*, errors) between pairs of corresponding pixels. This mean squared error (MSE) is a popular similarity measurement because it satisfies convexity, differentiability, and is computationally fast. However, the strength of the error signal is often not sufficient for quantifying image likeness in accord with human perception. To this end, a method such as the structural similarity index measure (SSIM) [85] is explored. SSIM processes spatially proximate pixels as patches and quantifies the likeness between patches based on a multiplicative combination of patch statistics (*e.g.*, luminance, contrast, and structure). This work leverages the scikit-image [140] implementation of SSIM, which is tuned to measure the mean of the scores between corresponding 39×39 pixel patches ($\sim$3.9Å × 3.9Å).

The final pixel-based approach explored is visual information fidelity (VIF) [141]. VIF is considered an information-theoretic method; *i.e.*, it is rooted in classical information-theoretic concepts such as information entropy, mutual information, *etc.* Specifically, VIF represents the ratio of two information measures, one that attempts to quantify the loss of information solely to the human visual system (HVS) in the reference image, and the other which attempts to quantify the loss of information due to distortion (in the distorted, or corresponding image being compared). This measurement uses a combination of multiscale decompositions, distortion models, and models of the HVS. Figure 2.3 provides a concise system level overview of the VIF measurement. The pixel-domain variant [141], provided by the authors of the original VIF measure, is referred to as VIFP. Refer to [85] and [86] for more theory behind SSIM and VIF, respectively. Registration for all direct-pixel comparisons is performed using a combination of iterative rotation with normalized-cross correlation template matching in scikit-image [140].

To contrast the direct-pixel image comparisons, we engineered a feature-based approach based on nearest neighbor distance and integrated intensity for each atomic column in an image. Feature-based comparisons have an advantage over direct-pixel comparisons in that registration is not a prerequisite for matching. First, to transform an image into a set of atomic column-based features, the (x, y) coordinates for all of the atomic columns are recorded. For each of the P columns in the image, the $(k-1)$ nearest neighbor column distances are arranged in ascending order, and every entry within the vector is normalized by the k^{th} entry. In this form, a 2D $M \times N$ image becomes a $P \times (k-1)$ collection of features, where k is set to 25 for testing. This collection of image features is the S_X in Eq. 2.15 (S_Y is constructed from the image being compared to S_X).

This work relies on a technique called earth movers distance (EMD) to compute image similarity between image features (Eq. 2.16). Conceptually, EMD quantifies the minimum "cost" to transport the weights of a distributing set of features into a receiving set, based on the location of the distributing features, and the relative distance and capacity of each receiving features to store a portion of the weight. Specifically, EMD measures the minimum distance between the two collections, allowing weights associated with each feature to be permuted and partially distributed. For this study, distances between features were defined in "city-block" space [142] (*i.e.*, movements are constrained along one dimension of space at a time when transporting weights). The integrated intensity of an atomic column is the "weight" assigned to the feature. In general, the (x, y) coordinates anchoring the positions of the features are not limited to intensity peaks. Other objects such as diffraction peaks or particle centroids, could be represented in feature spaces, as long as it their coordinates can be extracted consistently across images.

3.2.2. Image Preprocessing

Raw imaging data is highly susceptible to missing values, noise, and inconsistencies in the acquisition process. In addition, without calibrating the detector used to form the image, the true values from experiments are unknown. As a result, preprocessing is important for improving data quality, and ensuring that conclusions drawn from comparisons characterize the data accurately. Two common preprocessing techniques explored are normalization and standardization. Normalization transforms the min/max values of the numeric range to a scale between 0 and 1. In figures, we refer to normalization by specifying the "[0,1]" bounds it enforces. Standardization, on the other hand, involves rescaling the collection of numeric values to have a zero mean and unit variance. In figures, we refer to standardization using "[μ,σ]" to represent the mean/variance criteria it enforces.

3.2.3. Dataset Construction

Structure Prototypes and Supercells. Crystal structure data for inorganic compounds were obtained via the Inorganic Crystal Structure Database (ICSD) [143]. 500 unique structure prototypes were selected at random from entries in the ICSD and were rotated 4 times such that the resulting viewing angle coincided with a unique projection along a random integer combination [uvw] direction (integers are restricted to a range [-2,2]). The result is a collection of 2,000 distinct oriented supercells (composition + viewing angle) with dimensions 25 × 25 × 25 Å. The chemistry of the structure prototypes, symmetry information, *etc.*, are all summarized as part of the full *atomagined* dataset hosted by the Materials Data Facility (MDF), and are accessible via [144].

Image Simulations and Synthetic Distortions. Each oriented supercell is the input to HAADF image simulation performed using the Prismatic simulation code [39, 41]. The output 3D scattering signal, is integrated radially within a range of 100-150 mrad, to create a virtual detector image. Three benchmark conditions, namely *pristine*, *clean*, and *experiment* (presented in Figure 3.1), represent the output of a perfect image simulation, a proxy for a denoised experimental image, and a proxy for a raw experimental image, respectively, and are used to test combinations of preprocessing and similarity measurements.

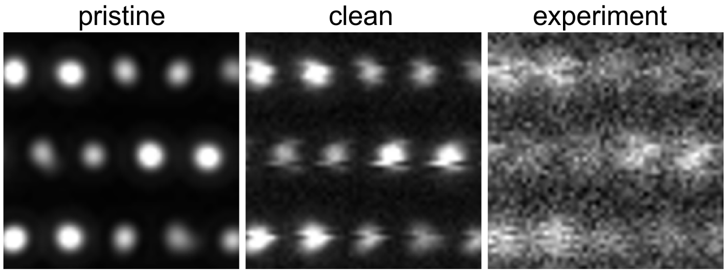


Figure 3.1. The *pristine* condition is a raw STEM image simulation, and the *clean* and *experiment* conditions refer to simulations with increasing amounts of postprocessed distortion. The *clean* condition mimics a denoised experimental image, and the *experiment* condition mimics a raw experimental image, without noise reduction.

Several different classes of *distortions*, applied to images of structures present in the benchmark conditions, are used for image similarity measurements. Here, *distortion* is a general term used to describe noise applied to the image simulation process ("image distortion") and/or the input structure used as input to the simulation ("input structure distortion"). The distortion types are illustrated in Figure 3.2, and were chosen to mimic some of the most commonly observed distortions in HAADF STEM images (see Section 3.5.2). The *background*, *shot*, and *blur* distortions in Figure 3.2 are all considered image distortions because they affect the size, shape and/or relative position of the pixel intensity distribution. The *background* distortion mimics STEM imaging under the condition of long probe tails, and is applied by adding a constant value to all pixels in the image and clipping on the high intensity end. This compresses the pixel intensity distribution (shifts the mode) in the direction of brighter pixel values. In most atomic-resolution HAADF STEM images, a majority of pixels in the image belong to the "background", thus the position of the mode (the most frequently occurring intensity value) is an important feature of this distortion type. The *shot* distortion adds signal-dependent Poisson noise, applied

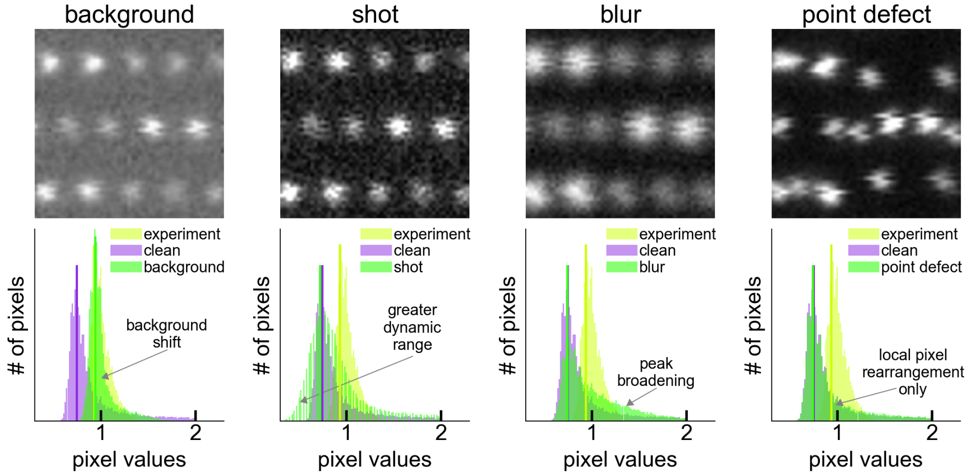


Figure 3.2. The image panels show examples of the *background*, *shot*, *blur*, and *point defect* distortion types. At the bottom of each panel is a histogram illustrating how the distortion concentration alters the distribution of pixel values from the *clean* and *experiment* benchmark conditions.

as "electrons per probe" dose. Higher levels of distortion, corresponding to smaller doses, produce increasingly discretized images with larger dynamic ranges (*i.e.*, fewer distinct pixel values in an image when dose is smaller, divided by dose, will have less unique pixels values that span a greater total range). The *blur* distortion emulates the spreading of the atomic column intensities from beam broadening and/or source size effects. This increase in the pixel intensity values in the vicinity of the atomic columns is visible in the pixel intensity distribution as a heavier right tail. Finally, the *point defect* distortion, generated with the same image postprocessing conditions as the *clean* benchmark, is applied as a localized structure modifications to the input supercell (*i.e.*, an "input structure distortion"). Moreover, despite visible degradation of the local geometry with increasing levels of *point defect* distortion, the distortion is a local pixel rearrangement, therefore the pixel intensity distribution is unaffected.

Dataset Splits. When two images are compared, the *target* image is considered to be the image that is fixed, or rather, is the image being matched to. This is consistent with the mention of *target* images in the descriptions of image matching from Section 2.3.1. Because the *target* image designation can change depending on the specific scenario, the dataset is split according to the benchmark and distortion classifications. The split for the benchmark class contains a total of 6,000 images, which is formed from the 2,000 unique oriented supercells simulated to include the *pristine*, *clean*, and *experiment* benchmark conditions. The distortion split of the dataset contains a total of 4,000 images constructed from 400 unique oriented supercells selected at random, simulated with the four distortion types (Figure 3.2) at two concentration levels, and an additional condition that combines the *experiment*-type benchmark image with two concentration levels of the *defect*-style distortion ($400 \times [(4 \times 2) + (1 \times 2)] = 4{,}000$). An exhaustive comparison between all images across both the benchmark and distortions splits involves a total of $6{,}000 \times 4{,}000 = 24{,}000{,}000$ comparisons.

3.3. Results and Discussion

We examine a series of relevant scenarios that address the suitability and/or capability of the preprocessing and image similarity measurements introduced: (1) constrained image matching (*i.e.,* a database search), (2) structure and parameter refinement, and (3) feature-based matching and clustering.

3.3.1. Constrained Image Matching and Retrieval

For constrained image matching, images from the *pristine* and *experiment* benchmark classes were registered, separately, against all images belonging to a given distortion class. The distortion class image with the smallest similarity distance was considered a "match" to the given benchmark *target.* This search was repeated for each different class in the distortion split. If the match contained an identical structure as the *target*, it was considered a true positive (TP) match. Naturally, a false positive (FP) is then an image of the incorrect underlying identified as the match. Matching is considered "constrained" in the sense that a TP match exists for each *target*, and it implies that the comparisons are made between images that have been registered, or can be registered in a straightforward manner, using scaling metadata with out-of-the-box registration algorithms. In other words, it is an attempt to ensure high matching quality and eliminate potential artifacts of sub-optimal registration.

The results of this constrained image matching depend heavily on the choice of preprocessing technique and image similarity measurement (and much less on the specific distortion class). For this reason, we do not further distinguish the specific distortion class for each *target*, but rather, present the results across all distortion classes for a fixed

Table 3.1. Constrained image matching results across all distortion classes, grouped according to preprocessing technique within a fixed *pristine* and *experiment* benchmark class.

	pristine		experiment	
	[0,1]	[μ,σ]	[0,1]	[μ,σ]
MSE	0.79	0.99	0.04	**0.99**
SSIM	0.81	0.99	0.65	**0.99**
VIFP	**1.00**	N/A	**0.99**	N/A
EMD	0.79	0.90	0.89	0.91

benchmark class. Table 3.1 shows the matching accuracy scores for the task (*i.e.*, the ratio of TP matches found for each pristine and experiment *target* to the whole population of *targets*). Again, the matching choices for each *target* were restricted to all images within a given distortion class. VIFP appears to be a robust option for HAADF STEM image matching, resulting in the highest matching accuracy scores for both the *pristine* and *experiment* images as *target*, and standardization (where applicable) enhances matching accuracy across the board – though it is likely not the most intuitive choice for preprocessing because it introduces negative values into a construct (imaging) which lacks definite meaning or interpretation for values in this range. With standardization, traditional, direct image similarity measurement techniques such as MSE and SSIM are competitive with an information-theoretic method such as VIFP which employs mathmatical models of the HVS as a component of the measurement.

It seems the success of standardization over normalization for MSE and SSIM may be rationalized using the concept of Weber (object) contrast [145] (*i.e.*, the ratio of the object/background luminance difference to background luminance). Weber contrast provides a way to quantify the visibility of the object in the image, relative to the background. Empirically, we observe that the value of the numerator in the Weber contrast measurement (one way to enhance the overall contrast) tends to improved with standardization. We leave further quantitative assessment of Weber contrast an avenue for further exploration because, strictly speaking, the extension of standard contrast measurements to accommodate negative background luminance is non-trivial, and negative background luminance is exceedingly common across in many of the standardized HAADF STEM images. The improvements with standardization are not as drastic for EMD and the above hypothesis is generally less applicable because only groups of raw pixel values for the columns themselves are considered (*i.e.*, integrated intensity) in the measurement. It is apparent, at least for these trials where registration is easily solved, that the feature-based EMD comparisons are not practical for such detailed comparisons without fine-tuning - that is, without further tuning some of the hyper-parameters associated with the construction of the features (*e.g.*, number of nearest neighbors, choice of weights for each feature, *etc.*).

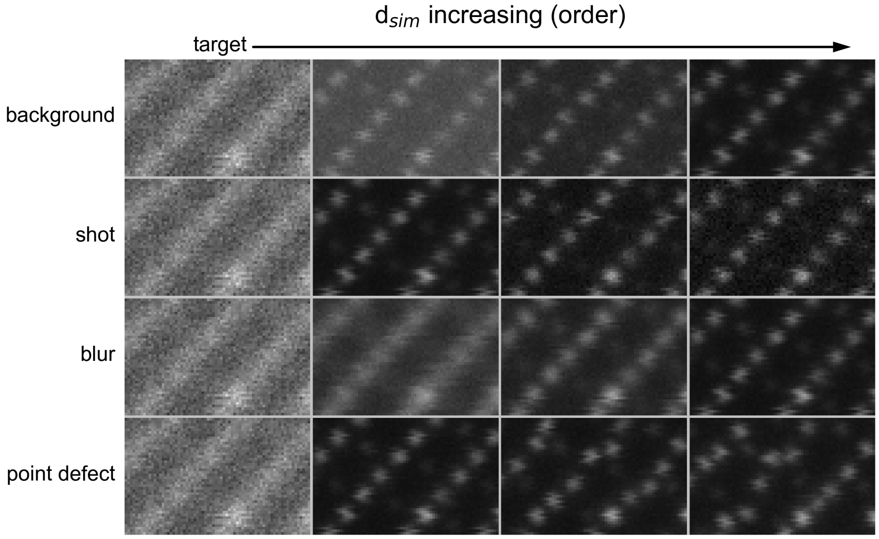


Figure 3.3. Grid showing the anticipated ordering of the similarity distances for each distortion class, relative to the *experiment* benchmark (leftmost column). Traversing the grid from left-to-right within a row (distortion class) provides a visualization of increasing similarity distance.

3.3.2. Structure and Parameter Refinement

For structure and/or acquisition parameter refinement, *structure* refers to the oriented supercell that is the input to an image simulation, and *parameters* are experimental acquisition parameters or forward model parameters that can be used in simulation to recreate aspects of an observed image. This scenario assumes that the imaged structure is mostly known, and the purpose is to quantify the extent to which parameters of an image simulation, microscope acquisition conditions, or small changes to the input structure file used simulated an image, impact certain quantitative aspects of image similarity. For example, to better characterize the thickness of an imaged structure, one approach is to

Table 3.2. Ranking accuracy results, grouped according to preprocessing technique within a fixed distortion class.

	background		shot		blur		point defect	
	[0,1]	[μ,σ]	[0,1]	[μ,σ]	[0,1]	[μ,σ]	[0,1]	[μ,σ]
MSE	0.67	0.01	0.02	0.69	0.71	0.95	0.32	0.92
SSIM	0.93	0.01	0.00	0.73	0.98	0.98	0.56	**0.97**
VIFP	**1.00**	N/A	**0.93**	N/A	**1.00**	N/A	0.95	N/A
EMD	0.78	0.43	0.01	0.15	0.57	0.35	0.90	0.95

sample intensity configurations produced by simulating images at varying thicknesses with lateral positions fixed, and determine which thickness produces column intensities that are most consistent with the observed experiment. If aspects of the structure are in question (*i.e.*, the interface region of a grain boundary, or the structure in the region surrounding a point defect), then sampling structural configurations to match an experimental image is necessary, where both lateral positions of the simulated structure, and thickness is varied. In both cases, the similarity measurement must adequately quantify what are often visually subtle changes in the statistics of the image, in order to rank the image appropriately against several other images containing similar subtle changes.

Figure 3.3 demonstrates the anticipated ordering of image similarity distances for a comparison with the *experiment* benchmark. The images in the leftmost column are the experimental benchmark *targets* (duplicated for visual purposes) and the subsequent columns to the right contain images of the same structure with progressively higher similarity distance (organized to show a different distortion class across each row). If the image similarity distance between the left panel and the successive panels in a row is strictly increasing, then the individual trial is a success. Ranking accuracy, or the ratio of

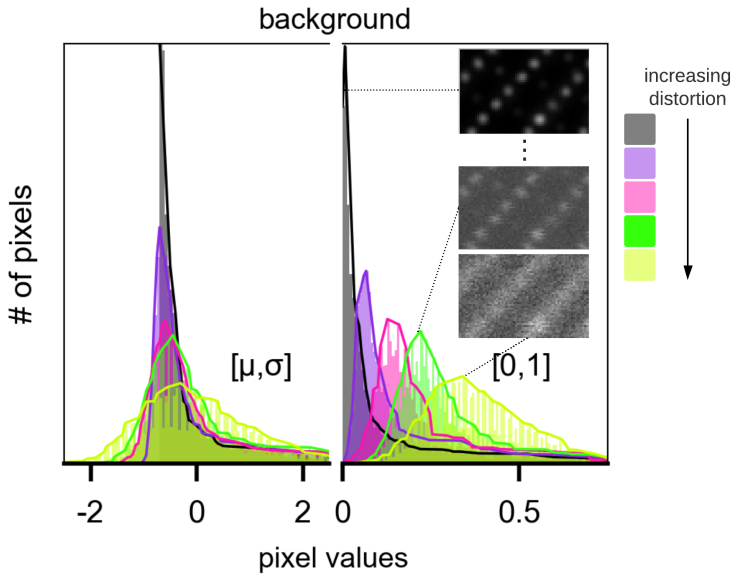


Figure 3.4. The effects of normalization and standardization on the pixel intensity distributions provide a means to begin rationalizing how preprocessing has an effect on ranking accuracy. Here, standardization reduces the distance between the modes of dissimilar distributions, which decreases the likelihood of ranking success across a given trial.

"success" trials, to the total number of attempted trials, is used to quantify performance on this ordering task.

Table 3.2 provides the ranking accuracy (by distortion class) for each image similarity measurement and the associated preprocessing techniques with the *experiment* benchmark as the *target*. Overall, normalization results in lower MSE and SSIM ranking accuracy compared to standardization, except for in the case of *background* distortion. Figure 3.4 illustrates how normalization, applied to images of the dataset with different levels of background noise, reinforces separation and proper ordering of the histogram modes, whereas standardization appears to collapse the positions of the modes (or at least reduces the distance between the modes of dissimilar distributions), which to some extent, negates the influence of this important feature on the overall similarity measurement. As mentioned in the discussion of Figure 3.2, it is the position of the mode that encodes the amount of this particular “background” distortion, so it follows that the relative repositioning of the modes due to preprocesssing is likely problematic. This is similar to performance decrease for shot distortion ranking accuracy with MSE and SSIM when normalization is used, that is to say that the distortion amount is related to the dynamic range of pixel values (the way it is applied), and normalization negates this range feature.

Overall, VIFP achieves the highest scores for all distortions affecting the the structure of the actual image (*i.e.*, *blur*, *shot*, *background*) and not the imaged structure, as in the *point defect* condition. For the *point defect* condition, SSIM provides the greatest accuracy when standardization is used, and the feature-based approach with EMD is competitive (tying the VIFP ranking accuracy). The strength of the feature-based approach in this scenario is in its ability to achieve commensurate ranking accuracy for the *point defect* class (the kind of distortion that it is designed for) without registration. This has interesting implications not only for situations where registration is non-trivial, but also for clustering and partial matching – topics explored in the next section. There are also differences to consider in the comparison techniques when it comes to profiling real time execution (refer to the supplementary information – Section 3.5.1).

3.3.3. Feature-Based Matching and Clustering

Throughout the previous scenarios, a feature-based image representation with EMD has shown the capacity for effective similarity comparisons between images with different concentrations of point defects, but subpar matching and ranking accuracy performance across the other distortion conditions tested suggests that it is not the most sensible choice when registration is tractable. However, to the extent that registration is non-trivial (*e.g.,* it is not uncommon for atomic-scale images to contain regions of a local nanostructured phases, grain boundaries, or dislocation cores, *etc.*), the "in-the-wild" search capabilities that feature-based representation with EMD offers appears promising for matching based on partial similarities (partial-matching), and/or for clustering similar atomic columns based on relative intensity of the columns and/or local coordination. The scenario involving twin boundaries in [110] CdTe (Figure 3.5a) underscores the value of feature-based approach for partial-matching. In the absence of an exact matching twin boundary structure in the dataset, the ideal TP match is the image containing the correct corresponding bulk structure. In this situation, feature-based representation with EMD is the only comparison method successful in immediately identifying the TP match, and beyond this, a similar diamond cubic Si structure that was present in the dataset. The [110] CdTe was eventually returned with VIFP and SSIM comparisons, but it was not considered the actual closest match.

Additionally, making use of the fact that each image is transformed into a set of features before an image similarity measurement with EMD, provides for a way to cluster the atomic columns into structurally, and sometimes chemically, meaningful groups. For example, Figure 3.5b highlights snapshots of three different oriented supercells. In the

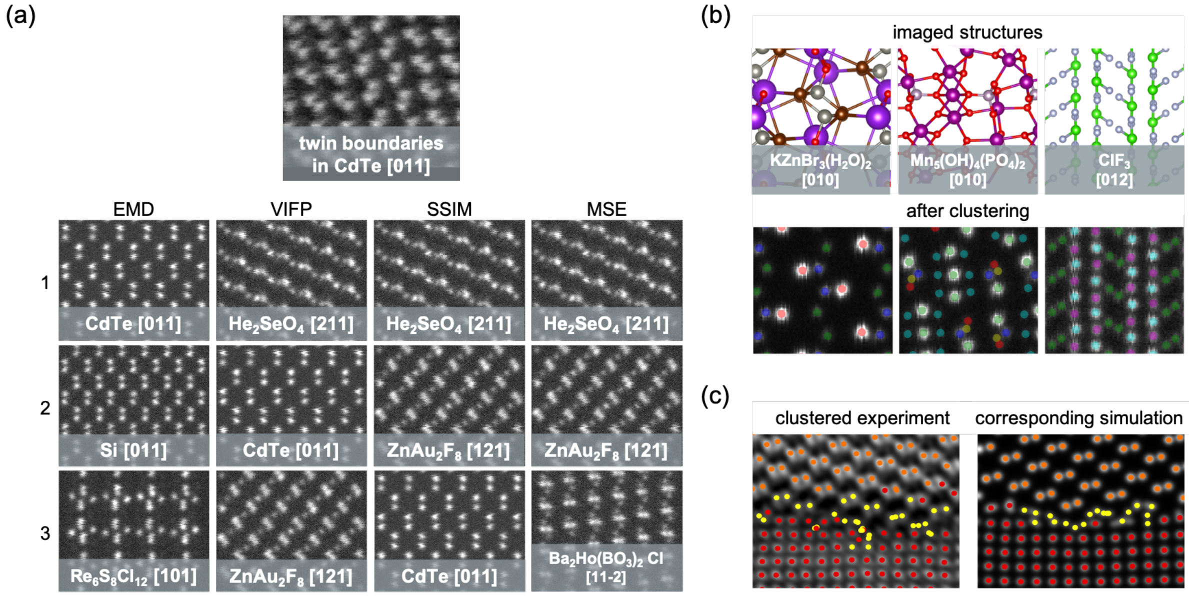


Figure 3.5. (a) The image grid shows the first three images returned (reading down each column) for the image similarity measurements tested when the twin boundary image is the target. The feature-based approach with EMD has the useful characteristic of being able to identify structures that are not identical but contain similar parts (first image), and match across different scales without scaling or alignment between the images being compared (second image). The pixel-based approaches show some success in finding the associated bulk structure, but appear more vulnerable to FP's that show similarity on a pixel level, but do not correspond structurally to the target. (b) Clustering with the "atomic column" features shows promise in its ability to group similar atomic columns in a chemically and/or functionally logical way, and also in (c) identifying useful structural correspondences between experiments and image simulations.

complex bromide compound (left image of Figure 3.5b), each atom type that gives rise to a distinct atomic column in the image is given a separate group (color) when clustered in the feature space. For the phosphate compound (center) and chlorine trifluoride (right), atoms of the same chemical species are even further divided into separate groups based on their local coordination. The images in Figure 3.5c illustrate the clustering of an experimental image alongside the clustering of a proposed simulation with the columns assigned to groups consistently across the images (*i.e.*, top columns belong to the bulk of grain 1, bottom columns belong to the bulk of grain 2, and columns in the middle are part of the interface). This sort of clustering is useful because it facilitates a degree of automation in the comparison of specific regions of an image, enabling one to isolate critical regions (*i.e.*, carry out structure-based image segmentation), and compare images on the basis of these regions. In other words, the value of developing feature-based comparisons is rooted in their flexibility to function as both a means for useful image comparison across scale and orientation (when used with EMD), as well as a mechanism for automatically clustering atomic-resolution images into chemically and/or structurally relevant parts.

3.4. Conclusion

The present study demonstrates the effectiveness of various image preprocessing and comparison techniques in context of a large simulated dataset of atomic-resolution HAADF STEM images. This is accomplished by exploring a series of relevant comparison scenarios: (1) constrained image matching, (2) structure and parameter refinement, and (3) feature-based matching and clustering. First, in constrained image matching, VIFP with normalization achieves the highest matching accuracy scores for both the *pristine* and *experiment* images as *targets*, but MSE and SSIM with standardization are also competitive, achieving equivalent (or near equivalent) matching accuracy. Though standardization is not the most intuitive preprocessing technique for images because it introduces negative pixel values into the imaging construct, for the dataset tested, it proves especially effective for comparisons with MSE and SSIM. The magnitude of the pixel-wise error and the concept of object contrast, both affected by standardization, provide likely explanations for the enhanced MSE and SSIM performance. Second, for structure and/or acquisition parameter refinement scenarios (*i.e.*, sampling different intensity or structural configurations), VIFP again achieves the highest ranking accuracy scores for all distortions that affect the specific pixel values encoding a fixed structure (*e.g.*, *blur*, *counts*, *background*). When the structure being imaged (simulated) is modified and the difference in the actual local atomic structure is the feature to be quantified in the image comparison, SSIM with standardization as preprocessing provides the greatest accuracy, and the feature-based approach with EMD is competitive (tying the VIFP ranking accuracy). Finally, we highlight some of the advantages of using a feature-based image representation with EMD as both a means for matching based on partial similarities, or for clustering similar atomic columns based on relative intensity of the columns and/or local coordination.

A quantitative image similarity measurement is the missing link to establishing automated interpretation strategies requiring comparisons (*i.e.*, comparison between regions within an image, across frames, to physics-based simulation, to images published across literature, *etc.*), and all image similarity and preprocessing methods tested have certain conditions under which their performance is optimal or near-optimal. It is therefore important to understand how a given situation mimics one of the comparison scenarios outlined in order to decided which comparison and preprocessing method most appropriate.

3.5. Supplementary Information

3.5.1. Timing Considerations

Table 3.3. Profile of execution time for image comparison techniques and registration.

comparison/operation	MSE	SSIM	VIFP	EMD	registration
average time (sec)	1.31E-4	4.62E-3	1.75E-2	1.96	1.98

Timing (*i.e.*, profiling real time execution) - recorded in Table 3.3 - is another important consideration in the assessment of image comparison techniques. The times recorded represent an average time for each operation on a single thread of an Intel Broadwell processor, taken after a minimum of 4.5 million measurements. Under these conditions, the current VIFP implementation is ∼3.7x slower and ∼133x slower than SSIM and MSE, respectively. Depending on the size of the dataset over which the comparisons are made, this could have significant timing consequences. Another consideration is that a full direct-pixel comparison that requires registration is on the same order of magnitude (timing-wise) as a comparison in feature space using EMD. The time recorded for EMD assumes that the average set of features for an image are 343 x 24, which means that a typical image contains 343 atomic columns and is represented in nearest neighbor space looking at the 24 closest neighbors. The time recorded for EMD does not include the time it takes to construct the features. This can be done as a separated process offline. Any significant deviation from average number of atomic columns will have an impact on the timing as the measurement can be very slow (exponential worst case complexity), especially for image with many atomic columns.

3.5.2. Image Simulation and Synthetic Distortion

The approach for creating idealized STEM noise and distortions was originally developed by Colin Ophus in MATLAB. The Python version, `STEMnoise.py`, is available as part of the *atomagined* dataset [144]. In addition to a raw image simulation, the input values given in the description of each distortion, can be used in `STEMnoise.py` to replicate the distortion conditions tested in the *atomagined* dataset. The '–1' or '–2' appended to the distortion name indicates a lower or higher concentration of the specified distortion, respectively.

Table 3.4. Distortion condition descriptions with corresponding `STEMnoise.py` inputs at low and high concentration levels.

distortion name	description	`STEMnoise.py` input array (1: low conc., 2: high conc.)
pristine	raw simulation for pristine structure	$\emptyset$
clean	trace amounts of distortion (*e.g.*, denoised experimental image)	[0.8, 400, 0.75, 2]
experiment	higher concentration of distortion (*e.g.*, raw experimental image)	[2.8, 400, 1.25, 10]
blur	source size broadening effect with Gaussian blur	1: [1.8,400,0.75,2] 2: [2.8,400,0.75,2]
shot	noise to capture Poisson characteristics of the signal	1: [0.8,100,0.75,2] 2: [2.8,40,0.75,2]
background	constant additive background (from long tails of STEM probe)	1: [0.8,400,0.75,6] 2: [0.8,400,0.75,10]
point defect	point defect in center of view, + local relaxation around defect	1: defect-1 + [0.8,400,0.75,2] 2: defect-2 + [0.8,400,0.75,2]
combination	combination of multiple distortions	1: defect-1 + [1.8,100,1.00,6] 2: defect-2 + [2.8,100,1.25,6]

CHAPTER 4

Fusing Atomic-Scale Image Simulations into Experiments

To fully leverage the power of image simulation to corroborate and explain patterns and structures in atomic resolution microscopy, an initial correspondence between the simulation and experimental image must be established at the outset of further high accuracy simulations or calculations. Furthermore, if simulation is to be used in context of highly automated processes or high-throughput optimization, the process of finding this correspondence itself must be automated. In this work, we introduce ingrained, an open-source automation framework which solves for this correspondence and fuses atomic resolution image simulations into the experimental images to which they correspond. We describe herein the overall ingrained workflow, focusing on its application to interface structure approximations, and the development of an experimentally rationalized forward model for scanning tunneling microscopy simulation.

4.1. Introduction

Materials image simulations are becoming an integral part in the structural analysis of complex materials systems. Having a three-dimensional atomistic structure of a materials system is valuable, both for understanding and for property prediction through first principles simulations. For scanning transmission electron microscopy (STEM), the electron-matter interactions governing the image formation process are well-codified in numerical "multislice" simulations [37], and the combination of aberration-corrected STEM images with these multislice simulations have been used effectively in a variety of contexts for structural determination with atomic precision [146–151]. Image simulations have also proven useful in scanning tunneling microscopy (STM) in order to help solve for surface structure or adsorption geometries [6, 152–154]. The success of these "simulation-to-experiment" comparisons is in their ability to link information about the underlying mechanisms generating the experimental observation to parameters and/or specific structures used in simulation. However, to utilize simulation for this purpose, a mapping between simulation and the visual or measurable expectation from experiment must be explicitly established (*i.e.*, pixels from one image are mapped to corresponding pixels in another). This process creating this mapping is referred to as image registration.

Image registration workflows are often divided into coarse and fine alignment steps [155]. Coarse alignment typically involves identifying salient points and their correspondences (so-called landmarks) across images. With landmarks in place, a point-set registration algorithm such as the iterative closest point (ICP) [156] ensures that the distance between corresponding landmarks is minimized, and thus roughly aligned. Landmark identification is used in both pycroscopy [157], and the 'TurboReg' plugin for ImageJ [158] (image processing platforms for microscopy) as a recommended initial step before full registration is executed. In many contexts, this landmark selection procedure is manual, so the overall coarse registration is technically considered a semi-automated procedure. There are ways to fully automate the selection and pairing of landmarks borrowing from computer vision (SIFT + RANSAC [159, 160]), but this is generally expensive. In addition to landmarks, intensity correlation is another approach to semi-automated coarse alignment, and fortunately, it does not rely on artificial markers placed in the field of view. In the simplest cases, an approach such as phase cross-correlation [161] can automatically remove translation (and rotation [162]) offset in images collected from pre-aligned sensors, or from images collected in rapid succession as part of an image stack (video).

These are the initial steps taken by the popular SmartAlign [163] tool, which provides general-purpose image processing for atomic-resolution time-series data from STEM.

In cases where registration is cast as a two-step procedure, an affine linear transformation is often sufficient for initial coarse alignment. This assumes that major spatial discrepancies between images can be corrected by a combination of rotation, translation, scaling, and shear. For the final “fine alignment” step, a straightforward intensity-based approach proceeds as an iterative optimization of a “similarity measure” which takes into account the explicit pixel values in each image [164], and in some cases, even subpixel shifts [165, 166]. If the proper alignment cannot be achieved with rigid deformations, a non-rigid registration approach can resolve local discrepancies in image content with a set of local deformations. Non-rigid registration has been used for purposes ranging from scan instability corrections in STEM imaging [167], to registration of MRI brain images to help capture brain shift during surgery [168]. The above examples use registration to compensate for spatial discrepancies that exist between two images, with the assumption that the objects in the images are similar enough to be overlayed on top of each other. Here registration can provide a quantitative measure of how the structures differ, or a means to direct comparison. Contrast this with registration needs for simulated materials characterization images that are used to corroborate experimental finding. With this, the goal often involves more than just a solution for a single spatial transformation, but also, the flexibility to modify the structure and/or imaging parameters to the forward model in the loop (*i.e.*, structure and parameter iteration are valuable additions to the overall registration framework).

In most standard contexts, registration of a simulated materials characterization image with an experimental image is considered separate from the forward simulation itself. This is a sufficient approach when enough is known about the parameters of the forward model to produce a reasonably appropriate image, but this is often not trivial. For example. STM measurements are routinely used to probe thin film surface morphology at atomic resolution. Simulations of STM images from density functional theory (DFT) charge densities using the Tersoff and Hamann approximation [42] often accompany these measurements in order to explore various surface geometries in a systematic way, but can be challenging to match to experiment because the overall appearance of the image is greatly influenced by small changes in parameters such as the charge density value to construct the isosurface or the vertical distance of orbitals below the surface that are considered to be accessible by the STM tip. These parameters are difficult to determine quantitatively from experimental conditions and DFT results, and thus a decoupled forward modeling and registration paradigm involving manual trial-and-error is less than ideal for these characterization techniques. In fact, if one knows all the optimal parameters to create a simulated image that matches an experimental image (*i.e.*, has the correct pixel size, atomic structure, *etc.*), then many of the popular image processing and computer vision tools such as OpenCV and Hyperspy (in addition to others already been mentioned), could be used to facilitate 2D correlation-based registration between a fixed pair of images. But knowing these parameters, or even constructing an initial structural model so that a simulated image can be created, is often not trivial.

In this study we introduce *ingrained*, an automated framework for image registration which allows for the fusion of atomic-resolution materials imaging simulations into the experimental images to which they correspond. The framework is modular, allowing for plug-and-play implementation of forward models for image simulations when an experimental complement exists, and provides tools for programmatic construction of periodic bicrystal interfacial structures from materials database queries. In addition to a framework overview, we outline two valuable use cases for image registration with *ingrained*: (1) an experimentally-informed initial bicrystal structure for further interface structure refinement through heuristic search algorithms (*e.g.* basin hopping or genetic algorithm), or high-accuracy structure refinement with multislice and high-angle annular dark-field STEM comparisons); and (2) an experimentally-rationalized forward model for STM image simulation that involves fine-tuning of imaging parameters. The *ingrained* toolkit has recently been used to support the identification of rectangular hydrogenated borophene – synthesized for the first time – from STM images [7]. Examples and instructions to access *ingrained* are available on GitHub (https://github.com/MaterialEyes/ingrained).

4.2. Methods

4.2.1. Ingrained Framework Overview

The *ingrained* framework requires an experimental atomic-resolution microscopy image as input. Conventional preprocessing operations such as Wiener filtering are often useful for simple restoration of the original image if it has undergone significant degradation during acquisition. The main components of the *ingrained* framework are depicted in Figure 1. The first step is structure initialization, where initial input parameters are

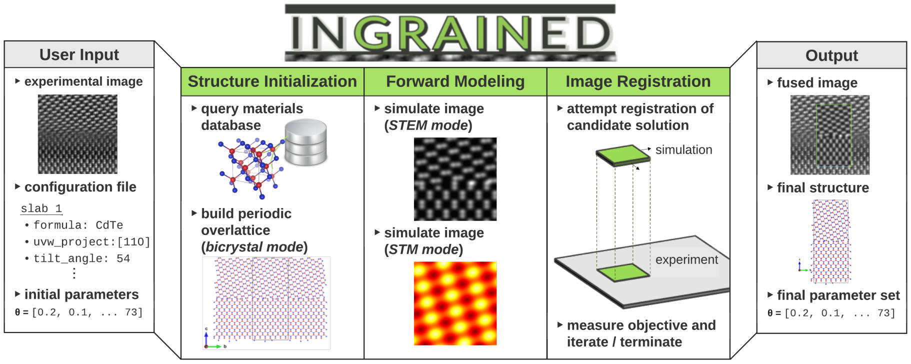


Figure 4.1. The series of panels provides an overview of the *ingrained* framework applied to an experimental interface image. The user input, which comprises an experimental image, configuration file, and initial set of parameters, is processed sequentially by the structure initialization, forward modeling, and image registration modules. The resulting output is a simulation with the best fit inside the experimental image (*i.e.*, the 'fused image'), and a final structure with a parameterized fit-to-experiment.

used to assemble a starting structure (a bicrystal in this instance, but other non-interface structures are possible). Next, a forward model produces a simulated image of the starting structure, which is fused with the experimental image after an iterative optimization procedure which registers the two images. The only manual step in the workflow, outside of potential image preprocessing, involves setting up the parameters and/or constraints used for structure initialization and forward modeling - no landmarks or manual image manipulation are necessary to achieve suitable image registration. The final registration result is a simulated image with a parameterized fit-to-experiment, as well as the structure approximation (*i.e.*, the structure that was used as the basis for the matching simulation). The following sections provide additional input and implementation details.

4.2.2. Structure Initialization

Depending on the nature of the imaged structure and its associated imaging modality, the *ingrained* framework offers two methods for structure initialization. In the simplest case, a database query tool can be used to programmatically download chemical structure files from the Materials Project (MP) database [169] based on chemical formula and space group information provided by the user in a configuration file (JSON format). With this, if the goal is to register a bulk crystalline structure to the experimental image and the user specifies a viewing direction, the structure is also rotated to the prescribed orientation as part of the initialization. The second method of structure initialization involves the automatic construction of bicrystal interfaces. In this mode, the configuration information specifies the composition and viewing orientation for two crystal structure files (grains) in the image, as well as constrain certain dimensions of the "over lattice" constructed around the overall composite bicrystal. In both the single structure query and the bicrystal construction, the user specifies orientation by providing the *uvw_projection* direction (*i.e.*, direction from viewpoint to screen), the *uvw_upwards* direction (*i.e.*, upwards direction on the screen) and the tilt angle, which is a misorientation applied after the previous vector constraints have been satisfied. The required construction parameters include: *min/max depth*, *min/max width*, *interface width*, *etc.* Currently, the size of the over lattice (which defines the simulation cell for the bicrystal) is controlled by these restrictions on the real-space dimensions. Image recognition tools aimed at identifying chemistry, scaling, and orientation information from the images directly is an avenue of further research that could enhance the automation in this step.

Proceeding in this bicrystal mode, the information in the configuration file is used to construct two oriented grains that are combined into a single bicrystal structure, satisfying periodicity conditions through application of uniform strain (*i.e.*, small discrepancies in individual dimension requirements are removed by strain). This is necessary, particularly for interface structures with low symmetry, as it is otherwise intractable to create atomic structures that are small (for computational efficiency in simulations) but remain periodic in at least two dimensions (required for some simulation approaches). The procedure for ensuring periodicity involves estimating repeat length from the grains, and then using the near-coincidence site lattice approach (CSL) with subspace search outlined in Buurma *et al.* [170] to determine the appropriate dimensions for the subsuming over lattice. The bicrystal can then be used in materials modeling context and for image simulation.

4.2.3. Forward Modeling

All the previous configuration parameters are specific to the assembly of the initial structure. Note that this structure initialization feature is not required and can be bypassed in situations where an initial structure and/or partial charge density data (STM simulations) is already provided. The relevancy of this initialization step is entirely dictated by the input structure requirements of the proceeding forward modeling step. In general, the forward model simulates an image from an atomistic structure and requires a set of simulation parameters. These simulation parameters are kept separate from the structure initialization parameters (where applicable) because they are specific to the forward model being implemented.

Currently, the *ingrained* toolkit provides forward modeling options for both high-angle annular dark-field (HAADF) STEM image simulation, as well as for STM. HAADF STEM image simulation is performed as a simple convolution of the atomic coordinates with a point spread function for the microscope, using Kirkland's incostem code [35]. The specific parameters that control image formation based on physical principles (*i.e.*, defocus, sample thickness, *etc.*) are consistent with the parameters discussed in [35]. This convolution approach is convenient because the calculation of an image is performed as a simple multiplication. This provides a tremendous speed advantage over other more quantitatively accurate techniques and in many cases, is capable of capturing many of the same pertinent features [35]. One should note, however, that convolution is in some cases not quantitatively accurate, especially in reproducing the contrast between different elements. The ability of ingrained to distinguish between different structural models is directly limited by the accuracy of image simulations. To improve the accuracy of the image simulations, the multislice or the more computationally efficient PRISM approach can be employed using other STEM simulation codes such as Prismatic [39, 41]. The interface of additional python-based image simulations codes with ingrained is straightforward. Alternatively, the initial correspondence obtained using ingrained with convolution simulation should be carefully checked with more accurate image simulation approaches. Refer to Sections 2.1.1 and 2.1.2 for further details and limitations.

For the STM mode, we developed a forward model to generate a simulated STM image from electronic charge densities data. Experimentally, the constant current STM images are obtained by moving the tip in parallel lines across the surface while the tip height is adjusted height to maintain a constant current using a feedback loop. Based on the Tersoff and Hammann approximation [42], the surface charge densities near the Fermi level correlate to the tunneling current observed in the experimental STM images. The simulated images generated by the STM forward model are the isosurfaces of charge densities near the Fermi level plotted when observed from the top view. The energy-selected charge density file from a DFT calculation (in the PARCHG file format in VASP [43, 171]) which contains the volumetric data of the partial charge densities in the entire slab structure, is the only required input structure. Beyond this, there are four modeling parameters required to simulate an STM image: electron density value (*r_val*) and an associated tolerance (*r_tol*), and the vertical distance above (*z_above*) and below the surface (*z_below*), which constrain the electron densities assumed to be available to the STM tip. The simulated STM image shows the $x - y$ grid points where the electron densities are within the specified isosurface density range (*r_val* ± *r_tol*) and the brightness at each grid point corresponds to its height. Several distinct simulated STM images can be generated from one PARCHG file with different combinations of these input parameters. *Ingrained* performs constrained optimization of these model parameters together with image processing parameters such as shear, strain, pixel size, sigma for Gaussian blur, *etc.*, to provide a final STM simulation.

In an experimental setup, the charge densities accessed by the STM tip changes with applied bias voltage. This tunneling current varies according to the density of states at the energy level which corresponds to the bias voltage. To address this, we recommend writing the partial charge densities from the DFT calculation at the energy window corresponding to the bias voltage. Since scanning transmission spectroscopy (STS) results adequately locate the band edges and a DFT calculation identifies the valence band edge (highest occupied state), the correspondence can be reasonably ascertained for negative bias. For positive bias, the fact that DFT with local or semilocal functionals tend to underestimate the band gap may cause errors with the correspondence, but hybrid functionals or a rigid shift of the bands can mitigate these errors. Multiple images can be used at different imaging voltages in order to confirm the *ingrained*-derived atomic structure – *i.e.*, the voltage-dependence of STM images provides an opportunity for a more reliable *ingrained* interpretation. The optimization for different voltages are independent and hence can be carried out in separate instances of *ingrained*.

The above STEM/STM simulations techniques all represent approximate forward models with well-known limitations. It is important that the users understand both the nuances of the experimental data they are trying to fit to (*e.g.*, only reproducible STM images obtained with multiple, independently prepared tip should be considered), as well as the simulation techniques that comprise the *ingrained* registrations, before making any conclusions about the results. Ultimately, *ingrained* is a tool to assist in finding a plausible atomistic correspondence between simulation and experiment. For electronic structure or more precise chemical information interpretation, more detailed theoretical analysis is required.

4.2.4. Image Registration

With *ingrained*, image registration is cast as an iterative optimization problem. It is assumed that scaling discrepancies between the experiment and simulation are minimal, which is reasonable when a raw microscopy image containing scaling metadata is available. The iterative optimization is local in the sense that it is restricted to solutions that are close to the initial guess, so a multi-start approach is recommended.

After an image is simulated, the first step of image registration is coarse alignment, which implies that both the simulated and experimental images are downsampled (number of actual pixels reduced) and quantized (number of unique pixel values reduced). The alignment step itself is actually an iterative procedure that takes patches of the experimental image that are highly correlated with and the same size as the simulation, and attempts to find one with zero translative offset relative to the simulation. Zero translative offset implies that no additional offset in the x or y direction is needed to improve the fit between the experimental image patch and the simulation. The translative offset is computed using an efficient phase cross-correlation function from scikit-image [140], which finds a position of maximum correlation between the two images in the frequency domain (*i.e.*, maximum correlation yields minimum translative offset). In this step, it is assumed that two atomic resolution images presented at the same scale with no translative offset between them, is a sufficient proxy for geometric consistency at the boundary of the simulation and experimental image. This is based on observation and holds for all cases presented in the results. If an experimental patch is found that satisfies the conditions of coarse alignment, a fine alignment step is applied to the simulation and experiment at native (higher) resolution. The purpose of this step is to search experimental patches in the local vicinity of the matching coordinates from coarse alignment, to find the higher resolution experimental patch with minimum translative offset. After the fine alignment step, the quality of the registration is assessed based on a custom objective function and the entire process is repeated for a new parameter set until the objective satisfies the convergence criteria set by the overall optimizer. Powell's method [172] is the default optimizer for registration, but, in theory, any derivative-free optimization method included as part of SciPy [173] could be used with very minimal revision to the current setup.

With this, the goal of optimization is to find a set of parameters for the forward model, θ, that produce a simulated image that can be arranged inside the experimental image in such a way that minimizes the objective, referred to as the the "mismatch":

$$M(\theta) = \alpha d_{trans}(\theta) + \beta d_{sim}(\theta) \tag{4.1}$$

where $d_{trans}(\theta)$ is the translation offset computed during fine alignment, $d_{sim}(\theta)$ is the similarity distance (the default is one minus the Structural Similarity Index Measure (SSIM) [85]), which quantifies the visual similarity between the simulation and experiment patch, and α, β are weights chosen to balance importance of each criteria ($\alpha = 0.1$, $\beta = 1$ are default values). This mismatch balances the importance of geometric consistency across the boundary of the simulated image, d_{trans} (*i.e.*, the atomic columns at the boundaries of the simulation are the same size, shape, and have the same orientation as the experimental columns they are replacing), and image content consistency within the boundaries of the simulated image $d_{sim}(\theta)$ (*i.e.*, the shapes and relative intensities are aligned and self-consistent across images). With the default weights and the default SSIM similarity for $d_{sim}(\theta)$, we find the following approximate interpretation of the mismatch values to hold for many of the samples observed: for M $\geq$ 1, a significant translative offset with respect to the simulation/experiment boundary exists, usually due to scaling discrepancies or local distortions; for 1 $>$ M $\geq$ 0.5, smaller translative offsets, if any, remain (typically, at least one of the bulk regions is well aligned with its respective experimental counterpart); for M $<$ 0.5, translative offset is increasingly rare, and the similarities between simulation and experiment image content is often noticeable. The solutions with mismatch values $<$ 0.2 represent the highest quality matches. In some cases, values this low are not attainable as much of this is influenced by the quality of the initial experimental input and level of structural disorder in the images sample. These observations are summarized in Figure 4.2. To provide additional confidence in the mismatch measurement's capacity to adequately assess fit-to-experiment, we conducted image similarity experiments on the *atomagined* synthetic STEM dataset [144] (see Section 3.3 for more details). Using this SSIM-based mismatch, 99% of the 2000 simulated STEM target images (postprocessed to mimic experimental noise conditions) were matched to a pristine simulation of the correct structure from a population of 4000 candidate images. The 4000 candidate choices contained a variety of images of the same crystalline structures (as the target) at different viewing angles. With the very small number of false positive matches, we are confident that the mismatch values are reliable at selecting the proper orientation of a given crystalline target if multiple are tested as part of the registration. We also address the suitability the defined mismatch for different crystalline structures of the same material in context of results presented for CdTe grain boundary systems below.

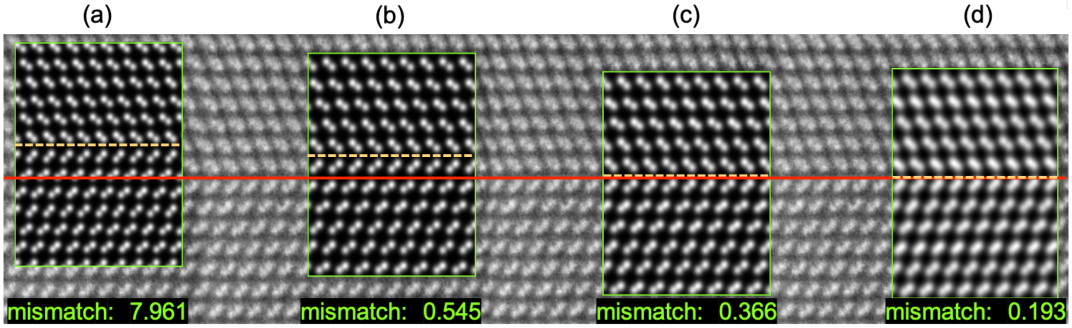


Figure 4.2. Sequence of images illustrating how mismatch values are indicative of the fit-to-experiment (with the default $\alpha = 0.1$, $\beta = 1$ weights). (a) Mismatch values > 1 implies a significant translative offset with respect to the simulation/experiment boundary (d_trans). In this case it is due to a scaling discrepancy. (b) Mismatch values between 0.5 and 1 are usually indicative of alignment between one or both grains and the respective bulk regions in the experiment, however, the positioning of the simulated interface relative to experimental interface is often unsuitable. (c) Mismatch values < 0.5 often reveal a suitable geometric fit. (d) With very low mismatch values (< 0.2), details such as the size of the atomic columns and the amount of blur begin to resemble levels found in the experimental image.

Finally, we note that if the interface can be obtained by a simple geometric combination of the grains, the resulting structure potentially matches across all portions of the image. However, this is often not the case, and further local structure operations are often needed to match the geometry of the interface more precisely. The use of atomistic modeling such as with DFT or interatomic potentials can aide the determination of relaxations at the interface. In addition, further structural sampling to refine the stoichiometry and configurations at the interface can be performed using iterative sampling methods such as Monte Carlo, basin hopping, and genetic algorithms. Moreover, a truly accurate representation of sample depth would involve further comparison with multislice simulations. Therefore, the process of obtaining these ingrained structures is considered an approximation or initial step, as opposed to exact structure determination.

4.2.5. Output

While the optimization proceeds, the default setting is for the current parameter set and mismatch score to print to screen (and is also recorded in a *progress.out* file so that specific iterations can be revisited). An example of the optimization progress information provided is included in the following snippet:

```
Iteration 1:
        • pix_size (Å)              :  0.125
        • interface width (Å)       :  0.0
        • defocus (Å)               :  1.0
        • (x, y) shear (frac)       :  (0.0, 0.0)
        • (x, y) stretch (frac)     :  (0.0, 0.0)
        • img_size (pixels)         :  (285, 171)
        >>> mismatch                :  0.6228565824012005

Iteration 2:
Warning - Solution contains significant translative offset (dtrans = 8)
        • pix_size (Å)              :  0.4
        • interface width (Å)       :  0.0
        • defocus (Å)               :  1.0
        • (x, y) shear (frac)       :  (0.0, 0.0)
        • (x, y) stretch (frac)     :  (0.0, 0.0)
        • img_size (pixels)         :  (129, 129)
        >>> mismatch                :  2.186105511593367
```

At the end of the optimization, both the structure whose image was simulated, and the parameters used to fuse the images together are accessible. These parameters are valuable because they codify the transformation needed to go from an atomic structure to a simulated image, that now has an explicit association to the experimental image. In the next sections, we highlight applications involving structure and simulation/experiment correspondence output from *ingrained*. Further tools are included as part of the main repository that allow users to create videos from selected iterations recorded in *progress.out*.

4.3. Results - Applications of *Ingrained*

The following examples showcase the capabilities of *ingrained* as both a tool for finding useful approximations of the structures in experimental imaging (for grain boundary and interfaces in particular), as well as for the development of experimentally-rationalized forward models in materials imaging. The case presented for forward model development involves STM simulation.

4.3.1. Case #1: Coherent and incoherent grain boundaries in CdTe

In this first example, we show one of the more straightforward structure initializations: a coherent {111} twin boundary in cadmium telluride (CdTe). The configuration file specifies a viewing direction along <110> and zinc-blende “F-43m” to form the bulk (to distinguish it from wurtzite and other less common metastable phases included in the MP database). In general, twin boundaries are extremely common in crystalline materials, and often form readily in response to thermal stress or applied deformation [174]. Figure 4.3a illustrates the resulting structure alongside both the final fused image and the interface comparison that results from the optimized fit-to-experiment. Because the crystals share a common plane of lattice points and mirror each other on either side of the interface, the resulting structure requires no strain to achieve coincidence along the width or depth and matches all portions of the image. For this reason, both the overall mismatch and d_{sim} values for the interfaces are particularly low (*e.g.*, < 0.20).

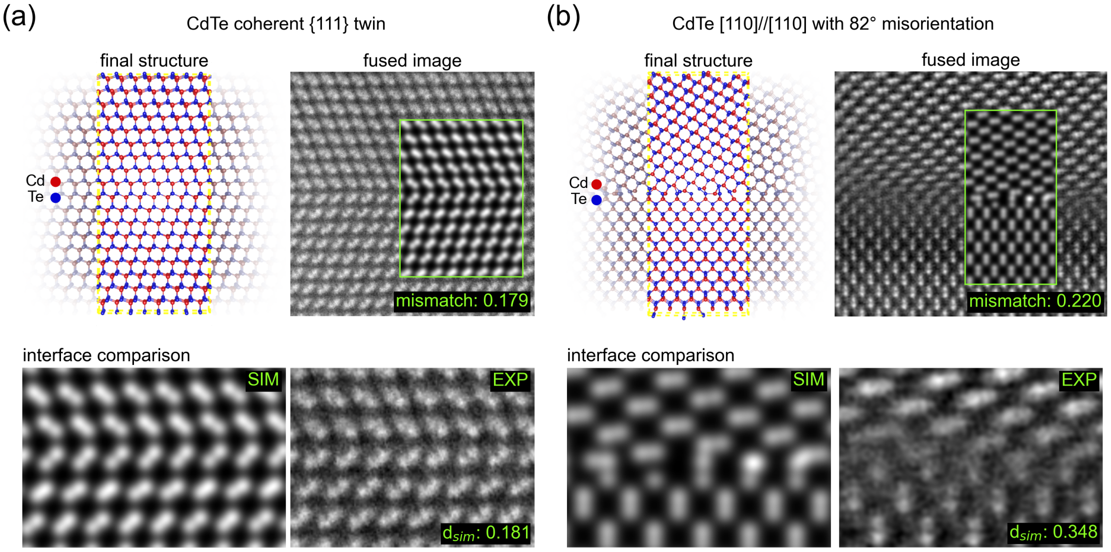


Figure 4.3. (a) The experimental image for the first collection of CdTe results contains a coherent {111} twin boundary viewed along the <110> direction. The final structure is given with the periodicity along the width highlighted. The fused image has a very low mismatch score, indicating a high-quality fusion, which is confirmed in a comparison of the simulated and experimental interfaces. (b) The experimental image for the second collection of CdTe results contains an incoherent [110]||[110] tilt boundary with 82° misorientation angle. The quality of the resulting structure – as far as matching the bulk regions – is high, and even with the natural ambiguity of the interface structure, the simulation at the interface maintains close visual resemblance. The experimental images were obtained from the authors of [4]

In the second example, we again use zinc-blende CdTe viewed along the <110> direction, but since the interface is now incoherent, the over lattice must be strained along the width to create a structure that is both computationally useful (in context of energetic calculations) but still fully periodic. Here, the magnitude of the strain is ∼1 %. Fortunately, because the viewing direction is common and the compound identical for both grains, like the twin boundary, this structure can also be constructed without strain along the depth. By observing diffraction patterns of the bulk crystalline regions, the misorientation angle between the crystals is measured at 82 degrees and is used to specify the tilt of the top in relation to the bottom. Figure 4.3b highlights a remarkable fit between simulation and experiment at the conclusion of image registration, notwithstanding the unresolved structural details of the experimental interface. Guo *et al.*[4] use this initial structure to explore the role of Se and Cl segregation in the reduction of midgap states in CdSeTe, and even after DFT relaxation is used to further optimize the interface, the initial correspondence established by ingrained is still applicable. It is also pertinent to mention that registration of wurtzite and trigonal CdTe structures were attempted in the bulk regions of the <110> zinc-blende CdTe in Figure 4.3b, and the closest mismatch to the correct CdTe bulk was 0.56 (wurtzite along <211>), which is nearly triple the mismatch scores for simulations of the correct zinc-blende structure which all scored in the range of 0.20, depending on the region of the experimental image selected for fitting.

4.3.2. Case #2: Interphase interfaces and significant localized strain

In the previous CdTe examples, both crystalline grains were identical, and except for the presence of in-plane tilt in the incoherent case, this greatly simplified construction of the periodic over lattice. The collection of results in Figure 4.4a demonstrates how *ingrained* can be used to confirm the geometric compatibility of a specific boride precipitate/Ni matrix interface in an experimentally observed structure [5], where lattice mismatch must be resolved along both the width and depth dimensions. A M_3B_2-type precipitate/Ni interface (where "M" a transition metal element *i.e.*, Cr, W, Mo, *etc.*) can be constructed from the tetragonal (P_4/mbm) Mo_3B_2 structure available in the Materials Project database. The specific structure tested in Figure 4.4a is the interface between (010) Mo_3B_2 and (1-30) Ni (viewed along <001> in each respective crystal). The ability to programmatically test geometric compatibility between different crystal phases in context of an experimentally observed structure is particularly useful for studies investigating complex interphase interfaces. Despite an excellent geometric fit (mismatch = 0.262), one can observe that the light-heavy pattern of column intensities in the experimental image, is not well-reproduced in the simulation. This suggests that the Mo_3B_2 structure, though a good candidate from the perspective of geometric compatibility, is likely not consistent with the chemistry of the structure observed. It is possible that these intensity discrepancies would be resolved with higher accuracy multislice simulations, or perhaps this suggests that the sample has mixed cations in the M-site. Once a structure is suggested by *ingrained*, sampling of different cation orderings can be performed on the structure, which do not affect the overall alignment. The final collection of results, illustrated in Figure 4.4b, is based on a HAADF STEM image of a low-angle grain boundary in a thin Si nanowire [175]. Notice a significant amount of localized strain, blurring at the interface, and the reduced spatial resolution and signal levels (compared with some of the experimental images presented in the previous examples). Image registration, as implemented, only accounts for rigid affine transformation between the simulation and experiment. Therefore, a coarse association between simulation and experiment can still be made, but only to the extent that a rigid transformation applied to simulation can capture some these distinctly localized distortions. In the case of the Si nanowire in Figure 4.4b, the overall fit is perhaps

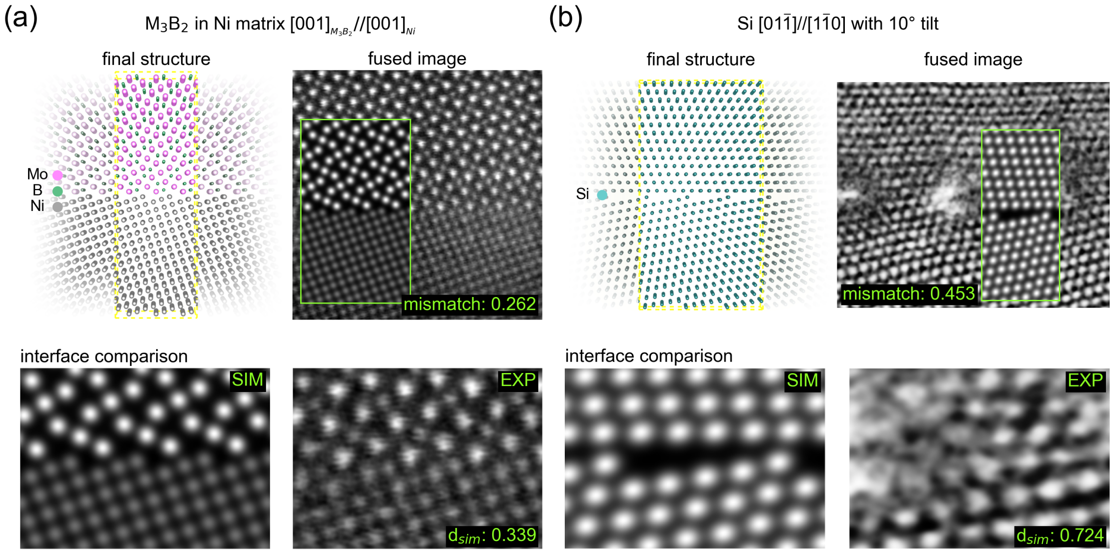


Figure 4.4. (a) The experimental image associated with the final registration contains an interface of an interphase M_3B_2 boride precipitate in a Ni-based super alloy ($[001]_{M3B2}//[001]_{Ni}$). The overall mismatch score reflects excellent geometric consistency across the boundary and within the image despite some inconsistencies in the relative intensities of the atomic columns. The experimental image was obtained from the authors of [5] (b) The final registration for a tilt grain boundary in Si [01-1]//[1-10] illustrates the difficulty in fitting a simulation with rigid affine transformations when the experimental image to be matched to contains significant local distortion and structural ambiguity at the interface.

adequate given the conditions, however, the similarity distance at the interface is unsurprisingly high ($d_{sim} = 0.724$). Though further local structure manipulations and energetic calculations are necessary to better capture the non-rigid characteristics and distortions observed particularly around the interface, this does not diminish the value of having an experimentally-informed approximate structure on which to base further analysis of the observed system. In addition, further improvements can be made to ingrained in the future to capture local distortions as measured using geometric phase analysis.

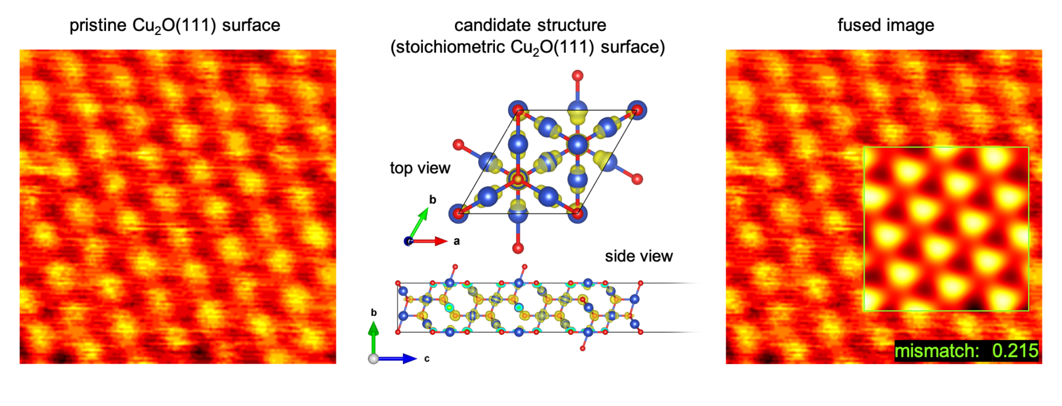

pristine $Cu_2O(111)$ surface
candidate structure
(stoichiometric $Cu_2O(111)$ surface)
top view
side view
a
b
b
c
fused image
mismatch: 0.215

Figure 4.5. High resolution STM image showing a pristine Cu_2O (111) surface is used as the experimental input to the *ingrained* framework. A DFT calculation for the proposed candidate structure, in the center, is used to create a simulated image of the surface, and *ingrained* confirms that the proposed structure is in fact consistent with experimental image, as described in [6].

The results in Figure 4.4a and 4.4b illustrate that the efficacy of ingrained is, unsurprisingly, limited by the information on composition, accuracy of the image simulation, and quality of the experimental image. Further improvement to the results may be possible by: iterative structural search to determine compositions; the use of energetic information such as density functional theory or interatomic force fields to improve the structural model, especially at interfaces; multislice simulations to improve image simulations; sampling multiple regions of the same interface, image denoising, or using machine learning approaches to improve information extraction from experimental images.

4.3.3. Case #3: STM mode and parameter optimization for Cu_2O (111)

In the previous structure initialization cases, the registration of the HAADF STEM simulation to experiment was used to verify the geometry of a plausible bicrystal structure, much of which was decided by the selection of the specific grain chemistries, orientations, and tilt outside of the *ingrained* optimization. Updates to the forward modelling parameters had minimal effects on the geometric appearance of the imaged structure. This is because the observed intensity peaks were always consistent with the presence of an atomic column in the atomistic structure, and the image formation parameters only really served to adjust the height and width of those peaks. This was the assumption when convolution imaging was basis of the forward model. For other imaging modalities, the observed intensity is not always consistent with the presence of an explicit atomic column at that site, and what is visible instead complicates the interpretation of the intensities in the image. For example, artifacts such as halos, and shade-off, commonly observed in phase contrast microscopy, complicate edge interpretation in that the appearance of an edge in the image does not necessarily mean that a true object edge exists at that location [176]. In these instances, a simulated fit-to-experiment is necessary for rationalizing the image output of the forward model.

In this example, we examined STM images of a pristine $Cu_2O(111)$ surface and proposed structure variants from Zhang *et al.* [6]. Using *ingrained* in combination with the DFT calculated partial charge densities, we were able to confirm that the atomic structure of the pristine (111) surface of a Cu_2O bulk crystal was consistent with experimental image. The partial charge densities (PARCHG) near the Fermi level of a surface slab can be obtained through DFT calculations using widely used VASP code. Figure 4.5 shows the experimental image, proposed structure, and the final image fusion with the optimized imaging parameters. The parameterized fit between simulation and experiment is similar to what was outlined in Zhang *et al.* [6].

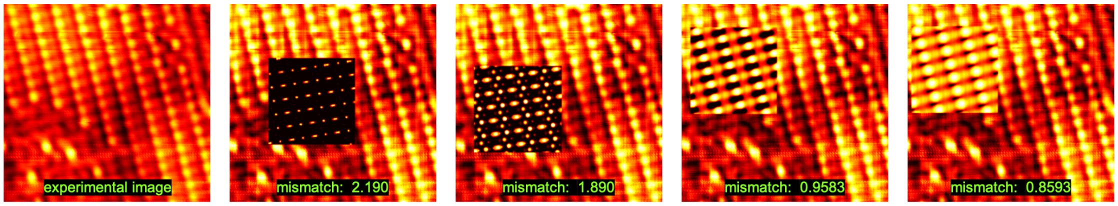


Figure 4.6. Progression of snapshots taken in the course of *ingrained* STM optimization shows improvements in both image structure (parameterization of the forward model) and in overall registration, suggesting that minimizing mismatch values is sufficient for capturing fit-to-experiment. The optimized STM simulation from *ingrained* and excellent experimental match was taken as evidence in support of a proposed borophane structure [7].

4.3.4. Case #4: STM mode drives materials discovery

The prior STM case provided validation of *ingrained* based on a known structure. When an initial structure is unknown, a multi-start approach, which implies that a variety of initial parameter configurations are tested on a population of candidate structures, can be used as means to filter out or focus in on certain structures of interest. In the case of STM, partial charge density information from several DFT calculations is the requisite "population" input, and the structures exhibiting the smallest mismatch values are interpreted as the most likely candidates. For example, Figure 4.6 depicts a progression of visual image improvements obtained through iteration of the *ingrained*, during the search for a hydrogenated borophene structure [7]. Among several candidate structures, the rectangular-2H model reported in the study, showed the lowest mismatch which helped support its identification as the structure of rectangular borophane. Again, *ingrained* provides only the means for a constructive starting point, and in this specific case, techniques such as scanning tunneling spectroscopy (STS), inelastic electron tunneling spectroscopy

(IETS), in-situ local work function measurement of synthesized sample, and further DFT calculations were all utilized in addition to the *ingrained* STM simulation, to confirm and augment any intuitions favored by the suggested atomistic correspondence. This case is included, not just as a way to illustrate the capabilities of *ingrained* in the realm of materials discovery, but also as an example of the sort of added rigor that the authors expect for users to begin making conclusions about precise chemical information from this tool.

4.4. Conclusion

Formulating materials imaging simulations in such a way as to corroborate fundamental and nuanced aspects of experimental imaging is a critical challenge that must be addressed to fully harness the power of simulation and modeling in context of materials characterization. The *ingrained* framework presented here is a tool for atomic-resolution imaging that helps establish this simulated fit-to-experiment in an automated and robust way, using a coarse-to-fine image registration approach cast as an iterative optimization problem. Through examples of STEM images of grain boundaries and interfaces, and STM images of a surface, we showcase the power of *ingrained*, not only in its ability to forge an explicit association between simulation and experiment, but also in its versatility (*i.e.*, numerous different imaging tasks can be improved with this approach). All the code for *ingrained* and the example cases explored in this work is available on GitHub. It is our hope that both computational researcher and microscopists alike will find practical use cases to add to the existing collection of examples outlined.

CHAPTER 5

Determining CdTe Grain Boundary Structures

Polycrystalline CdTe-based photovoltaics have had a major commercial impact in recent years. Despite this, an inability to adequately model highly-disordered grain boundary structures observed under atomic-resolution imaging has proved a hindrance to the understanding of device efficiency. Within the FANTASTX (Fully Automated Nanoscale To Atomistic Structures from Theory and eXperiment) framework [52], we leverage a grand canonical form of basin-hopping optimization, along with simulation and comparison of high-angle annular dark-field (HAADF) scanning transmission electron microscopy (STEM) images to experiment, to propose a collection of plausible high-angle grain boundary structures in CdTe. First principles density functional theory (DFT) calculations of energetic and electronic properties of these grain boundaries along with multi-slice STEM simulations suggest that this inexpensive materials modeling approach, coupled with automated image comparison, is of benefit in the further study of realistic grain boundary and interface structures.

5.1. Introduction

Cadmium Telluride (CdTe) solar cell technologies have had a major commercial impact in recent years [177], as its direct band gap (∼1.5 eV), high absorption coefficient ($>5\times10^5$ cm^{-1}), and general ease of processing (*e.g.*, high-deposition rates, substrate flexibility, facile doping), have enabled energy conversion efficiencies exceeding 22% in polycrystalline cells [178]. However, the interfaces in the polycrystalline cells greatly influence the mechanical, electronic, and transport properties of the materials systems, and despite improvements in spatial resolution and depth estimation for imaging crystalline materials [179], it is exceedingly difficult to determine the atomic structure of interfaces from experiments alone. First-principles methods such as density functional theory (DFT) can supplement experimental results in a way that enables improved structure determination [180–182], but an extension of such methods for simulating long-range behavior across interfaces of practical size is not straightforward. And though machine learning has played a significant role in the development of crystal structure and material property descriptors [183] to extend length scales and serve as proxy for expensive calculations [184–186], reliable descriptors and data-driven modeling paradigms are lacking for interfacial systems. Further work is needed to transform existing trial-and-error approaches to CdTe device modeling [187], in order to generate interfacial structures that can be tested, in part, by comparison to observational evidence from scanning transmission electron microscopy (STEM) images – a valuable probe of atomic structure and chemistry.

Bottlenecks preventing the facile generation and comparison of candidate structures include (1) energy calculations for interfaces and (2) STEM image simulation. For energetic modeling of practical interfaces, interatomic potentials provide a computationally efficient alternative to electronic structure methods such as DFT, while often capturing the same essential physics. Global structural optimization schemes based on energetic minimization sample a space of low energy solutions and are often sufficient if the task is to find the *most* stable structures. This is problematic if the system of interest contains possible metastable configurations, as is the case with the interfaces in grain boundary systems, because these schemes work to minimize energy, often without regard to the ‘observable’ similarity between the experimental image and a simulated image based on the proposed structure. Therefore, an additional constraint that involves STEM image simulation and subsequent comparison with experiment, ensures that the space of structures explored is consistent, to some extent, with observation. For electron microscopy simulations, enforcing true quantitative consistency typically requires simulations based on the popular multislice method of Cowley and Moodie [37], where intensity is calculated as an electron beam is propagated through slices of a sample. Unfortunately, transmission and propagation of the electron beam must be computed for each probe position, requiring computation times that are not amenable to extensive iteration. Moreover, of the modern available solutions [39, 41, 188], image convolution provides perhaps the most convenient, often qualitatively informative alternative, supporting simulation times that are commensurate with energy calculations using interatomic potentials. Once the image is simulated, recent advances in computer vision (CV) applied to atomic-resolution tomography [189], microscopy image search (Chapter 3), crystalline defect recognition [76], and nanoparticle optimization [51], that have shed light on the potential for automated image assessment in complex structure optimization problems.

With image simulation and energy calculation schemes in order, it is necessary to codify the rules of iteration such that realistic structure solutions can be obtained. Basinhopping [19] (BH), an iterative random search technique that tries to find the global minimum of an objective function, has been successfully applied to real materials systems to solve for complex minimum-energy atomic and molecular structures using DFT [183, 184], density functional tight binding [190], and empirical potentials [55]. First, a random point is chosen in the space of possible structure solutions, and from there, BH attempts to improve the structure by small perturbations (*i.e.*, “hopping”), in the positions of atoms. Over numerous iterations, BH progresses towards better structure solutions. In this study, we leverage a grand canonical form of BH, which allows for atom addition and subtraction operations at the interface, along with simulation and matching of HAADF STEM images to experiment, to propose a collection of plausible high-angle grain boundary structures in CdTe thin films. DFT calculations are performed for a select number of pareto-optimal candidate structures to demonstrate how structures obtained from cheaper interatomic potential method of evaluation, yield qualitatively consistent DFT energies. Finally, we further validate the structures obtained by performing multislice simulations and comparing the resultant simulated images with experimental images.

5.2. Methods

A (13°)(110)‖(001) CdTe bicrystal was constructed as a model grain boundary using wafer bonding as described in [191]. The HAADF STEM image of the CdTe grain boundary in Figure 5.1a shows the extent of the disorder in the interfacial region, and serves as the experimental input for BH structure optimization. In BH optimization, atomic configurations of high-angle CdTe grain boundary structures are iteratively modified and evaluated in order to find a collection of solutions with both low energy and high simulated-to-experiment image similarity. The BH optimization is part of a package under development at the Center for Nanoscale Materials at Argonne National Laboratory, titled FANTASTX (Fully Automated Nanoscale to Atomistic Structure from Theory and eXperiments) [52], which performs structure search based on a combination of energetic information from DFT or empirical potentials, and similarity measurements between simulations and experimental inputs from microscopy, scattering, and spectroscopy modalities.

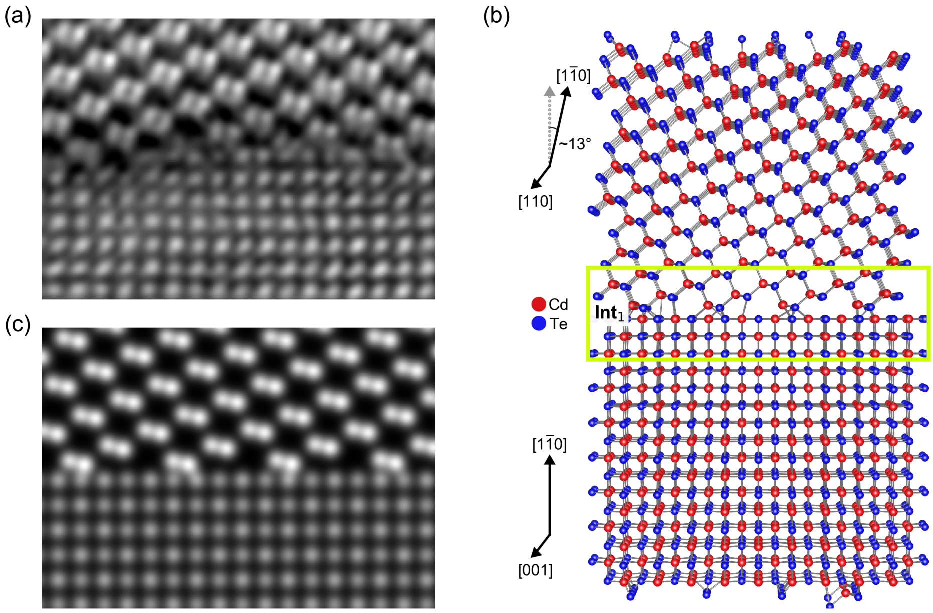


Figure 5.1. (a) Experimental HAADF STEM image of a (13°)(110)‖(001) CdTe grain boundary system. (b) A quasi-periodic CdTe grain boundary structure. The bulk portions are consistent with orientation estimates from the experimental image. The highlighted interface (Int1) indicates the specific volume region over which the optimization occurs. (c) Simulated convolution HAADF STEM image of the initial grain boundary structure with an optimized fit to the experimental image (see Chapter 4)

5.2.1. Initialization

The initialization procedure incorporates orientation information for the bulk crystalline regions of the experimental image, to find quasi-periodic boundaries around a bicrystal structure (Figure 5.1b). Strict periodicity requirements are relaxed to provide reasonable computational efficiency in the downstream modeling and simulation steps. The volume region containing 8% of the total length of the bicrystal above and below the central plane between the two bicrystals is considered the interfacial region. For the bicrystal structure in Figure 5.1b, this is 12Å in width (6Å above and below the interface). The final step of initialization is to solve for the spatial correspondence (image registration) between the new bicrystal and the experimental image, which ensures that the bulk portions of the structure are aligned to the experimental image, leaving the interface region as the area for refinement. With this, the structure in Figure 5.1b becomes the initial candidate solution, and its simulated image in Figure 5.1c, is used as the starting point for image comparisons. Initial structure creation and the correspondence-to-experiment tasks are accomplished using the the *ingrained* toolkit, as outlined in Chapter 4.

5.2.2. Update Steps

After the initial structure is generated, new potential candidates are obtained in two different ways. One is through a jump operation, which is a random displacement of individual or groups of atoms considered part of the interface region defined in Figure 5.1b. With a jump, a given atom is displaced according to a uniform distribution anywhere in a cube, centered on the current atomic coordinate that extends in all three directions for a distance (edge length) plus and minus the jump size, in this case, between -3Å and +3Å, which is an approximate average bond length when considering all pairs of elements. Jumps that displace an atom outside the interface boundary (into the bulk) are forbidden. Because the simulations are grand canonical, the other way a new candidate structure is obtained is through an atom addition or subtraction operation. In these scenarios, an atom of a random type is inserted or removed from the interface region, while the rest of the structure remains fixed. A 0.8 Å minimum distance constraint and periodic boundary conditions are enforced for both the jump and add/subtract operations. The completion of an update step constitutes what is referred to as a BH step.

5.2.3. Evaluation of Candidate Structures

After a BH step, evaluation of the candidate structure involves two calculations: (1) the total grand canonical energy of the relaxed structure, and (2) the similarity distance between a simulated HAADF STEM image of the candidate structure, and the original experimental image. For the energy calculation, the candidate structures is relaxed using the Stillinger-Weber (SW) potential for CdTe evaluated with LAMMPS [192], and the total energy, considering the grand canonical approach, is computed as:

$$E_{total} = E_{relaxed} - \mu_{Cd}N_{Cd} - \mu_{Te}N_{Te} \tag{5.1}$$

where $E_{relaxed}$ is the total energy of the relaxed interfacial structure, μ_{Cd} and μ_{Te} are Cd and Te chemical potentials, set to mimic a Cd-rich synthesis environment, and N_{Cd} and N_{Te} are the numbers of Cd and Te atoms in the model, respectively. At the Cd-rich limit, the grain boundary is assumed to be in equilibrium with metallic Cd and bulk CdTe. Therefore, $\mu_{Cd} = \mu_{Cd}^{0(bulk)}$ and $\mu_{Te} = \mu_{CdTe}$ - μ_{Cd}^{0}, where μ_{CdTe} is the total energy of the bulk CdTe evaluated with the SW potential, and $\mu_{Cd}^{0(bulk)}$ is the binding energy of elemental Cd metal. In this convention, E_{total} is zero for stoichiometric bulk CdTe.

A HAADF STEM image is simulated using a convolution model (see Section 5.2.4), and the similarity distance, d_{SSIM}, between the simulated image, I_{sim}, and the original experimental image, I_{exp}, is calculated using Structural Similarity Index (SSIM) [85] as:

$$d_{SSIM} = 1 - SSIM(I_{exp}, I_{sim}) \tag{5.2}$$

SSIM values range between 0 and 1, with 1 indicating a perfect match. The choice of SSIM for an image similarity measurement is based on the findings from Chapter 3, which show that SSIM provides an appropriate balance of efficiency and performance for comparisons between HAADF STEM simulations and experimental images.

With this setup, the overall image similarity objective is a minimization of the difference between simulated and experimental images, which we call similarity distance. Ultimately, to represent the complete objective of the problem (minimize energy and similarity distance), we construct the main objective function as a weighted summation of the energy and image similarity contributions.

$$Z_{total} = \alpha E_{total} + \beta d_{SSIM} \tag{5.3}$$

where Z_{total} is the objective score and α and β are the weights associated with E_{total} and d_{SSIM} from Equation 5.1 and Equation 5.2, respectively (a restatement of Eqn. 2.11)

All calculations use weights of $\alpha = 1$, $\beta = 200$, which given the approximate range of E_{total} values from 150-200 eV, and the approximate range of d_{SSIM} values from 0.25-0.65, places a ~20-45% weight on the image similarity measurement in the overall objective decision. Prior attempts at an equal weighting of the variables heavily favored d_{SSIM} solutions (*i.e.*, structures matching an image well in projection), but with disregard for coordination chemistry. After each update step, the change in objective due to structure updates is computed, and a candidate structure is accepted or rejected as a global solution based on the metropolis criterion with $T = 0.1$. If the structure is not accepted, the existing global configuration serves as the starting point for the next step taken. This continues until 50,000 BH steps elapse without an improvement to the global solution.

Table 5.1. Input parameters for HAADF STEM simulation codes.

parameter	*incostem* (convolution)	*autostem* (multislice)
V0 (kV)	200	200
Cs3 (mm)	0.002	0.001844
Cs5 (mm)	0	-0.7503
df (Å)	1.5	-28.17
apert (mrad)	29	29
spec trans func N_x, N_y (pix)	N/A	1024, 1024
probe wave func N_x, N_y (pix)	N/A	512, 512
min,max angles (mrad)	90, 175	90, 175
slice thickness (Å)	13.66	1.706
total sample depth (Å)	13.66	204.9
temperature in degrees (K)	N/A	300
# of configurations	N/A	25
source size (FWHM in Å)	1.50	0.78

5.2.4. Image Simulations

Kirkland's incostem and autostem codes are used for convolution and multislice image simulations, respectively [35]. The parameters for each simulation type are provided in Table 5.1. Parameter settings for the multislice image are informed primarily from the summary of image acquisition parameters provided by the STEM software. The convolution parameter settings differ from multislice largely because of the need to compensate for realistic depth effects and general blurring that occurs as a result of averaging over several thermal configurations in the multislice calculations.

5.2.5. Density Functional Theory (DFT) Calculations

DFT calculations are performed using the Vienna Ab Initio Simulation Package (VASP) [171] with supplied Projector augmented wave (PAW) potentials [43]. The generalized gradient approximation (GGA) exchange correlation functionals parameterized by Perdew-Burke-Ernzerhof (PBE) [193, 194] was used. A kinetic energy cutoff of 343 eV was used. The Brilluoin zone was sampled with the Γ point for dislocation core models. All atomic positions are relaxed to give an energy convergence of 10^{-5} eV/atom.

5.3. Results and Discussion

The optimization progress is shown in Figure 5.2a. In addition to the solution obtained at algorithm termination (211,153 BH steps), we are also interested in solutions that reside on the pareto front, the leading edge of the solution space (*i.e.*, the space containing solutions as function of their objectives). These pareto-optimal solutions cannot be improved in a single objective without penalty to another. From the existing pareto-optimal solutions, the final collection of *critical* structures selected included the solution with (1) minimum energy, (2) minimum similarity distance, and (3) minimum main objective score Z_{total} (*i.e.*, the best "*energy* solution", the best match-to-experiment "*matching* solution", and the best "*objective* solution", respectively). Figure 5.2b shows the critical structures alongside other BH solutions plotted in a favorable region of the solution space. The energies (denoted with γ_{GB}) are given as interfacial energy per unit area (mJ/m^2) to compare with values reported in Sen *et al.* [187] for high-angle CdTe grain boundaries. The values in this study are substantially higher (~3-4x), however that is to be expected given the magnitude of the disorder at the interface.

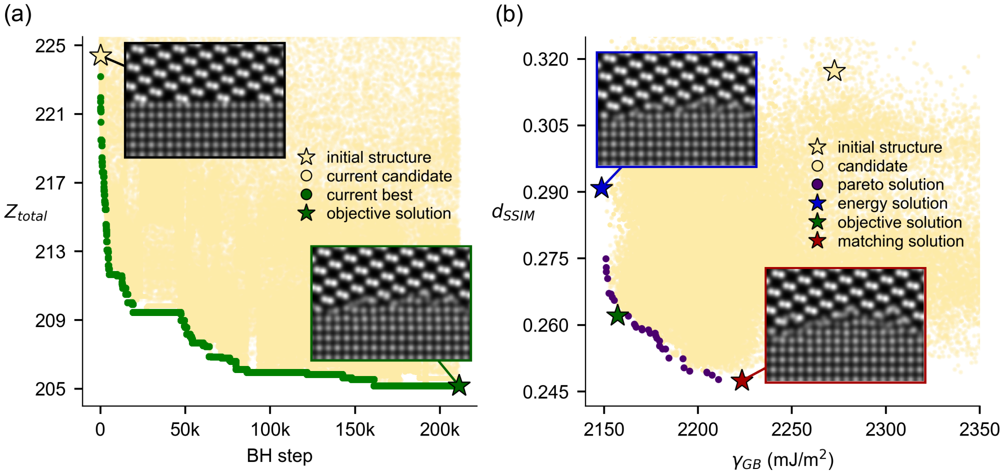


Figure 5.2. (a) Plot illustrating the progress of the objective value as optimization proceeds. The green points track the solution with the current best objective value. Convolution HAADF STEM images show both the initial structure and solution with the best objective value (*objective* solution) after termination. (b) Plot highlighting the *critical* solutions and their associated simulated images in solution space. The purple points represent pareto-optimal solutions. In addition to the *objective* solution (green star), both the blue and red stars represent pareto-optimal solutions with the best energy (*energy* solution) and best match-to-experiment (*matching* solution).

In order to further examine the energetic and electronic properties of the *critical* structures, we perform DFT calculations. These calculations are expensive, as each structure contains nearly 1200 atoms. The DFT-relaxed *energy* solution is presented in Figure 5.3a, with two bulk and two interface regions highlighted. The simulated STEM images for the starting and optimized structures are shown in Figure 5.3b, revealing a rearrangement of atoms in the optimized interface "Int1" region. This rearrangement occurred for the interfaces of the remaining *critical* structures. Further, we computed the electronic density of states (DOS) for each structure and projected them onto the bulk and interface region atoms, which is shown for the *energy* solution in Figure 5.3c. Each calculation leads to

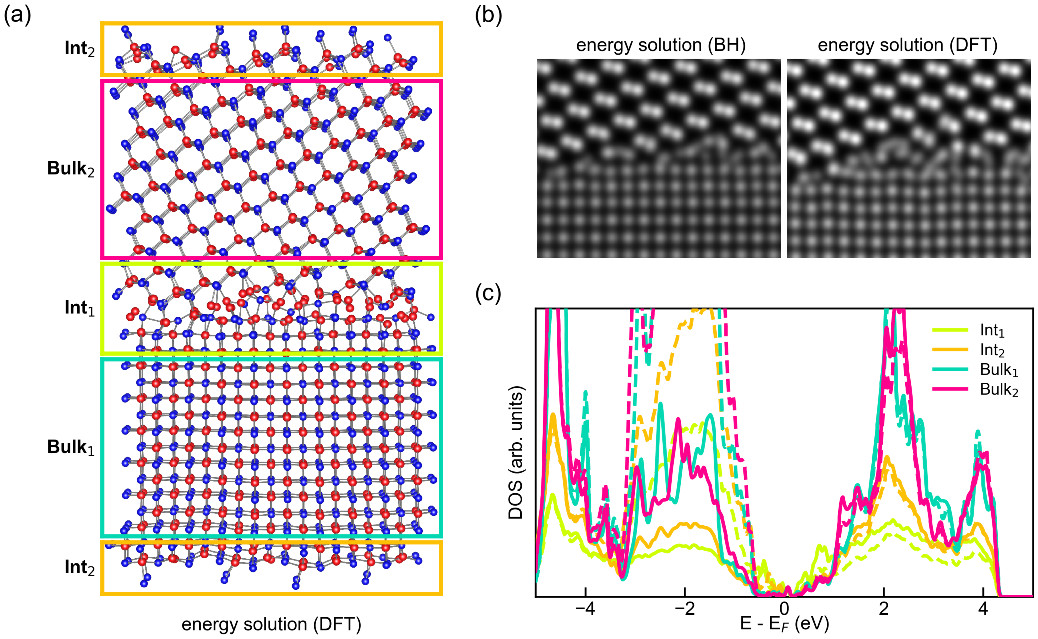


Figure 5.3. (a) The DFT-relaxed structure with bulk and interface divisions of the interface highlighted. (b) The simulated STEM images (convolution) before and after DFT relaxation. (c) The computed electronic density of states for the energy solution.

a qualitatively similar DOS, indicating a similarity in electronic structure for each of the *critical* solutions highlighted in Figure 5.2b. There are some mid-gap states that originate mostly from the interface atoms, as has been studied and reported in the past [4, 195]. Upon comparing the energetics of the three solutions following DFT relaxation, we observed that the *energy* solution had the lowest total DFT energy, while the objective and *matching* solutions were 0.3 meV/atom ($\Delta\gamma_{GB}$ = 5.0 mJ/m^2) and 1.1 meV/atom ($\Delta\gamma_{GB}$ = 18.3 mJ/m^2) higher in energy, respectively. Despite the interface atom rearrangement due to DFT relaxation and the known limitations of the SW potential [196], the energy ordering of the three structures remained the same, meaning the significantly cheaper interatomic potential method of energy evaluation were qualitatively correct.

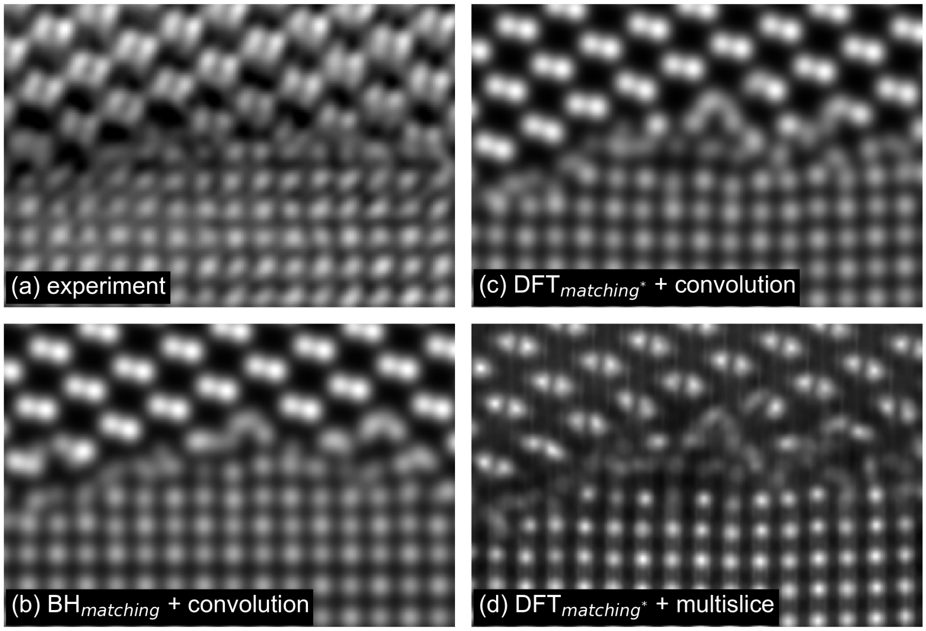

(a) experiment
(c) DFT$_{matching^*}$ + convolution
(b) BH$_{matching}$ + convolution
(d) DFT$_{matching^*}$ + multislice

Figure 5.4. A panel of STEM images highlighting the combinations of energy-relaxation and image simulation techniques applied to the final matching solutions. (a) The initial experiment image. (b) A convolution simulation of the original matching solution from BH, referred to as the "BH + convolution" image. (c) A convolution simulation of the matching solution after DFT relaxation, referred to as the "DFT + convolution" image. (d) A multislice simulation of the matching solution after DFT, referred to as the "DFT + multislice" image. The "matching*" designation is used to separate the initial matching structure from the DFT-relaxed matching structure, which is often differs due to a rearrangement of interface atoms.

The multislice method is widely accepted as a quantitatively accurate approach for simulating contrast in HAADF STEM images. The convolution images used to compute the similarity distance in the main objective (Z_{total}) capture the approximate appearance of the STEM images; however, they do not replicate the depth-specific contrast modulations that occur, particularly near the interface. To this end, we applied Kirkland's multislice code [35] to the *critical* solutions after DFT relaxation in the hope that a more quantitative simulation was capable of replicating certain intensity features of the experimental image. The simulation parameters were set in accordance with the known experimental microscope/acquisition settings (see 5.2.4). Figure 5.4 highlights combinations of energy-relaxation and image simulation techniques applied to the final *matching* solution, and Table 5.2 shows an increase in the similarity distance (*i.e.*, more mismatch) moving from the convolution image of all of the *critical* solutions “BH + convolution”, to the final multislice image of the DFT-relaxed solutions “DFT + multislice”. This general trend of increasing similarity distance from “BH + convolution” to “DFT + multislice” solutions within the structures tested is expected; DFT leads to a rearrangement of interface atoms, and the multislice calculation incorporates realistic sample depth and microscope settings - neither of these additional steps were explicitly considered in the optimization loop. Table 5.2 also confirms that across the combinations of energy-relaxation and image simulation techniques applied, the ordering of the similarity distances between the three structures (*i.e.*, $d_{SSIM}^{matching} < d_{SSIM}^{objective} < d_{SSIM}^{energy}$) remained the same. In general, the preservation of the ordering for energy, similarity distance, and intensity correlation across the interface substantiates the value of cheaper energetic and image simulation proxies in the context of iterative structure optimization; the similarity of the electronic DOS across the pareto solutions tested further suggests that perhaps the BH solutions chosen from the pareto front are representative of the true grain boundary structure corresponding to the input experimental STEM image.

Table 5.2. Similarity distances for selected energy-relaxation and image simulation techniques applied to critical structures.

	BH + convolution	DFT + convolution	DFT + multislice
$d_{SSIM}^{matching}$	0.236	0.281	0.348
$d_{SSIM}^{objective}$	0.257	0.301	0.355
d_{SSIM}^{energy}	0.267	0.358	0.402

5.4. Conclusion

Iterative optimization, with physically motivated energetic and image similarity objectives as proxies for expensive higher-fidelity calculations, is used to propose a collection of plausible high-angle grain boundary structures. The ordering of the proposed structures, both energetically and in terms of the similiarty distance measured from the associated image simulation to the experiment, is preserved after both DFT relaxation, which leads to further atom rearrangement at the interface, and multislice simulation, which provides a better model of image contrast. This implies that the significantly cheaper interatomic potential method of energy evaluation with image convolution as the basis for comparison is qualitatively correct for this system. Additionally, the solutions examined lead to a qualitatively similar DOS, and reveal the presence of some previously observed mid-gap states that originate mostly from the interface atoms. Perhaps this indicates that not only is there a similarity in electronic structure among structures on the optimization's pareto front, but also that the structures discovered are in fact representative of the true grain boundary structure corresponding to the input experimental STEM image.

CHAPTER 6

Constructing Self-Labeled Microscopy Datasets from Literature

*Due to recent improvements in image resolution and acquisition speed, materials microscopy is experiencing an explosion of published imaging data. The standard publication format, while sufficient for data ingestion scenarios where a selection of images can be critically examined and curated manually, is not conducive to large-scale data aggregation or analysis, hindering data sharing and reuse. Most images in publications are part of a larger figure, with their explicit context buried in the main body or caption text; so even if aggregated, collections of images with weak or no digitized contextual labels have limited value. To solve the problem of curating labeled microscopy data from literature, this work introduces the EXSCLAIM! Python toolkit for the automatic **EX**traction, **S**eparation, and **C**aption-based natural **L**anguage **A**nnotation of **IM**ages from scientific literature. We highlight the methodology behind the construction of EXSCLAIM! and demonstrate its ability to extract and label open-source scientific images at high volume.*

6.1. Introduction

Journal articles have long been the standard for communicating important advances in scientific understanding. As sophisticated measurement and visualization tools render scientific communication more intricate and diverse, the visual presentation of scientific results as figures in these articles is noticeably more complex [197]. With this complexity, which is often a byproduct of the "compound" layout of the figures (*i.e.*, figures containing multiple embedded images, graphs, and illustrations, *etc.*), the meaning of each standalone image is not always apparent. The result is that individual images are not only unsearchable, but the effort required to extract them into a machine-readable format is significant. This plays a major factor in the relative scarcity of general materials characterization images in the development and testing of new algorithms in deep learning (DL), an emerging field characterized by the use of deep neural networks – hierarchical, multi-layered structures of processing elements – to learn representations of input data (often images) that reveal important characteristics of its content or overall appearance. The current surge of interest in DL stems from recent success in applications such as facial recognition [198], self-driving cars [199], and complex game playing [200]. Much of this success is the byproduct of having large labeled datasets available for training [201], and in order for current materials imaging classification and recognition tasks [202–209] to reap the benefits afforded by DL pipelines, having access to substantial labeled data is crucial. Fortunately, the incentive to publish is nearly ubiquitous across all scientific disciplines, and with a mechanism for automatic dataset construction that includes both separating out individual images from compound figures, as well as providing concise annotations describing key aspects or classification of the image content, much more of the scientific imaging data in literature could be utilized for training and developing DL models.

The effort to automate the construction and labeling of datasets from general web data has garnered broad attention from the computer vision, language technologies, and even chemistry/materials informatics communities [210–213]. In chemistry and materials informatics, most of the focus has been on the development of text-mining tools adapted for "chemistry-aware" natural language processing (NLP), and have been used to create datasets of material properties and synthesis parameters from journal article text [24, 25, 115, 127, 214]. For imaging datasets, standard computer vision approaches take the top images retrieved from a keyword query of an image search engine (*e.g.*, Google, Flikr, *etc.*) and train classifiers to further populate datasets based on a keyword [213]. Unfortunately, this approach is problematic for scientific figures because of their compound layout. Current works that address this layout problem (what we refer to as problem of "figure separation") rely on hand-crafted rule based approaches [215–218], or adapt neural-networks to interpret figure separation as an object detection problem [219, 220]. While hand-crafted techniques generally work well for sharp image boundaries, and neural networks capture irregular image arrangements, neither of these methods are designed such that explicit references to the caption text are considered as part of the separation. This is problematic for both figure separation and labeling because when the two are not consistent with each other (*i.e.*, there are more/fewer images than subfigure labels) the intended description for an image is not always clear. Despite efforts to advance figure separation and summarization related to automatic dataset creation [215, 216, 221–224], tools that do this in query-to-dataset fashion, capable of extracting individual images from figures, classifying image content, recognizing scaling information, and assigning a relevant descriptive label based on specific user search requirements, are currently lacking.

In this work, we present a tool for the automatic **EX**traction, **S**eparation, and **C**aption-based natural **L**anguage **A**nnotation of **I**mages (EXSCLAIM!) for scientific figures and demonstrate its effectiveness in constructing a self-labeled electron microscopy dataset of nanostructure images. The EXSCLAIM! tool is developed around materials microscopy images, but the approach is applicable to other scientific domains that produce high-volumes of publications with image-based data (*e.g.*, graphs, illustrations, *etc.*) as the basis for assigning keyword labels is dictated by user search terms.

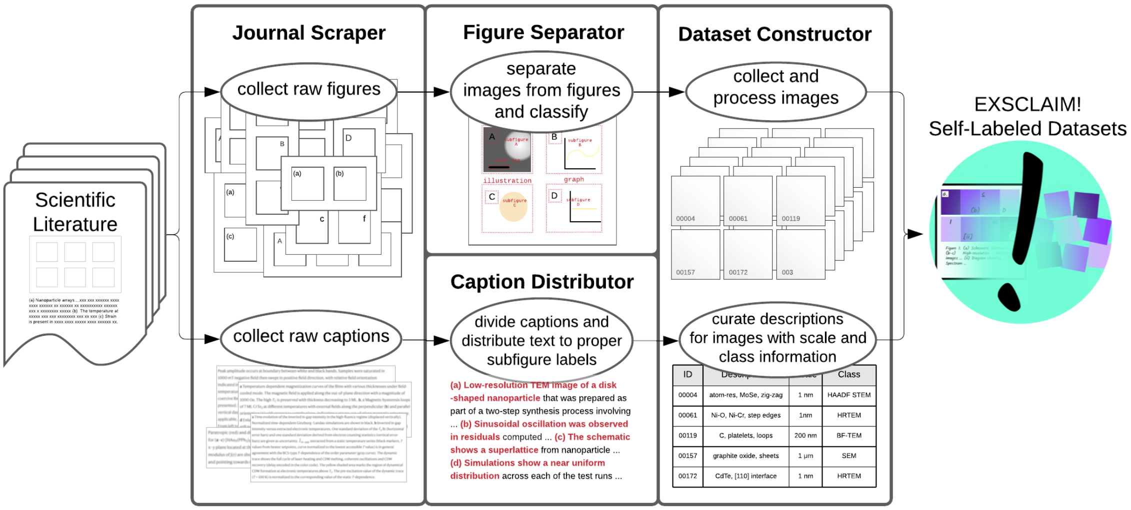

Scientific Literature
Journal Scraper
collect raw figures
collect raw captions
Figure Separator
separate images from figures and classify
illustration
graph
Caption Distributor
divide captions and distribute text to proper subfigure labels
(a) Low-resolution TEM image of a disk -shaped nanoparticle that was prepared as part of a two-step synthesis process involving ... (b) Sinusoidal oscillation was observed in residuals computed ... (c) The schematic shows a superlattice from nanoparticle ... (d) Simulations show a near uniform distribution across each of the test runs ...
Dataset Constructor
collect and process images
curate descriptions for images with scale and class information
ID
Class
00004 atom-res, MoSe, zig-zag 1 nm HAADF STEM
00061 Ni-O, Ni-Cr, step edges 1nm HRTEM
00119 C, platelets, loops 200 nm BF-TEM
00157 graphite oxide, sheets 1 μm SEM
00172 CdTe, [110] interface 1 nm HRTEM
EXSCLAIM! Self-Labeled Datasets

Figure 6.1. An overview of the EXSCLAIM! pipeline. The *journal scraper* extracts raw figure/caption pairs from journal articles, the *caption distributor* divides caption text into segments that are consistent with the images in the figure, reducing them to single keywords if possible (*i.e.*, self-labeling), and the *figure separator* computes bounding boxes that separate and classify the individual images from the full figure. The *dataset constructor* processes all prior information to create a structured self-labeled dataset.

6.2. Methods

6.2.1. Design Overview

The main goal of the EXSCLAIM! toolkit is to provide researchers with a collection of modules that (1) enable comprehensive keyword searches for scientific figures within open-source journal articles, and (2) facilitate the extraction and pairing of images from within figures, to the appropriate descriptive keywords and phrases from caption text. Figure 6.1 provides an overview of the pipeline, highlighting the role that the *journal scraper*, *caption distributor*, *figure separator*, and *dataset constructor* modules serve in transforming search queries made to scientific journal platforms, into structured self-labeled image datasets.

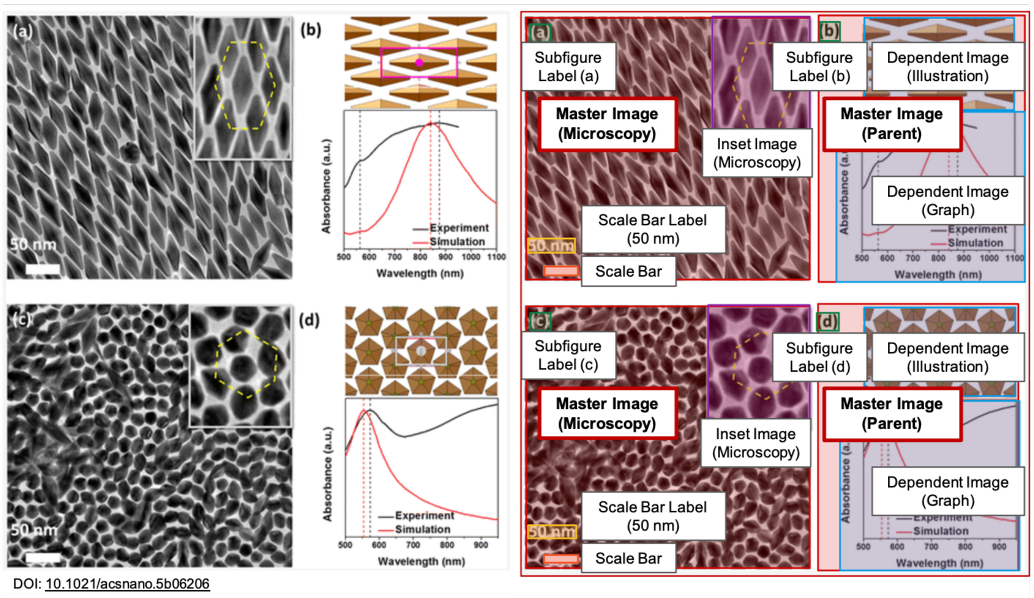
The search queries define the scope and extent of the search, and are JSON objects with attributes populated by the user. Attributes such as keywords, synonyms, journal family, *etc.* are required, along with total number of journal articles to extract content from. The above modules incrementally populate a separate output JSON object, which provides a final record of the relevant figure URLs, bounding boxes that encode the size and location of the individual images on the figure, and proper sequences of caption text to serve as labels for the images. The *dataset constructor* then uses this output file to construct a high-quality self-labeled image dataset.

6.2.2. Journal Scraper

This module performs the first extraction step in the pipeline. It is responsible for retrieving figure/caption pairs from articles that fit the parameters of the search query, and uses the Python Requests library [225] to handle all of the HTTP requests. First, article URLs are extracted from a collection of individual searches formed from all possible combinations of search field keywords and associated synonyms. With an ordered list of open-source article URLs, the scraper contains a method that sends further GET requests to retrieve each article and its figure/caption pairs. Finally, the output JSON is populated with all the general article information returned during the extraction steps, including the full caption data for each figure, and URLs to the articles and their associated figures – providing basic provenance to the data workflow. Full article text can also be included if desired. Currently, the *journal scraper* has the functionality to parse open-source journals from the Nature, American Chemical Society (ACS) and Royal Society of Chemistry (RSC) families by default, but is designed in such a way that users can create new parsers with straightforward modifications of the base “journal family” class. The reason these three were chosen as defaults is related to both their impact and the fact that their contents are served in fundamentally different ways. Nature and ACS have static content, while RSC’s content is served dynamically. Consequently, with this collection of default parsers, we cover two common ways information could be obtained from a website, which leaves the door open for this tool to be easily extended to any open source journal. All the figure and caption extraction is performed directly from the HTML version of the article and does not require PDF downloads. This is a design choice based on the observation that most open-source content is available directly from the HTML text and the ability to download the information in PDF form is more of a convenience rather than a necessity.

6.2.3. Caption Distributor

With all figure and caption information recorded, the next step is to distribute subsequences of the caption text to the respective subfigure that they reference. This first involves sentence tokenization with Parts-of-Speech (POS) tagging. POS tagging deconstructs sentence text into small units (tokens), which are given a tag that identifies their part-of-speech in the sentence. For this, the natural language processing (NLP) tokenization tools from the spaCy library [119] are extended to properly assign “subfigure identifier” tags, to patterns that indicate the presence of a subfigure description. This is achieved through custom rule-based matching and includes all alphanumeric symbols as well as directional phrases such as “top right” or “bottom left”, *etc.* With this custom tagging, the “(a)” in a phrase such as “(a) Nanoparticles deposited on . . . ”, is properly interpreted as the subfigure identifier instead of a determiner surrounded by parenthesis.



DOI: 10.1021/acsnano.5b06206

Figure 6.2. The left panel is the initial figure before annotation, the right is a copy that has been properly annotated according to the Master-Dependent-Inset (MDI) model. The presence of a subfigure label is necessary for an image to be given a "master image" designation. Because the master image corresponding to subfigure label "b" governs two distinct images (illustration and graph), it is classified as a parent. Additional image features such as the scale bar and scale bar label are also identified.

From there, a regular expression style of pattern matching is performed on the list of the custom POS tags, with a dictionary of reference sequences collectively representing a "standard syntax" for typical subfigure image descriptions in caption text. Refer to Chapter 2 for a specific example of POS-tagging performed on caption text (Table 2.1), and for the "standard syntax" pattern matching used to extract proper image descriptions (Table 2.2). Finally, words from the initial search query that appear in the segment of caption text for an image are designated as "keywords" for the image.

6.2.4. Figure Separator

A scientific figure is often composed of multiple individual images that must be separated before further processing and pairing with the caption text. This figure separation step involves subdivision of the extracted figures into “master images”, according to what we establish as the Master-Dependent-Inset (MDI) modeling paradigm. In the MDI paradigm, the master image is defined relative to a subfigure label (*e.g.*, “(a)”, “(b)”,*etc.*), and the subfigure label is the functional element bridging the visual image content to the text describing it. All visual components (*i.e.*, all images, drawings, clarifying annotations) belonging to the complete entity referenced by the subfigure label, are collectively referred to as the “master image”. In addition to defining the master images in context of the full figure, the *figure separator* both classifies the image according to the nature of the image content (*e.g.*, microscopy, diffraction, graph *etc.*), and extracts scaling information that is present in the form of a scale bar on images within the figure.

Figure 6.2 provides a detailed view of the MDI model applied to a standard figure. Both insets and scaling information are shown in subfigure “(a)”, and subfigure “(b)” is useful for demonstrating the need for a “master image” classification, as it is clear that it is referencing more than one distinct image. Moreover, when a master image contains multiple dependent images, it is classified as a parent. The detection and classification of master images is the primary task of the *figure separator*, which follows a two-stage framework outlined in Jiang *et al.* [226]. The first stage identifies subfigure labels within the compound figure. This is achieved using a combination of object localization (YOLOv3-style object detector [94]) and object recognition (ResNet-152 [227]) neural networks. In the second stage, a binary mask is created to provide visual anchors for the subfigure locations in the full figure. The binary mask is then concatenated with the standard RGB input channels and fed into the master image detection module. Taken together, these neural networks locate and classify master images within the figure, while preserving the association between master images and their respective subfigure labels.

After the master images are detected, localized, and classified, the final step is to extract scaling information, which also relies on a two-stage neural network. First an object detection neural network (Faster R-CNN [228]) is used to detect the bounding boxes of scale bar labels and scale bar lines that exist in a given figure. Next, the detected scale bar labels are fed into a Convolutional Recurrent Neural Network (CRNN [229]) in order to perform text recognition, making the scale bar label text machine readable. The result of the CRNN is processed by a rule-based search to ensure the output is a valid scale bar label (*i.e.*, a number followed by a unit). Multiple scale bar lines and scale bar labels in a single figure are matched by assigning the scale bar labels to scale bar lines greedily based on the distance between the center of their respective bounding boxes. Each matched scale bar-scale bar label pair is assigned to the subfigure in which it is contained. Using the length in pixels of the scale bar line and the subfigure and the scale bar label text, the real space size of the subfigure can be determined.

6.2.5. Dataset Constructor

The above modules incrementally transform the search query into a JSON output structure, which contains all the information necessary to create a dataset annotated from caption descriptions in the literature. The *dataset constructor* allows users to process

URLs, bounding boxes, and annotation information in the saved JSON output, to gather the appropriate figures, extract their individual images, and align the keyword labels. The separated images can then be stored locally along with a *.csv* that contains row entry descriptions and keywords for each image. This mode of operation lends itself to users looking for quick access to custom data curation. For more ambitious users, additional features exist for easy connection to a local MongoDB server. Postprocessing tools aimed at further interpreting and removing the pixel annotations (*e.g.*, subfigure labels, scale bars, and any additional text, symbols, or shapes added to further clarify the image content) are under active development.

6.2.6. Crowdsourcing Figure Annotation with Mechanical Turk

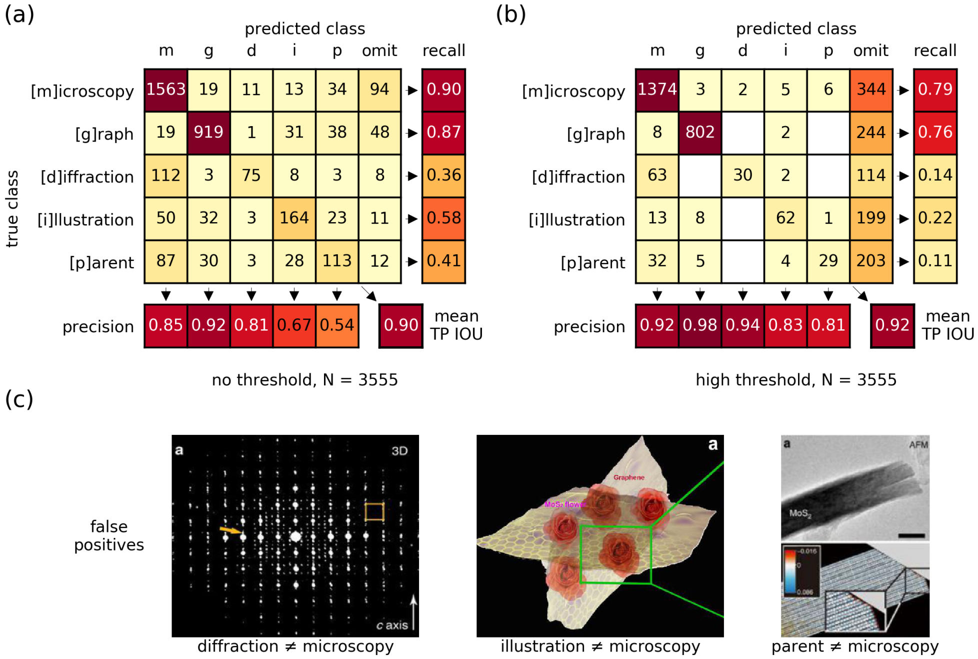
In order to achieve sufficient accuracy, the above models must be trained on images that are deemed as proper references, or have been verified as representing the correct way to locate and classify master images (*i.e.*, ground truth). Because the demands of this task are unique in that figure separation does not fall explicitly within the canon of standard computer vision training tasks, we needed an approach to scale the annotation effort to ensure the best accuracy on the figure separation task. For this, we used the crowdsourcing platform from Amazon called Mechanical Turk (MTurk). Though proper interpretation of a scientific image often requires an expert to understand the nuances of the image content, identifying where the master images are located, as well as their proper classification, can be formulated so that those without a rigorous science background can annotate the images with only a very modest amount of instruction. As such, we designed a custom figure annotation GUI (snapshots of the GUI are included in the supplementary information) within the MTurk platform, and asked workers to draw bounding boxes around each master image in the dataset, and then classify them. This allowed us to quickly create a dataset of > 3000 MDI annotated figures ($\sim$ 18,000 separate images). The basis for training the current version of the *figure separator* involves augmentation of a random sampling of 2000 of the annotated images from MTurk and is described in more detail in Jiang *et al.* [226].

6.3. Results and Discussion

There are several components of the EXSCLAIM! pipeline that must be considered in evaluating overall performance. Here, we (1) validate classification accuracy for the *figure separator* using precision and recall metrics obtained on a reference dataset, (2) examine the various scenarios for how caption text is assigned, quantifying accuracy for the case where a single keyword is used to describe the image, and finally (3) provide suggestions for how to create new labels or general topics to associate with images that are left un- or under-annotated. In total, the results emphasize the attention placed on accuracy and extensibility of the EXSCLAIM! toolkit.

6.3.1. Validation of the Figure Separator on MTurk Dataset

To validate the classification and bounding box prediction accuracy of the *figure separator*, 784 figures (3555 separated images) from the crowdsourced MTurk dataset – withheld during training – were used as part of the validation set. The validation dataset is available from the Materials Data Facility [230, 231] (DOI: 10.18126/a6jr-yfoq [232]). The results are shown in Figure 6.3. Images were scraped from Nature publishing sources (the



(a)

predicted class

true class	m	g	d	i	p	omit	recall
[m]icroscopy	1563	19	11	13	34	94	0.90
[g]raph	19	919	1	31	38	48	0.87
[d]iffraction	112	3	75	8	3	8	0.36
[i]llustration	50	32	3	164	23	11	0.58
[p]arent	87	30	3	28	113	12	0.41
precision	0.85	0.92	0.81	0.67	0.54	0.90	mean TP IOU

no threshold, N = 3555

(b)

predicted class

true class	m	g	d	i	p	omit	recall
[m]icroscopy	1374	3	2	5	6	344	0.79
[g]raph	8	802		2		244	0.76
[d]iffraction	63		30	2		114	0.14
[i]llustration	13	8		62	1	199	0.22
[p]arent	32	5		4	29	203	0.11
precision	0.92	0.98	0.94	0.83	0.81	0.92	mean TP IOU

high threshold, N = 3555

Figure 6.3. The confusion matrices highlight the nature of the mistakes made in each classification scenario at two confidence thresholds (a) no threshold, and (b) high-threshold for N=3555 images. In both cases, the precision scores are adequate, particularly in the case of the microscopy, graph, and diffraction images. Recall suffers across the board as correctness is emphasized over completeness in the design of the pipeline. Images in (c) highlight some of the more easily rationalized examples of false positive microscopy classifications. DOIs for articles containing the example images (left to right): 10.1038/ncomms14925, 10.1038/srep08722, 10.1038/ncomms4631.

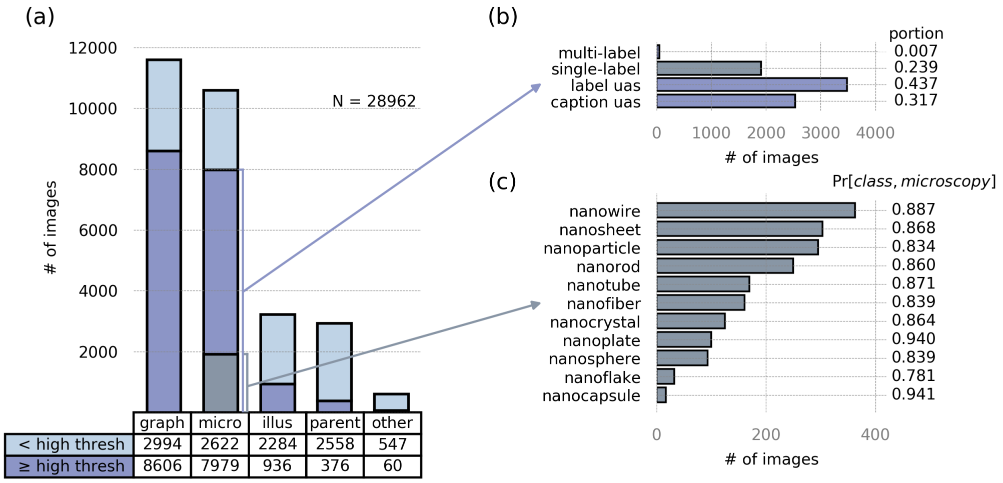
code for the scrapers is easily extendable to other journal sources), and positive predictive value (precision), which is loosely the correctness of the positive classification, is always prioritized over a measure of completeness, such as recall. Confusion matrices summarizing important aspects of the classification performance are given in both a “no threshold” (Figure 6.3a) and “high-threshold” (Figure 6.3b) condition. In the “no threshold” condition, the most likely classification in the final output of the neural network is accepted, regardless of its magnitude. In the “high-threshold” condition, only classifications with values of magnitude greater than or equal to 0.99 are accepted. In this context, the threshold is a proxy for classification confidence in the *figure separator*. Both microscopy and graph classification with “no threshold” and “high-threshold” specification are favorable from a precision perspective with all scores more than 0.80 – and furthermore, the “high-threshold” precision for graphs is ∼0.98.

One of the primary use cases for the EXSCLAIM! toolkit is the construction of self-labeled datasets, and in the assumption of data abundance (*i.e.*, the opportunity cost of passing up on an image is low), the low recall values from false negatives (particularly diffraction images identified as microscopy images) are not as detrimental to the integrity of the set as images that are incorrectly identified as an image type they are not (*i.e.*, false positives). Figure 6.3c highlights some of the interesting false positive trends where a sizable population of diffraction, illustration, and parent classes are being incorrectly assigned as microscopy images. In the case of the diffraction example (left), the diffraction patterns show periodicity resembling atomic-resolution microscopy images, so the microscopy assignment seems logical. The false positive for an illustration (middle) maintains some features commonly associated with a microscopy image, such as a darker background, and even the coloring of the graphene sheet resembles a texture present in microscopy images, but globally is clearly not a microscopy image. Finally, the parent on the far right (Figure 6.3c) actually does contain a microscopy image, but because the image below it does not contain a subfigure label and is “semantically” tied to the microscopy image in “a”, the most appropriate classification for this image is “parent”.

6.3.2. Sample Query: Electron Microscopy Images of Nanostructures

We illustrate the utility of EXSCLAIM! for labeling materials imaging datasets with an example of electron microscopy images of nanostructures. Open source Nature articles were collected from a "Sort By Relevance" search related to the collection of queries formed from the following lists of keywords: ("electron microscope", "electron microscopy"), and ("nanoparticle", "nanosheet", "nanoflake", "nanorod", "nanotube", "nanoplate", "nanocrystal", "nanowire", "nanosphere", "nanocapsule", "nanofiber"). This specific query mimics a wildcard-style search for nanostructures imaged in an electron microscopy modality and returned a total of 13,450 open-source articles with 83,504 figure-caption pairs. For the purpose of quantifying overall retrieval performance on microscopy images, which involves both an assessment of image classification and keyword labeling accuracy, we restrict the following quantitative measurements to articles in the top 10% of the relevancy ranked list, which is a reasonable simplification because the labels defined by the search query ("nanoparticle, "nanowire", *etc.*) depend on the presence of the keyword in the caption, and the median keyword frequency decays exponentially across article rank (refer to the supplementary information). This collection of articles has a yield of 29,096 separate images. The full dataset returned from this search query of nanostructure images is available from the Materials Data Facility [230, 231] (DOI: 10.18126/v7bl-lj1n [233]).

Graphs and microscopy images are among the most popular image types for this specific search query. The abundance of microscopy images is expected, as a result of including microscopy-relevant keywords as a separate word family, whereas the high frequency of graphs is most likely a result of how authors choose to format scientific results. Figure 6.4a highlights the distribution of predicted image types in the top 10% of retrieved



	graph	micro	illus	parent	other
< high thresh	2994	2622	2284	2558	547
≥ high thresh	8606	7979	936	376	60

Figure 6.4. The example query is used to extract electron microscopy images of general nanostructures from Nature journals. The bar plot in (a) shows the distribution of image types extracted at two different thresholds. The bar plots in (b) and (c) further subdivide the population of high confidence microscopy images. In (b), the distribution of label types are recorded. In this context, single and multi-label refer the existence of a keyword label. Label unassigned (uas) means that caption text has been distributed to the image, but no keyword label from the initial query exists. Caption unassigned (uas) refers to a scenario where the caption distributor was not able to confidently distribute a proper substring to caption text. The bar plot in (c) represents the distribution of labels in the top 10% of retrieved microscopy images and provides a estimate of the joint probability that an image classified as microscopy and given the corresponding subsequent label is a true microscopy image represented by the given label.

articles, and indicates the threshold used to assign the classification with further color-coded divisions of each bar. These thresholds ("no threshold" and "high-threshold") again act as proxy for classification confidence and are consistent with the definitions given in the evaluation of the *figure separator* on the MTurk dataset. Within both graphs and microscopy images, 75% of the total population receives a high-level of confidence associated with the prediction of its image type (*i.e.*, high-threshold condition), which is likely a consequence of having a more precisely defined image type. This distribution of predicted image types in Figure 6.4a is useful for quantifying the approximate number or percentage of high-confidence extractions (image classification) that one could expect to obtain for a given query. Moreover, taken with the results on the MTurk images from Figure 6.3, it is likely that a large fraction of the high-confidence classifications, particularly in the case of microscopy images and graphs, are actual instances of microscopy images or graphs because of the high precision scores.

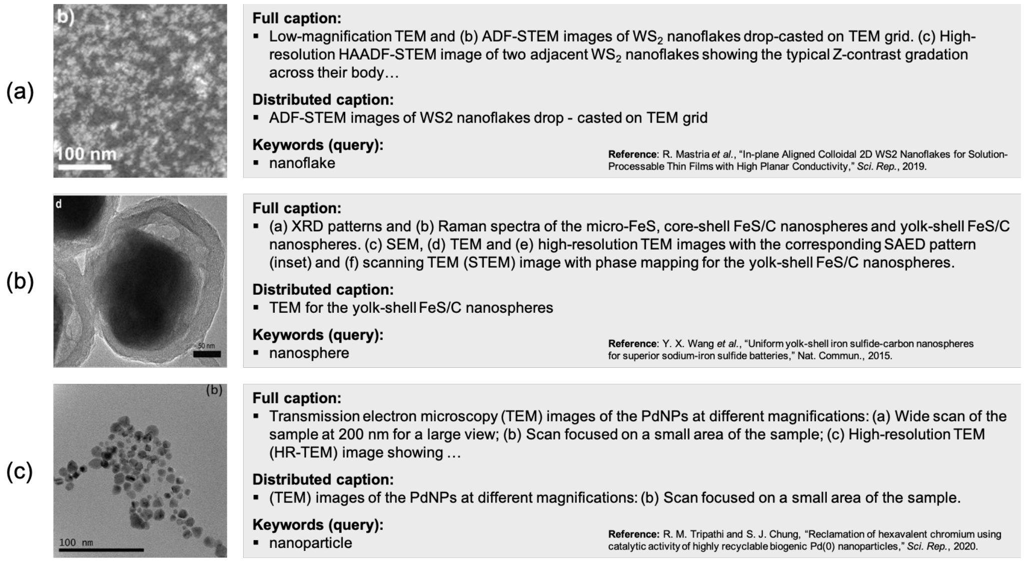
Beyond just image classification confidence, it is important to start examining both the frequency and quality of the processed caption text from the *caption distributor*, because it is this text that is used to ultimately describe the image content (*i.e.*, it is the difference between constructing a generic dataset of microscopy images vs. a highly specific dataset of atomic resolution microscopy images of Ag nanoparticles). First, we examine the frequency at which the processed text is assigned to images and look further at the nature of the text extracted (*e.g.*, is it a single keyword? multiple keywords? sentence or phrase?). Figure 6.4b identifies four categories of possible image labeling. The "label uas" (label unassigned) represents a population of the separated images that have received a portion of the caption text as a description, but do not contain any of the keywords from the search query. If this is problematic from a throughput perspective, it is advised to include more search terms as part of the user query. Fortunately, with the modular design of the pipeline, any of the modules can be re-run on existing extraction results, so this inclusion of more words to increase the percentage of assigned labels could be resolved in an efficient manner (*i.e.*, doesn't require user to start dataset construction from scratch). The "caption uas" (caption unassigned) category, contains separated images that have not received a portion of the caption text at all. These captions left unassigned typically have sentence structure complexities, such as multiple compound subjects, or long intervening phrases, *etc.*, that make caption distribution somewhat ambiguous. And while the percentage of "caption uas" images is somewhat high, this is to some degree intentional, as the data abundance assumption and "no information is better than bad information" design principles underscore many of the extraction steps in the pipeline for the purposes of keeping the data clean. Also, the current regular expression approach taken in caption distribution provides users with the ability to extend or fine tune the system to better suit individual use cases. The final conditions, identified in Figure 6.4b, are the single and multi-label conditions, which represent images containing assigned caption text with explicit reference to one or more of the keywords from the search query.

The single-label condition is further broken down in Figure 6.4c, and its relation to the initial full set of extracted microscopy images is emphasized with the gray coloring in Figure 6.4a (∼24% of the images in the high-threshold group have a single keyword label). The bar chart shows the distribution of images assigned to each keyword, as well as a measure of the joint probability of both the predicted microscopy classification and keyword as being correctly identified. The average joint probability of the positive microscopy/keyword prediction computed across all classes is ∼87%, which means that of the 1909 single-label microscopy images, 1660 are true microscopy images with appropriate keyword labels. Across the entirety of the extracted images (beyond the results for the top 10% recorded here), there are approximately 4300 microscopy images with a single keyword explicitly related to the query. With the $\sim$ 87% joint probability of the positive microscopy/keyword prediction, we would expect a full dataset for this search query to contain approximately 3725 high-confidence microscopy images with the appropriate keyword label.

Current caption distribution prioritizes identification of appropriate keywords over language correctness, so only keyword accuracy is factored into the ground truth comparisons in Figure 6.4. There are, however, many examples where both the keywords are successfully extracted, and the distributed caption is grammatically sufficient, and this even extends to scenarios where the caption text assignment contains non-contiguous segments. For example, in Figure 6.5a, the distributed caption uses all the text from the subfigure identifier "(b)" to the period signifying the end of the sentence, however Figure 6.5b-c show that the current method is capable of parsing the sentence at a higher structural level and pick out the structurally closest sequence of text, even when it is not linearly the closest. For example, in the Figure 6.5c, the subject of the sentence "TEM images" is not matched to the linearly closest descriptions (adjectives) or even closest preposition, but rather, to the preposition at the end of the sentence. Structurally, this makes sense and is the proper division of this caption text.

6.3.3. Extraction of Scaling Information from Images

Magnification of objects is fundamental to microscopic imaging. With suitable estimation of object magnification, achieved by recognizing and interpreting the scale bar length and accompanying text in the separated subfigure images, researchers can begin quantifying image content based on size. In addition, searches through resulting databases can be refined to a particular size range, giving users greater control over results. Further modeling can combine scale and caption information to associate keyword terms with dimensions.



Full caption:
- Low-magnification TEM and (b) ADF-STEM images of WS_2 nanoflakes drop-casted on TEM grid. (c) High-resolution HAADF-STEM image of two adjacent WS_2 nanoflakes showing the typical Z-contrast gradation across their body…

Distributed caption:
- ADF-STEM images of WS2 nanoflakes drop - casted on TEM grid

Keywords (query):
- nanoflake

Reference: R. Mastria *et al.*, "In-plane Aligned Colloidal 2D WS2 Nanoflakes for Solution-Processable Thin Films with High Planar Conductivity," *Sci. Rep.*, 2019.

Full caption:
- (a) XRD patterns and (b) Raman spectra of the micro-FeS, core-shell FeS/C nanospheres and yolk-shell FeS/C nanospheres. (c) SEM, (d) TEM and (e) high-resolution TEM images with the corresponding SAED pattern (inset) and (f) scanning TEM (STEM) image with phase mapping for the yolk-shell FeS/C nanospheres.

Distributed caption:
- TEM for the yolk-shell FeS/C nanospheres

Keywords (query):
- nanosphere

Reference: Y. X. Wang *et al.*, "Uniform yolk-shell iron sulfide-carbon nanospheres for superior sodium-iron sulfide batteries," Nat. Commun., 2015.

Full caption:
- Transmission electron microscopy (TEM) images of the PdNPs at different magnifications: (a) Wide scan of the sample at 200 nm for a large view; (b) Scan focused on a small area of the sample; (c) High-resolution TEM (HR-TEM) image showing …

Distributed caption:
- (TEM) images of the PdNPs at different magnifications: (b) Scan focused on a small area of the sample.

Keywords (query):
- nanoparticle

Reference: R. M. Tripathi and S. J. Chung, "Reclamation of hexavalent chromium using catalytic activity of highly recyclable biogenic Pd(0) nanoparticles," *Sci. Rep.*, 2020.

Figure 6.5. The extracted images contain both a keyword label shared with the initial query, as well as a grammatically sufficient sequence of distributed caption text. Some caption distribution examples are simple, as in (a), where all the distributed words are linearly connected. However, the current *caption distributor* is also designed to capture more complex structural dependence relationships (b-c) where the subject is separated from the text completing the full consistent description.

To quantify the accuracy of the scale detection step, 440 figures containing 920 scale bar lines and scale bar labels from the MTurk dataset, withheld during training, were used. The predicted pixel length of the scale bar line differed from the ground truth value by a mean absolute error of 5.4%. This level of error is similar to that of humans performing the task and negligible when determining the approximate scale of the image. Scale bar label recognition overall is 92% accurate (both number and unit) (See Figure 6.9 in the SI) when the confidence threshold is 0.2. This is a reasonable accuracy given the variance in the quality of the scale bar text itself. It is not uncommon for authors to leave the default scale bar text untouched when a microscopy image is included as part of a compound figure – this is problematic because the native text size is often too small for high quality visualization, and there are many instances where the color of the text is similar to the image content it sits on top of (*i.e.*, recognition suffers from very low contrast). Refer to some of the examples included in the supplementary information which support these general observations. The supplementary information also includes confusion matrices to summarize the prediction results breaking each label down by number and unit.

6.3.4. Self-Labeling with NLP Tools

We have demonstrated the effectiveness of the EXSCLAIM! tool in situations where keywords are extracted from the caption text, and we even show that in some situations when the structure of the caption is complex, the *caption distributor* is still capable of extracting sentences and phrases in a way that preserves the intended meaning. However, in the current setup, keywords are defined explicitly by the initial query, so any other related topics that are prominent in the returned set, but not explicitly specified within the query, are ignored. The advantage of treating the problem of dataset construction in this manner is that it ensures a high degree of relevance in the images that are assigned a keyword label, but it does limit the scope of the self-labeling task in general. To this end, we first suggest including more search terms in the query (synonyms in particular), to better capture the breadth of language used to describe the concepts that the user deems important. Here though, we also explore how NLP tools can be leveraged to transform the phrases or sentences of the distributed caption text into a series of relevant, hierarchical labels for each image they are referencing. Specifically, the outlined approach involves the use of two popular techniques in NLP: word embeddings and topic modeling from documents. This gives users another option of generalizing this labeling to other scientific

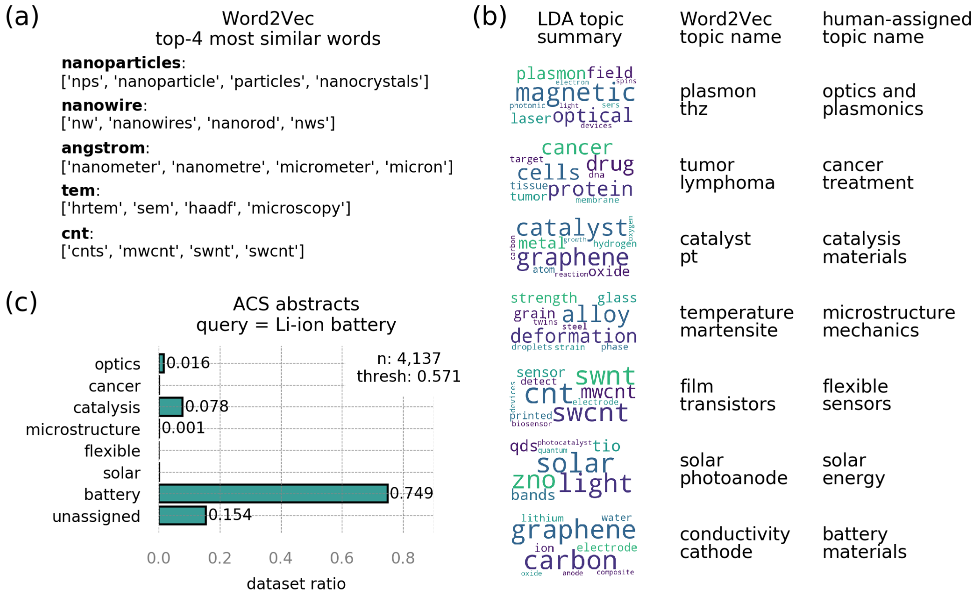


Figure 6.6. (a) Word embedding examples for a Word2Vec model trained on abstract and introduction texts from the nanostructure query. (b) LDA topic modeling applied to the introduction and abstract texts reveals some of the most popular technological applications of nanostructures. The Word2Vec topic name and human-assigned topic name represent further attempts to summarize the words of the topics into more concise titles. (c) Distribution of topics assigned to a group of 4237 abstracts collected from a query of American Chemical Society (ACS) journals for Li-ion batteries.

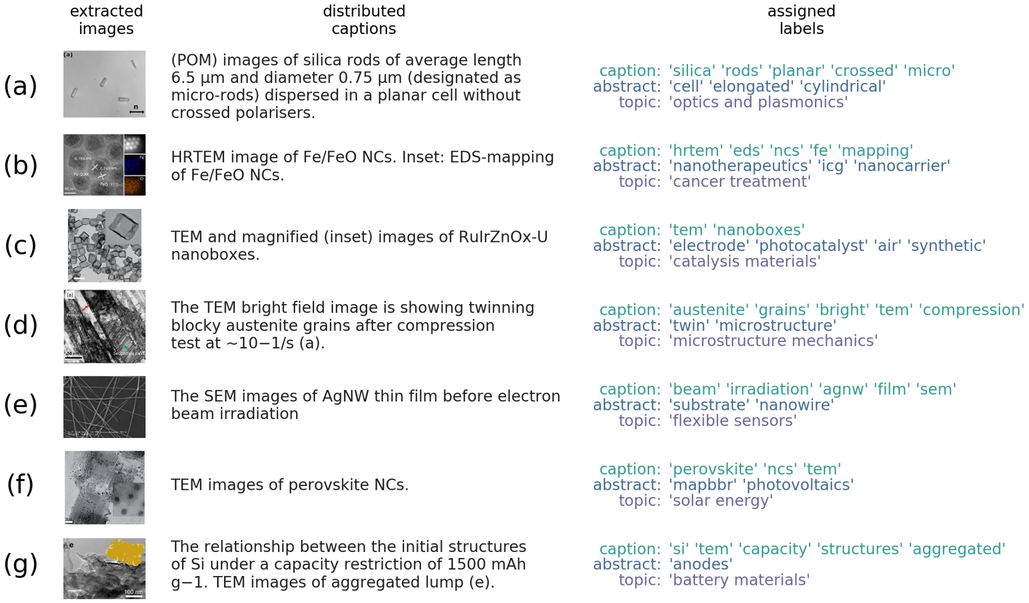
topics that are not explicitly part of the initial query.

Word Embeddings. The goal of word embeddings is to create vector space representations for individual words in such a way that similar words are located close to one another in vector space. EXSCLAIM! leverages the popular unsupervised Word2Vec technique [110] to learn high quality word vectors for the images returned in the nanostructure query, using the 26,683 abstract and introductory paragraphs (*i.e.*, all text before methods are described) from the source articles. Constraining the language to the topics present in the searched articles is appropriate for the goal of image self-labeling. For a more comprehensive embedding of the materials science literature with an information discovery focus, refer to [127]. To demonstrate how Word2Vec can be used for word associations in context of the abstracts and introduction texts, Figure 6.6a highlights simple word lookup examples. Without being explicitly associated, “nanoparticles” and “nanowire” are placed in close proximity to their abbreviations, “nps” and “nw”, respectively. Additionally, 3D “nanoparticles” are closely associated with another 3D nanostructure, such as a “nanocrystal”, and “nanowire” is placed near a similar 2D “nanorod”. The “angstrom” unit of length is placed closely to “nanometer” and “micrometer”, and interestingly, in this case, scale is even preserved in the ordering (i.e. angstrom is closer in scale to a nm than micrometer). These sorts of quantitative relationships are not uncommon and are described in further detail in the original paper [110]. The “tem” and “cnt” lookups provide further demonstrative evidence of the Word2Vec’s effectiveness in creating some notion of proper word associations and context.

Topic Modeling. When all the abstract and introductory texts related to the general nanostructures search query are collected together, certain topics (groups of related words) arise from both high-frequency word occurrences and common word orderings. Perhaps unsurprisingly, these topics reveal many of the popular technological applications of nanostructured materials, and when not explicitly included as part of the search query word families, can provide meaningful extra context. Figure 6.6b illustrates how Latent Dirichlet Analysis (LDA) [123], a popular technique used for topic modeling, is used in context of the nanostructure query. The word clouds, provided to visualize the

LDA output, illustrate the unsupervised clustering of related words into topics. Unfortunately, LDA does not provide a topic name to the words it clusters together. To this end, we illustrate how the trained Word2Vec model can be used to create topic names and how they closely mirror and/or are quickly resolved into rational human-suggested titles. For example, the topic containing “catalyst”, “metal”, “oxide”, “graphene”, “reaction”, etc. is given the Word2Vec topic name of “catalyst”, “pt”, which is easily understood to represent the general class of “catalysis materials”. To demonstrate that the LDA has indeed learned topics in some appropriate fashion, we collected 4137 abstract and introduction texts from a search query of ACS journal family for “Li-ion batteries” and observed that the majority of the documents were categorized explicitly as belonging to batteries, or the highly related/overlapping catalysis category.

Hierarchical Label Assignment. The task of resolving sentences or phrases into relevant image labels is handled using Word2Vec and LDA topic modeling at multiple structural levels. Figure 6.6 provides examples of how an image/caption pair is transformed into a series of hierarchical labels that describe and provide context for the image content. In all examples, the “caption” labels are determined using an iterative word dropout approach that removes words furthest away (measured by cosine similarity) from the center of the current group. The “abstract” labels are the 2-3 words closest to the center of the combination of abstract and caption words, and typically provide context for the image that is not found explicitly in the caption. Finally, the LDA model trained on the abstracts and introduction texts, assigns the best “topic” label to the document containing the image, if a confidence threshold of 0.80 is exceeded. Figure 6.6 provides several



	extracted images	distributed captions	assigned labels
(a)		(POM) images of silica rods of average length 6.5 µm and diameter 0.75 µm (designated as micro-rods) dispersed in a planar cell without crossed polarisers.	caption: 'silica' 'rods' 'planar' 'crossed' 'micro' abstract: 'cell' 'elongated' 'cylindrical' topic: 'optics and plasmonics'
(b)		HRTEM image of Fe/FeO NCs. Inset: EDS-mapping of Fe/FeO NCs.	caption: 'hrtem' 'eds' 'ncs' 'fe' 'mapping' abstract: 'nanotherapeutics' 'icg' 'nanocarrier' topic: 'cancer treatment'
(c)		TEM and magnified (inset) images of RuIrZnOx-U nanoboxes.	caption: 'tem' 'nanoboxes' abstract: 'electrode' 'photocatalyst' 'air' 'synthetic' topic: 'catalysis materials'
(d)		The TEM bright field image is showing twinning blocky austenite grains after compression test at ~10−1/s (a).	caption: 'austenite' 'grains' 'bright' 'tem' 'compression' abstract: 'twin' 'microstructure' topic: 'microstructure mechanics'
(e)		The SEM images of AgNW thin film before electron beam irradiation	caption: 'beam' 'irradiation' 'agnw' 'film' 'sem' abstract: 'substrate' 'nanowire' topic: 'flexible sensors'
(f)		TEM images of perovskite NCs.	caption: 'perovskite' 'ncs' 'tem' abstract: 'mapbbr' 'photovoltaics' topic: 'solar energy'
(g)		The relationship between the initial structures of Si under a capacity restriction of 1500 mAh g−1. TEM images of aggregated lump (e).	caption: 'si' 'tem' 'capacity' 'structures' 'aggregated' abstract: 'anodes' topic: 'battery materials'

Figure 6.7. (a-g) Examples of hierarchical label assignment for images containing a properly distributed caption. For each image, the caption labels are limited to caption text only. Conversely, the abstract labels are free to draw additional relevant words from the abstract text, and the topic labels come from the human-assigned topic names from the LDA topic summaries. — DOIs for articles containing the example images (natural reading order): 10.1038/s41598-019-40198-1,10.1038/s41467-019-12142-4, 10.1038/s41467-019-12885-0, 10.1038/s41598-019-55803-6, 10.1038/srep17716, 10.1038/am2016167, 10.1038/srep42734..

compelling examples of how this approach can provide useful contextual labels for the images from language outside that found explicitly in the distributed caption. For example, we learn things that we can confirm visually, such as the fact that nanorods are elongated and cylindrical (Figure 6.6a). We also learn things that an expert might know that are useful for understanding the function of the image content, such as the facts that: Fe/FeO nanocrystals function as nanotherapeutics (Figure 6.6b); RuIrZnOx-U nanoboxes are synthetic as opposed to biological catalysts (Figure 6.6c); austenite grains describe the "microstructure" of the image (Figure 6.6d); nanocrystals are MA lead halide perovskite (numerical characters get stripped in the text preprocessing, so 'mapbbr' refers to for MAPbBr3) (Figure 6.6f), and even in Figure 6.6e where the abstract context is more redundant than unique or complimentary, the topic provides useful context for where the specific image content appears from an application perspective.

There are a few subtle issues with some of the assigned labels that are a result of some of the known shortcomings in Word2Vec model training. Most notably is that in some cases, similarity is more indicative of how interchangeable/related words are, as opposed to measuring their actual likeness. For example, the nanoboxes in Figure 6.6c are inaccurately labeled as photocatalysts and should be described as electrocatalysts. While these words are highly related and often found in interchangeable contexts (*i.e.*, [photocatalysts/electrocatalysts] facilitate water-splitting ... *etc.*), they can present certain instances where the suggested labeling in problematic. Overall, the combination of Word2Vec modeling with LDA topic discovery provides a solid backbone for self-labeling imaging effort. Future work will involve finding ways to use language components and the imaging analysis jointly to describe image content.

6.4. Conclusion

We present EXSCLAIM!, a software pipeline for the automatic **EX**traction, **S**eparation, and **C**aption-based natural **L**anguage **A**nnotation of **I**mages from scientific figures. In this work, we detail the specific extraction tools and provide quantitative measures of performance for image classification and keyword labeling accuracy on both a crowdsourced-labeled dataset, and an extracted dataset of nanostructure figures from Nature family journals. In addition, we provided discussions and useful model implementations aimed at assigning image labels from complete sentence text. Successful consolidation and self-labeling of images from scientific literature sources will not only enhance the navigation and searchability of images spanning materials, medical, and biological domains, but is a vital first step towards introducing scientific imaging to the canon of training datasets for state-of-the art deep learning and computer vision algorithms.

6.5. Supplementary Information

6.5.1. Data Availability

A dataset used to validate the classification and bounding box prediction accuracy of the *figure separator* component of the EXSCLAIM! pipeline, as presented in Figure 6.3, can be found via the Materials Data Facility (https://doi.org/10.18126/a6jr-yfoq). A dataset illustrating how a sample query submitted to the EXSCLAIM! pipeline can be used to construct a sizable labeled dataset (> 280,000 images) of microscopy images from literature can be found via the Materials Data Facility (https://doi.org/10.18126/v7bl-lj1n). Analysis of image extraction and keyword relevance associated with this dataset is presented Figure 6.4 of the manuscript.

6.5.2. Keyword Frequency Across Retrieved Articles

6.5.3. Accuracy of Scale Bar Label Detection

6.5.4. Word2Vec and LDA Training

Word2Vec (from gensim library[234]) was trained on a corpus of 26,683 abstract and introduction paragraphs from the source articles of the nanostructure query. The following parameters were used for training: `min_count = 25`, `size = 200`, `iter = 500`. For the LDA topic modeling, abstract and introduction text was transformed into a corpus of TFIDF vectors and `LdaMulticore` was used with `num_topics=7`, `id2word=custom_dictionary`, `passes=64`, `workers=4`. The number of topics was selected based on coherence score (`u_mass`), which assess the quality of the learned topics.

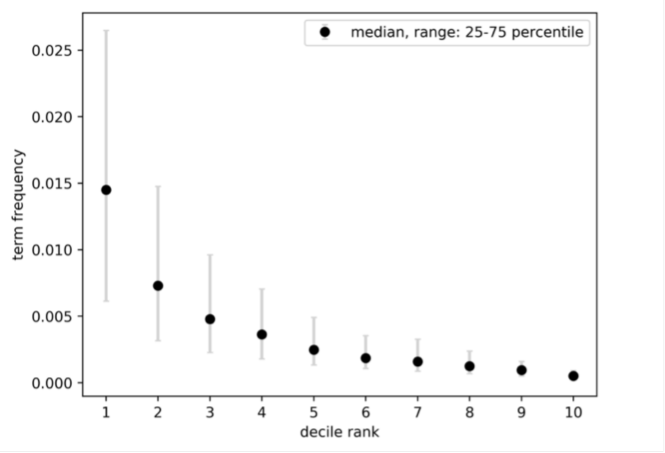


Figure 6.8. The explicit keywords (part of the nanostructure query) were counted in each retrieved article, and the median value of the keyword frequency for each ranked decile was plotted to highlight the exponential decay of keyword occurrences across the full retrieved dataset. This implies that articles returned in the top 10% have the greatest chance of being assigned a keyword label directly from the query.

6.5.5. MTurk GUI for Annotating Figure Separation Training Data

A snapshot of the graphical use interface (GUI) that is accessible via the MTurk worker platform (https://www.mturk.com/worker) is included in Figure 6.10. MTurk workers with an extensive record of positive work on the platform can access and complete the MDI annotation “HITS” (Human Intelligence Tasks) posted to the site. For more details on the MDI model, see Figure 6.2. Workers can select the categories at the bottom that are used to describe the objects in the figure, and after drawing a bounding box around the object, they are asked to either further classify the image, or transcribe the text (if applicable). When one of the members of the EXSCLAIM! project reviews the HIT submitted by the worker, if directions are not followed, or it is clear that the task was completed in a rushed or careless fashion, the HIT is rejected. This helps ensure the quality of the training dataset created for this task of figure separation.

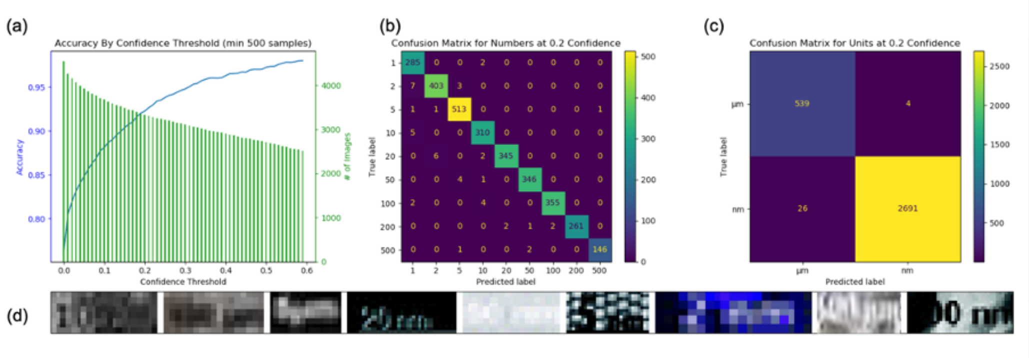
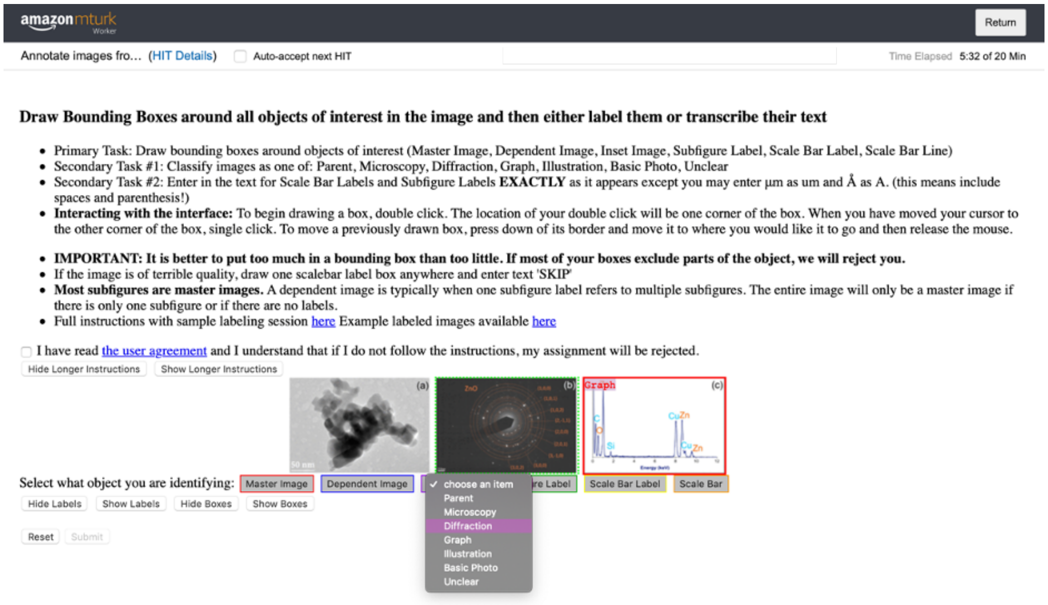


Figure 6.9. The plot in (a) shows how accuracy varies as a function of confidence threshold, including the number of images present at a given threshold. For all thresholds shown, there are at least 500 samples. When the confidence level associated with the scale bar label detection is ∼0.6, the overall accuracy (percentage of scale bar number and unit labels that are correct) is > 0.95. The confusion matrices in (b) and (c) highlight the accuracy of the predicted number and unit components for the scale bar label recognition. Labels for mm and cm were not adequately represented in the training set, so they are not part of the test set. Both number and scale recognition accuracy is high at the 0.2 threshold (∼92% and 99% respectively for the labels shown). The examples in (d) show common instances of the low-resolution and low-contrast conditions that are responsible for a majority of the prediction errors.


amazon mturk
Worker
Return
Annotate images fro... (HIT Details)
□ Auto-accept next HIT
Time Elapsed 5:32 of 20 Min

Draw Bounding Boxes around all objects of interest in the image and then either label them or transcribe their text

- Primary Task: Draw bounding boxes around objects of interest (Master Image, Dependent Image, Inset Image, Subfigure Label, Scale Bar Label, Scale Bar Line)
- Secondary Task #1: Classify images as one of: Parent, Microscopy, Diffraction, Graph, Illustration, Basic Photo, Unclear
- Secondary Task #2: Enter in the text for Scale Bar Labels and Subfigure Labels **EXACTLY** as it appears except you may enter µm as um and Å as A. (this means include spaces and parenthesis!)
- **Interacting with the interface:** To begin drawing a box, double click. The location of your double click will be one corner of the box. When you have moved your cursor to the other corner of the box, single click. To move a previously drawn box, press down of its border and move it to where you would like it to go and then release the mouse.

- **IMPORTANT: It is better to put too much in a bounding box than too little. If most of your boxes exclude parts of the object, we will reject you.**
- If the image is of terrible quality, draw one scalebar label box anywhere and enter text 'SKIP'
- **Most subfigures are master images.** A dependent image is typically when one subfigure label refers to multiple subfigures. The entire image will only be a master image if there is only one subfigure or if there are no labels.
- Full instructions with sample labeling session here Example labeled images available here

□ I have read the user agreement and I understand that if I do not follow the instructions, my assignment will be rejected.

Select what object you are identifying:

Figure 6.10. Screenshot of MTurk interface used for MDI annoation of figures used as training data for the *figure separator*.

CHAPTER 7

Summary and Future Directions

In this thesis, I present a collection of vision-based automation tools that can be used to assist in the prediction of materials structures and associated structural characteristics from experimental and simulated microscopy images. These automation tools are powerful because they eliminate many of the scaling bottlenecks related to manual image comparison found in traditional microscopy interpretation workflows. In particular, I focus on interpretation workflows involving atomic-resolution STEM and STM imaging modalities, as well as those involving general materials microscopy images scraped from open-source scientific literature. The goal of each workflow is to assist in the search of plausible structures to best explain experimental observations. With the developed tools, there is the potential to both realize structure in a direct sense, using image simulation and atomistic modeling, as well as in a more indirect sense, using images curated from literature as the foundation for deep learning or image retrieval tasks that attempt to make sense of observed structures in the context of learned patterns or existing published images. Here I summarize the main topics covered within the chapters of this thesis, and in the end, I provide several directions to take for both extending and improving the overall robustness of the tools developed.

7.1. Summary

One of the primary research challenges of this thesis involves quantifying the similarity between experimentally observed images and simulated materials characterization images. In Chapter 3, we introduce multiple image comparison scenarios in materials microscopy to test similarity measurements, and suggest ways in which pixel value distributions, in combination with a consideration of the search goal, can be used to inform preprocessing and comparison strategies. When the goal of search is to return images that contain identical or near-identical structures to the image that is used as the request for comparison (*i.e.*, the query), visual information fidelity in the pixel domain (VIFP), with normalization as the preprocessing method, achieves the highest matching accuracy scores; however, simple pixel-wise mean square error (MSE) or structural similarity index measure (SSIM) with standardization as the preprocessing method are competitive.

In the event that registration cannot be addressed in a straightforward manner (*i.e.*, out-of-the-box rigid registration algorithms fails), we go on to show how features, constructed to capture the local geometry information surrounding each distinct atomic column, can be used both as a basis for scale-invariant image comparison, and as a way to consolidate similar atomic columns in a chemically and/or functionally logical way (*e.g.*, group atomic columns with similar coordination or atomic columns that are all part of an interface). Finally, in addition to matching structures contained in the images, we provide tools for the creation of idealized STEM distortions and begin to decouple the possible roles that each distortion type plays in further image comparison scenarios.

As mentioned above, in order to focus on quantitative aspects of pixel-level matching highlighted in Chapter 3, assumptions were made to remove registration as a factor in the image similarity comparison. The specific problem of image registration for atomic-resolution imaging is explored in Chapter 4, in context of the *ingrained* toolkit. With *ingrained*, registration is used to find a proper mapping between the simulated image (both the underlying structure and associated simulation parameters) and the experimental image. Structure initialization for bicrystal interfaces is the introductory use case for the *ingrained* toolkit. In this case, *ingrained* iteratively optimizes the misorientation between two grain structures (obtained with programmatic queries of online materials structure databases), alongside a set of image simulation parameters, to produce a bicrystal structure whose simulated image is the best match to the experiment. Conversely, the STM use case introduces an experimentally-rationalized forward model for STM image simulation that relies on *ingrained* for fine-tuning of important imaging-related parameters. The term "experimentally-rationalized" implies that since the STM intensities for a single structure vary widely with small changes in the simulated imaging parameters, the mapping between the simulation and experiment is essential for justifying a proposed structure. In this sense, image registration is part of the forward model itself. Already, the STM use case has proven impactful as it was recently used with first-principles modeling to help determine the structure of hydrogenated borophane. [7].

Even when a satisfactory registration result is obtained (*i.e.*, the interface cases presented in Figures 4.3 and 4.4), the final structure itself often needs further refinement because there is no mechanism to programmatically add structural features and/or defects to compensate for any localized disorder that exists in the experimentally observed structure. This acknowledgment does not diminish the demonstrated capabilities of the *ingrained* toolkit, but rather, motivates the global structure search methods introduced in Chapter 5. In Chapter 5, a proposed registration solution from *ingrained* is used as the starting point in a global structure search workflow that incorporates image simulation and atomistic modeling together for the purpose of (1) enforcing visual consistency between simulation and experimental observation (similar to *ingrained*), and (2) assessing energetic feasibility in the form of atomistic calculations. Both of these components, image matching and structural feasibility assessment, are incorporated into the search objective function. Specifically, in Chapter 5, we show that a global structure search workflow can be used to propose a collection of plausible high-angle grain boundary structures in cadmium telluride (CdTe). In addition, we use first principles calculations of energetic and electronic properties of these grain boundaries along with multislice calculations, to suggest that this global optimization workflow is of benefit in the further study of realistic grain boundary structures. This work in Chapter 5 is a critical component of the FANTASTX software package [52], one of the core theory and modelling capabilities of the Center for Nanoscale Materials at Argonne National Laboratory, which provides a framework to determine atomistic-level structures from multi-modal experimental and theoretical data.

The final topic of this dissertation is a departure from the preceding chapters that follow bottom-up approaches to construct structure interpretation pipelines by combining atomistic modeling with image simulation. In Chapter 6, we introduce EXSCLAIM!, a tool for the automatic **EX**traction, **S**eparation, and **C**aption-based natural **L**anguage **A**nnotation of **IM**ages from scientific figures. Drawing from both the depth and diversity of open-source imaging content available in the scientific literature, EXSCLAIM! allows users the ability to construct high-quality self-annotated imaging datasets at volumes conducive to meaningful experimentation with deep learning algorithms, and thus ascertain structural insight from microscopy in a top-down, data-driven way. In Chapter 6, we summarize the main components of the pipeline, highlighting the role that natural language processing (NLP) and computer vision play in the construction of the components, and providing examples that illustrate the pipeline's overall effectiveness. We conclude with a case study that shows how the combination of word vectorizations and topic modeling can be used to further refine the procedure for assigning keywords to extracted images.

7.2. Future Directions

This thesis established an important foundation for future vision-based automation projects in materials microscopy. Each of the following directions extends one or more of the concepts or tools explored in prior chapters. In some cases, particularly with *ingrained* (Chapter 4) and EXSCLAIM! (Chapter 6), the directions identified are already being addressed in context of new spinoff projects that use these software tools.

7.2.1. CNN-based Peak Detection

In the atomic-resolution images explored in Chapter 3, the positions and intensities of the atomic columns (*i.e.*, "peaks"), were used to construct the image features. In the presentation of the methods, peak finding was trivialized by the fact that the coordinates of the atoms (and therefore the columns) could be found directly from the structures that were simulated in order to create the synthetic dataset. In order for the feature-based comparison results to translate to comparisons with real experimental images, successful peak detection both is critical, and non-trivial. Using morphological and thresholding operations, e.g., from off-the-shelf packages such as scikit-image, is often not reliable, and thus require a good deal of parameter tuning to reduce false positive and false negative peak detections. Given the abundance of open-source CNN-based architectures currently available, in conjunction with the fact that this work established the infrastructure for creating large-scale synthetic datasets of atomic-resolution microscopy images, it would be worth training a CNN for peak-detection on these images so that feature-based matching approaches that rely on peaks could be adequately utilized on a wide-variety of experimental atomic-resolution microscopy images.

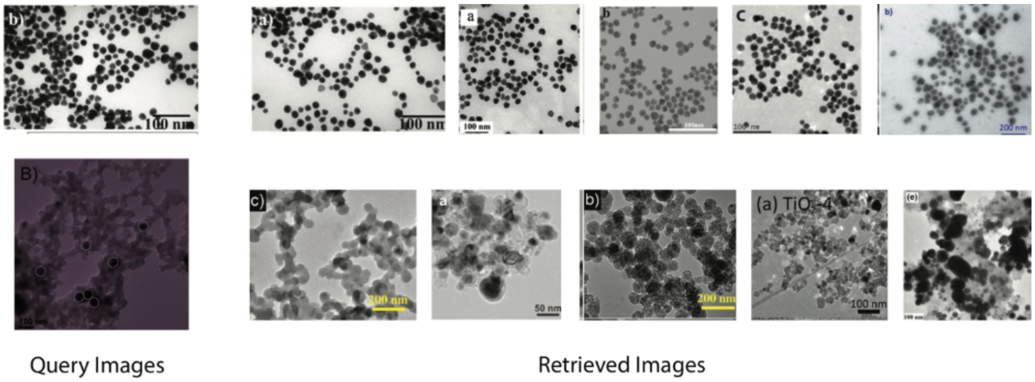


Figure 7.1. Microscopy images from literature are queried with the images presented on the left, and the top matches obtained using MSE comparisons with features computed used the default ResNet-101 backbone are presented on the right. The default ResNet-101 backbone is trained on millions of "natural" images from the popular ImageNet database [8]

7.2.2. Image Similarity Comparisons with CNN-based Features

In recent years, interest in hand-crafted feature-based approaches for image retrieval have subsided in favor of CNN-based approaches. With CNNs, image features are not explicitly defined, but rather, are learned in the course of training a network, and fortunately, training a network for generic image classification or recognition tasks on natural scenes can often yield feature representations that are fundamental enough to vision itself to be effective for more specialized recognition or retrieval tasks. This is the basis of what is referred to as transfer learning, and its value is remarkably apparent even with a simple anecdotal case applied to microscopy image retrieval. In Figure 7.1, we show how a simple forward pass of the ResNet-101 architecture trained on natural images, can be used to perform meaningful microscopy image retrieval with simple MSE comparisons on the fixed-length features that are output. We propose further investigation of CNNs for feature representation that involve fine-tuning of layers in the convolutional base.

7.2.3. Multi-start Strategies and Resource Allocation for *ingrained*

The optimization strategy that underlies the image registration in *ingrained* is local in the sense that it restricts the search to points that are close to the starting location. The optimization is not global so that if the desired solution is not close to the starting point, we do not obtain the global minimum in the objective function. One way this is typically dealt with is to implement a multi-start strategy that terminates the current optimization if progress stagnates, and restarts the optimization at a new starting location. The examples presented in Chapter 4 are the result of several manual randomized restarts, so in order to improve automation in the process, these restarts should be incorporated directly into the optimization workflow. For example, a new starting location could be selected at random when a local minimum is detected, or a memory-based strategy that retains prior solutions or portions of solutions that are deemed beneficial could be implemented. In addition to a multi-start strategy, overall optimization progress would benefit from a restructuring of the code that includes (1) multi-core or multi-thread approaches to better exploit concurrency, and (2) the ability to quantify optimization progress so that computation resources can be appropriately allocated. Fortunately, being an optimization problem that involves many distinct runs to explore a relatively large space of input parameters, the *ingrained* optimization strategy is “embarrassingly parallel”, which makes (1) and (2) more straightforward to address.

7.2.4. Accelerated image simulation algorithms for global optimization

Despite several oversimplifying assumptions, convolution is the default microscopy image simulation method for both *ingrained* (Chapter 4) and the grand canonical basinhopping procedure outlined in Chapter 5. In both cases, but particularly in the latter, even a lenient convergence criterion requires several of thousands if not hundreds of thousands of function evaluations, so simulation execution time is an essential consideration. There are two main directions that could be followed to address the current speed and/or accuracy concerns when image simulation is part of the optimization process. One is to reformulate the expensive multislice calculation (considered a simulation gold standard), to reduce the number of operations and/or the cost associated with each operation, which was approach taken in the original PRISM method [39, 41]. A multithreaded form of the PRISM method (calculations for each probe positions took advantage of a multithreaded work dispatch) was used to construct the *atomagined* simulated microscopy dataset introduced in (Chapter 3). The next possible direction involves hardware-specific accelerations with graphics processing units (GPU). Using GPUs to accelerate microscopy image simulation has been well-addressed [38–40, 188], however, to be of most use to the global optimization efforts, that is, for the entirety of the image simulation, energy calculations, and the governing optimization to benefit most from GPU acceleration, GPU-specific structure reorganizations of the current codes are likely necessary.

7.2.5. EXSCLAIM! Version 2.0

Here we cover some of the new features and testing planned for upcoming releases of EXSCLAIM!

Further Proof-of-Concept. To provide a clear example of how EXSCLAIM! is intended to function in the scientific literature ecosystem as far as its capacity to extract images at high-volume for deep learning applications, we propose using extracted image/keyword pairs to train a CNN for steel microstructure classification. The idea here is to show how the image/keyword pairs returned from queries of Nature, ACS, and RSC journal articles, either by themselves, or augmented with an existing experimental dataset [203], can be used as a basis for image classification performance enhancement. This is not a focus on a specific architecture, but rather, an attempt to show the performance benefit of training with a significant body of literature-based imaging data.

Development of a Web-Based GUI. In its current form, EXSCLAIM! can be instantiated from the command line or from a Python script, however, the visualization of the results, and customization of the parsers and components require some degree of familiarity with programming concepts, which may be challenging for novice users. To this end, we propose developing a web-based Graphical User Interface (GUI) to simplify the construction of a user query, improve the visualization and manipulation of the imaging results, and finally, to provide options for exporting the generated datasets in formats amenable to deep learning pipelines or for database storage (export to a database). A majority of these functionalities are available in the current command line version available on GitHub, however the GUI will hopefully attract researchers interested in using EXSCLAIM!, that would otherwise not go through the effort of downloading the software and learning how to use the command line interface.

Robust Caption Assignment with Deep NLP. In its current form, the *caption distributor* module of EXSCLAIM! is rule-based. As detailed in (Chapter 6), this means that the decisions underlying the extraction and pairing of strings of text to the images they describe uses both proximity to the caption delimiting elements with patterns in part-of-speech (POS) sequences to obtain appropriate descriptions of the images. In NLP, this grouping of words following POS tagging is referred to as “chunking” and is part of a general collection of “sequence labeling” NLP tasks. The current rule-based chunking approach adopted in EXSCLAIM! is consistent with early NLP systems and works to the extent that the dictionary of chunks (POS sequences) is representative of real caption sentence syntax. From testing the first version of the tool, we find that improvements are necessary, as the way authors describe images in the caption, despite being a relatively small body of text, is considerably complex. The presence of long noun phrases (what we refer to a noun chunks in Chapter 6) or the abundance of non-adjacent dependencies, are a few contributors to the overall complexity. State-of-the art approaches to solving sequence labeling problems now often involve variants of either a recurrent neural network (RNNs) or transformer-based architectures. Both are specifically designed to handle sequential data and employ concepts such as memory and attention when training on massive text corpora to achieve impressive results on a variety of higher-level language understanding

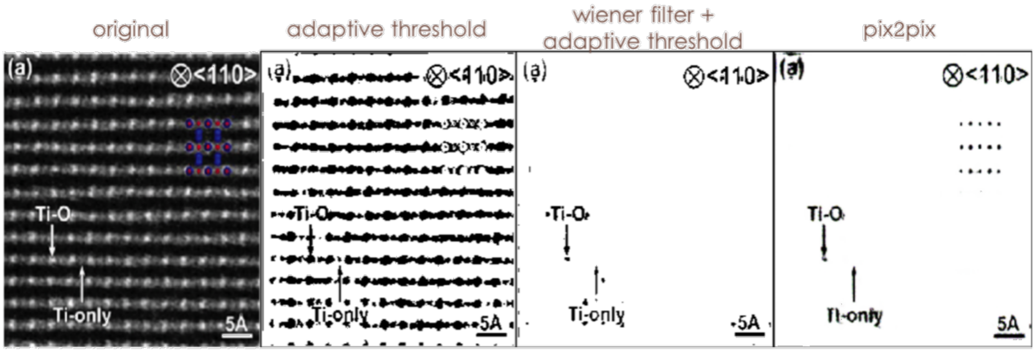


Figure 7.2. An annotated atomic-resolution HAADF STEM image of rutile TiO2 (original image on the left), adapted from [9], with filtering results that illustrate various attempts to separate the annotations drawn on the image (foreground) from the true image content (background). Adaptive thresholding appears to oversegment the image, while Wiener filtering with adaptive thresholding requires considerabel fine tunin. The pix2pix (cGAN approach) with minimal training produces the most accurate segmentation among the variants tested.

tasks – earning them the distinction of being considered “deep NLP”. Going forward, we propose adopting deep NLP methods into EXSCLAIM!, with the hope of being able to construct the best topical phrases and/or summaries of the imaging content. Part of this might also involve extending the scope of the text beyond the caption to also include the full-body text.

Image Annotation Segmentation and Inpainting. In addition to the descriptive text assigned to an extracted image from the caption, researchers often use image annotations as a way to further clarify the identity or function of an object within an image. These annotations, overlaid directly on top of the image content, can take the form of circles, rectangles, arrows, text, *etc*, and could be used to clarify a visualization or indicate heightened importance. In the short term, these annotations are considered problematic, out of a concern that their presence might promote false positive correlations to improve classification or recognition performance in context of training a CNN. For example, an "(a)" in the upper left corner may have some weight in the final classification decision if a majority of the images of that specific class happen to be the first image in a figure. To both avoid promoting false positive correlations when during training, as well as establish a foundation for image annotation interpretation tools going forward, we propose addressing image annotation segmentation (*i.e.*, identifying which pixels contain raw image content, and which pixels are the result of human annotation). Preliminary efforts are already underway to use conditional generative adversarial networks (cGANs), specifically the popular pix2pix setup [96], to construct custom "annotation" filters that can be used to identify portions of an image that contain additional human annotation on top of the raw microscopy output. Figure 7.2 shows some preliminary results of using pix2pix to segment the annotations from the image, and the approach appears promising relative to traditional filtering approaches. Once the annotation locations are identified on the image, they can either be ignored, to avoid the possibility of promoting false positive correlations, or inpainted over. Inpainting algorithms would fill in missing image pixels with values that attempt to appropriately complete the image.

Graph Digitization. Within the current scope of EXSCLAIM!, much of the focus has been on how the tool could be used to amass datasets of microscopy images as a way to facilitate meaningful experimentation with deep learning algorithms further down the pipeline. The current scope has largely ignored the potential contributions of graphical data. Extracting underlying numerical data from graphs (graph digitalization), and pairing it with its associated caption text, would be valuable to those looking to perform further analysis directly on the data themselves. In the literature, identifying distinct instances of objects that all belong to the same class (*e.g.*, a plotted line) in an image is a task referred to as instance segmentation. This is another important step in the effort to make EXSCLAIM! a general tool for image-based information curation in scientific literature.

References

[1] Y. Wu, “An introduction to computer vision,” *EECS 432-Advanced Computer Vision Notes Series 1*, 2017.

[2] Z. Wang and A. C. Bovik, “Modern image quality assessment,” *Synthesis Lectures on Image, Video, and Multimedia Processing*, 2005.

[3] R. Szeliski, “Computer vision : Algorithms and applications,” in *Computer Vision : Algorithms and Applications*, 2009.

[4] J. Guo, A. Mannodi-Kanakkithodi, F. G. Sen, E. Schwenker, E. S. Barnard, A. Munshi, W. Sampath, M. K. Chan, and R. F. Klie, “Effect of selenium and chlorine co-passivation in polycrystalline cdsete devices,” *Applied Physics Letters*, 2019.

[5] X. B. Hu, N. C. Sheng, Y. M. Zhu, J. F. Nie, J. D. Liu, X. F. Sun, and X. L. Ma, “Atomic-scale investigation of the borides precipitated in a transient liquid phase-bonded ni-based superalloy,” *Metallurgical and Materials Transactions A: Physical Metallurgy and Materials Science*, 2020.

[6] R. Zhang, L. Li, L. Frazer, K. B. Chang, K. R. Poeppelmeier, M. K. Chan, and J. R. Guest, “Atomistic determination of the surface structure of cu2o(111): Experiment and theory,” *Physical Chemistry Chemical Physics*, 2018.

[7] Q. Li, V. S. C. Kolluru, M. S. Rahn, E. Schwenker, S. Li, R. G. Hennig, P. Darancet, M. K. Chan, and M. C. Hersam, “Synthesis of borophane polymorphs through hydrogenation of borophene,” *Science*, vol. 371, pp. 1143–1148, 3 2021.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[9] N. Shibata, A. Goto, K. Matsunaga, T. Mizoguchi, S. Findlay, T. Yamamoto, and Y. Ikuhara, “Interface structures of gold nanoparticles on tio 2 (110),” *Physical review letters*, vol. 102, no. 13, p. 136105, 2009.

[10] O. L. Krivanek, N. Dellby, and A. R. Lupini, “Towards sub-Å electron beams,” *Ultramicroscopy*, 1999.

[11] N. Dellby, L. Krivaneka, D. Nellist, E. Batson, and R. Lupini, “Progress in aberration-corrected scanning transmission electron microscopy,” *Journal of Electron Microscopy*, 2001.

[12] P. E. Batson, N. Dellby, and O. L. Krivanek, “Sub-ångstrom resolution using aberration corrected electron optics,” *Nature*, vol. 418, no. 6898, p. 617, 2002.

[13] O. L. Krivanek, J. P. Ursin, N. J. Bacon, G. J. Corbin, N. Dellby, P. Hrncirik, M. F. Murfitt, C. S. Own, and Z. S. Szilagyi, "High-energy-resolution monochromator for aberration-corrected scanning transmission electron microscopy/electron energy-loss spectroscopy," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2009.

[14] P. W. Hawkes, *Advances in Imaging and Electron Physics: Aberration-Corrected Electron Microscopy*. Academic Press, 2009.

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," 2014.

[16] S. Goedecker, "Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems," *Journal of Chemical Physics*, 2004.

[17] A. Laio and M. Parrinello, "Escaping free-energy minima," *Proceedings of the National Academy of Sciences of the United States of America*, 2002.

[18] S. T. Call, D. Y. Zubarev, and A. I. Boldyrev, "Global minimum structure searches via particle swarm optimization," *Journal of Computational Chemistry*, 2007.

[19] D. J. Wales and J. P. Doye, "Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms," *Journal of Physical Chemistry A*, 1997.

[20] T. Back, *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford university press, 1996.

[21] P. Juhás, C. L. Farrow, X. Yang, K. R. Knox, and S. J. Billinge, "Complex modeling: A strategy and software program for combining multiple information sources to solve ill posed structure and nanostructure inverse problems," *Acta Crystallographica Section A: Foundations and Advances*, 2015.

[22] N. J. Mitra and M. Pauly, "Shadow art," *ACM Transactions on Graphics*, 2009.

[23] R. Alakbarov, "Looking at two cities from one point of view. 2002," 2007.

[24] M. C. Swain and J. M. Cole, "Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature," *Journal of chemical information and modeling*, vol. 56, no. 10, pp. 1894–1904, 2016.

[25] E. Kim, K. Huang, A. Tomala, S. Matthews, E. Strubell, A. Saunders, A. McCallum, and E. Olivetti, "Machine-learned and codified synthesis parameters of oxide materials," *Scientific data*, vol. 4, p. 170127, 2017.

[26] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," 2017.

[27] V. Kougia, J. Pavlopoulos, P. Papapetrou, and M. Gordon, "Rtex: A novel methodology for ranking, tagging, and explanatory diagnostic captioning of radiography exams," *arXiv*, 2020.

[28] M. Knoll and E. Ruska, "Beitrag zur geometrischen elektronenoptik. i," *Annalen der Physik*, 1932.

[29] M. Knoll and E. Ruska, “Beitrag zur geometrischen elektronenoptik. ii,” *Annalen der Physik*, 1932.

[30] S. J. Pennycook, “Structure determination through z-contrast microscopy,” *Advances in Imaging and Electron Physics*, 2002.

[31] P. Hartel, H. Rose, and C. Dinges, “Conditions and reasons for incoherent imaging in stem,” *Ultramicroscopy*, 1996.

[32] D. A. Muller, “Structure and bonding at the atomic scale by scanning transmission electron microscopy,” *Nature Materials*, 2009.

[33] S. Hillyard and J. Silcox, “Detector geometry, thermal diffuse scattering and strain effects in adf stem imaging,” *Ultramicroscopy*, 1995.

[34] D. E. Jesson and S. J. Pennycook, “Incoherent imaging of thin specimens using coherently scattered electrons,” *Proceedings - Royal Society of London, A*, 1993.

[35] E. J. Kirkland, *Advanced computing in electron microscopy: Second edition*. 2010.

[36] A. H. Combs, J. J. Maldonis, J. Feng, Z. Xu, P. M. Voyles, and D. Morgan, “Fast approximate stem image simulations from a machine learning model,” *Advanced Structural and Chemical Imaging*, 2019.

[37] J. M. Cowley and A. F. Moodie, “The scattering of electrons by atoms and crystals. i. a new theoretical approach,” *Acta Crystallographica*, 1957.

[38] Y. Yao, B. H. Ge, X. Shen, Y. G. Wang, and R. C. Yu, “Stem image simulation with hybrid cpu/gpu programming,” *Ultramicroscopy*, vol. 166, pp. 1–8, 2016.

[39] A. Pryor, C. Ophus, and J. Miao, “A streaming multi-gpu implementation of image simulation algorithms for scanning transmission electron microscopy,” *Advanced structural and chemical imaging*, vol. 3, no. 1, p. 15, 2017.

[40] M. Radek, J. G. Tenberge, S. Hilke, G. Wilde, and M. Peterlechner, “Stemcl–a multi-gpu multislice algorithm for simulation of large structure and imaging parameter series,” *Ultramicroscopy*, 2018.

[41] C. Ophus, “A fast image simulation algorithm for scanning transmission electron microscopy,” *Advanced structural and chemical imaging*, vol. 3, no. 1, p. 13, 2017.

[42] J. Tersoff and D. R. Hamann, “Theory of the scanning tunneling microscope,” *Physical Review B*, 1985.

[43] G. Kresse and J. Furthmüller, “Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set,” *Physical Review B - Condensed Matter and Materials Physics*, 1996.

[44] *Fifty Years of X-Ray Diffraction*. 1962.

[45] S. J. Billinge, “The rise of the x-ray atomic pair distribution function method: A series of fortunate events,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2019.

[46] S. Billinge, “The nanostructure problem,” *Physics*, 2010.

[47] S. J. Billinge and I. Levin, “The problem with determining atomic structure at the nanoscale,” *Science*, 2007.

[48] B. Narayanan, K. Sasikumar, Z. G. Mei, A. Kinaci, F. G. Sen, M. J. Davis, S. K. Gray, M. K. Chan, and S. K. Sankaranarayanan, "Development of a modified embedded atom force field for zirconium nitride using multi-objective evolutionary optimization," *Journal of Physical Chemistry C*, 2016.

[49] K. Sasikumar, B. Narayanan, M. Cherukara, A. Kinaci, F. G. Sen, S. K. Gray, M. K. Chan, and S. K. Sankaranarayanan, "Evolutionary optimization of a charge transfer ionic potential model for ta/ta-oxide heterointerfaces," *Chemistry of Materials*, 2017.

[50] A. Kinaci, B. Narayanan, F. G. Sen, M. J. Davis, S. K. Gray, S. K. Sankaranarayanan, and M. K. Chan, "Unraveling the planar-globular transition in gold nanoclusters through evolutionary search," *Scientific Reports*, 2016.

[51] M. Yu, A. B. Yankovich, A. Kaczmarowski, D. Morgan, and P. M. Voyles, "Integrated computational and experimental structure refinement for nanoparticles," *ACS Nano*, 2016.

[52] V. Kolluru, "Fantastx." `https://github.com/MaterialEyes/fantastx`, 2021.

[53] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, 1970.

[54] M. Jiang, Q. Zeng, T. Zhang, M. Yang, and K. A. Jackson, "Icosahedral to double-icosahedral shape transition of copper clusters," *Journal of Chemical Physics*, 2012.

[55] L. O. Paz-Borbón, V. T. Mortimer-Jones, R. L. Johnston, A. Posada-Amarillas, G. Barcaro, and A. Fortunelli, "Structures and energetics of 98 atom pd-pt nanoalloys: Potential stability of the leary tetrahedron for bimetallic nanoparticles," *Physical Chemistry Chemical Physics*, 2007.

[56] M. L. Mauldin, "Maintaining diversity in genetic search.," 1983.

[57] C. M. Fonseca and P. J. Fleming, "An overview of evolutionary algorithms in multiobjective optimization," *Evolutionary Computation*, 1995.

[58] E. J. Hughes, "Multiple single objective pareto sampling," 2003.

[59] O. L. De Weck, "Multiobjective optimization: History and promise," 0.

[60] J. Frank, R. Hegerl, W. Hoppe, M. S. Isaacson, D. Kopf, J. E. Mellema, W. O. Saxton, M. Utlaut, and R. H. Wade, *Computer processing of electron microscope images*, vol. 13. Springer Science and Business Media, 2012.

[61] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising with block-matching and 3d filtering," 2006.

[62] R. H. Bates and M. J. McDonnell, "Image restoration and reconstruction.," *Image restoration and reconstruction.*, 1986.

[63] P. Schiske, "Zur frage der bildrekonstruktion durch fokusreihen," pp. 145–146, 1968.

[64] R. W. Gerchberg and W. O. Saxton, "Practical algorithm for the determination of phase from image and diffraction plane pictures.," *Optik (Stuttgart)*, 1972.

[65] B. Girod, "What's wrong with mean-squared error?," *Digital images and human vision*, pp. 207–220, 1993.

[66] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Transactions on Communications*, 1995.

[67] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, 1981.

[68] B. D. Lucas and T. Kanade, "Iterative image registration technique with an application to stereo vision.," 1981.

[69] Z. Ni, C. Pacoret, R. Benosman, S. Ieng, and S. Régnier, "Asynchronous event-based high speed vision for microparticle tracking," *Journal of Microscopy*, 2012.

[70] E. Jurrus, M. Hardy, T. Tasdizen, P. T. Fletcher, P. Koshevoy, B. C. Chien, W. Denk, and R. Whitaker, "Axon tracking in serial block-face scanning electron microscopy," *Medical Image Analysis*, 2009.

[71] V. Abrishami, J. Vargas, X. Li, Y. Cheng, R. Marabini, C. O. S. Sorzano, and J. M. Carazo, "Alignment of direct detection device micrographs using a robust optical flow approach," *Journal of Structural Biology*, 2015.

[72] B. L. DeCost and E. A. Holm, "A computer vision approach for automated analysis and classification of microstructural image data," *Computational materials science*, vol. 110, pp. 126–133, 2015.

[73] A. Chowdhury, E. Kautz, B. Yener, and D. Lewis, "Image driven machine learning methods for microstructure recognition," *Computational Materials Science*, 2016.

[74] N. Lubbers, T. Lookman, and K. Barros, "Inferring low-dimensional microstructure representations using convolutional neural networks," *Physical Review E*, 2017.

[75] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 1998.

[76] M. Ziatdinov, O. Dyck, A. Maksov, X. Li, X. Sang, K. Xiao, R. R. Unocic, R. Vasudevan, S. Jesse, and V. S. Kalinin, "Deep learning of atomically resolved scanning transmission electron microscopy images: Chemical identification and tracking local transformations," *ACS Nano*, 2017.

[77] W. Li, K. G. Field, and D. Morgan, "Automated defect analysis in electron microscopic images," *npj Computational Materials*, vol. 4, no. 1, p. 36, 2018.

[78] W. Xu and J. M. LeBeau, "A deep convolutional neural network to analyze position averaged convergent beam electron diffraction patterns," *Ultramicroscopy*, 2018.

[79] J. M. Ede and R. Beanland, "Partial scanning transmission electron microscopy with deep learning," *Scientific Reports*, 2020.

[80] P. M. Voyles, "Informatics and data science in materials microscopy," *Current Opinion in Solid State and Materials Science*, 2017.

[81] M. Chen, ed., *Computer Vision for Microscopy Image Analysis.* Academic Press, 2020.

[82] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[83] L. Zheng, Y. Yang, and Q. Tian, “Sift meets cnn: A decade survey of instance retrieval,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[84] M. P. Eckert and A. P. Bradley, “Perceptual quality metrics applied to still image compression,” *Signal Processing*, 1998.

[85] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, *et al.*, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[86] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” vol. 3, pp. iii—709, 2004.

[87] Y. Rubner, C. Tomasi, and L. J. Guibas, “Earth mover’s distance as a metric for image retrieval,” *International Journal of Computer Vision*, 2000.

[88] D. Bertsimas, *Introduction to linear optimization.* 2012.

[89] A. Klein, S. S. Ghosh, B. Avants, B. T. Yeo, B. Fischl, B. Ardekani, J. C. Gee, J. J. Mann, and V. R. Parsey, “Evaluation of volume-based and surface-based brain image registration methods,” *NeuroImage*, 2010.

[90] W. R. Crum, T. Hartkens, and D. L. Hill, “Non-rigid image registration: Theory and practice,” *British Journal of Radiology*, 2004.

[91] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning.* MIT Press, 2016. \url{http://www.deeplearningbook.org}.

[92] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” 2009.

[93] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” 2016.

[94] J. Redmon and A. Farhadi, “Yolo v.3,” *Tech report*, 2018.

[95] M. Mirza and S. Osindero, “Cgan,” *CoRR*, 2014.

[96] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” 2017.

[97] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” 2015.

[98] N. Chomsky, “Three models for the description of language,” *IRE Transactions on Information Theory*, 1956.

[99] N. Hornstein and N. Chomsky, “Knowledge of language: Its nature, origin, and use.,” *The Philosophical Review*, 1988.

[100] N. Sager, “Syntactic analysis of natural language,” *Advances in Computers*, 1967.

[101] N. Sager, M. Lyman, N. T. Nhan, and L. J. Tick, “Medical language processing: Applications to patient data representation and automatic encoding,” *Methods of Information in Medicine*, 1995.

[102] P. Zweigenbaum, “Menelas: an access system for medical records using natural language,” *Computer Methods and Programs in Biomedicine*, 1994.

[103] B. Do Amaral Marcio and Y. Satomura, "Associating semantic grammars with the snomed: processing medical language and representing clinical facts into a language-independent frame.," *Medinfo. MEDINFO*, 1995.

[104] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," *Journal of the American Medical Informatics Association*, 2011.

[105] R. B. Lees and N. Chomsky, "Syntactic structures," *Language*, 1957.

[106] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2015.

[107] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," 2013.

[108] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," 2015.

[109] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.

[110] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations ofwords and phrases and their compositionality," 2013.

[111] P. Corbett and P. Murray-Rust, "High-throughput identification of chemistry in life science texts," 2006.

[112] L. Hawizy, D. M. Jessop, N. Adams, and P. Murray-Rust, "Chemicaltagger: A tool for semantic text-mining in chemistry," *Journal of Cheminformatics*, 2011.

[113] D. M. Lowe and R. A. Sayle, "Leadmine: A grammar and dictionary driven approach to entity recognition," *Journal of Cheminformatics*, 2015.

[114] "Leadmine." `https://www.nextmovesoftware.com/leadmine.html`, accessed 2020-04-10.

[115] C. J. Court and J. M. Cole, "Auto-generated materials database of curie and néel temperatures via semi-supervised relationship extraction," *Scientific data*, vol. 5, p. 180111, 2018.

[116] S. Mysore, Z. Jensen, E. Kim, K. Huang, H. S. Chang, E. Strubell, J. Flanigan, A. McCallum, and E. Olivetti, "The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures," 2019.

[117] E. Kim, Z. Jensen, A. Van Grootel, K. Huang, M. Staib, S. Mysore, H. S. Chang, E. Strubell, A. McCallum, S. Jegelka, and E. Olivetti, "Inorganic materials synthesis planning with literature-trained neural networks," *Journal of Chemical Information and Modeling*, 2020.

[118] C. J. Court and J. M. Cole, "Magnetic and superconducting phase diagrams and transition temperatures predicted using text mining and machine learning," *npj Computational Materials*, 2020.

[119] "spacy." `https://spacy.io`.

[120] G. Salton and J. McGill, Michael, "Information retrieval: an introduction," in *Introduction to modern information retrieval*, 1983.

[121] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, 1990.

[122] T. Hofmann, "Probabilistic latent semantic indexing," 1999.

[123] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, 2003.

[124] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, 2019.

[125] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[126] T. Mikolov, W. T. Yih, and G. Zweig, "Linguistic regularities in continuous spaceword representations," 2013.

[127] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain, "Unsupervised word embeddings capture latent knowledge from materials science literature," *Nature*, 2019.

[128] C. Kisielowski, B. Freitag, M. Bischoff, H. Van Lin, S. Lazar, G. Knippels, P. Tiemeijer, M. van der Stam, S. von Harrach, M. Stekelenburg, *et al.*, "Detection of single atoms and buried defects in three dimensions by aberration-corrected electron microscope with 0.5-Åinformation limit," *Microscopy and Microanalysis*, vol. 14, no. 5, pp. 469–477, 2008.

[129] S.-Y. Chung, S.-Y. Choi, T. Yamamoto, and Y. Ikuhara, "Atomic-scale visualization of antisite defects in lifepo 4," *Physical review letters*, vol. 100, no. 12, p. 125502, 2008.

[130] P. M. Voyles, D. A. Muller, J. L. Grazul, P. H. Citrin, and H.-J. Gossmann, "Atomic-scale imaging of individual dopant atoms and clusters in highly n-type bulk si," *Nature*, vol. 416, no. 6883, p. 826, 2002.

[131] V. A. Crewe, J. Wall, and J. Langmore, "Visibility of single atoms," *Science*, vol. 168, no. 3937, pp. 1338–1340, 1970.

[132] L. D. Marks, "Direct atomic imaging of solid surfaces: I. image simulation and interpretation," *Surface Science*, vol. 139, no. 1, pp. 281–298, 1984.

[133] C. M. Breneman, L. C. Brinson, L. S. Schadler, B. Natarajan, M. Krein, K. Wu, L. Morkowchuk, Y. Li, H. Deng, and H. Xu, "Stalking the materials genome: A data-driven approach to the virtual design of nanostructured polymers," *Advanced functional materials*, vol. 23, no. 46, pp. 5746–5752, 2013.

[134] H. Xu, R. Liu, A. Choudhary, and W. Chen, "A machine learning-based design representation method for designing heterogeneous microstructures," *Journal of Mechanical Design*, vol. 137, no. 5, p. 51403, 2015.

[135] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," pp. 2169–2178, 2006.

[136] S. Somnath, C. R. Smith, V. S. Kalinin, M. Chi, A. Borisevich, N. Cross, G. Duscher, and S. Jesse, "Feature extraction via similarity search: application to atom finding and denoising in electron and scanning probe microscopy imaging," *Advanced structural and chemical imaging*, vol. 4, no. 1, p. 3, 2018.

[137] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[138] N. Laanait, M. Ziatdinov, Q. He, and A. Borisevich, "Identifying local structural states in atomic imaging by computer vision," *Advanced structural and chemical imaging*, vol. 2, no. 1, p. 14, 2017.

[139] P. Gao, A. Kumamoto, R. Ishikawa, N. Lugg, N. Shibata, and Y. Ikuhara, "Picometer-scale atom position analysis in annular bright-field stem imaging," *Ultramicroscopy*, vol. 184, pp. 177–187, 2018.

[140] S. Van Der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, "Scikit-image: Image processing in python," *PeerJ*, 2014.

[141] H. Sheikh and A. Bovik, "Visual information fidelity (multi-scale pixel domain implementation)," *https://live.ece.utexas.edu/research/Quality/VIF.htm*, 0.

[142] P. Black, "Dictionary of algorithms and data structures." `http://www.nist.gov/dads`.

[143] G. Bergerhoff and I. D. Brown, "Inorganic crystal structure database," *Acta Crystallographica Section A Foundations of Crystallography*, 1981.

[144] E. Schwenker, F. Sen, C. Wolverton, C. Ophus, and M. K. Chan, "A simulated atomic-resolution haadf stem imaging dataset containing unique icsd structure prototypes," 2020.

[145] D. G. Pelli and P. Bex, "Measuring contrast sensitivity," *Vision research*, vol. 90, pp. 10–14, 2013.

[146] A. Rosenauer, K. Gries, K. Müller, A. Pretorius, M. Schowalter, A. Avramescu, K. Engl, and S. Lutgen, "Measurement of specimen thickness and composition in alx ga1 - x n / gan using high-angle annular dark field images," *Ultramicroscopy*, 2009.

[147] T. Grieb, K. Müller, R. Fritz, M. Schowalter, N. Neugebohrn, N. Knaub, K. Volz, and A. Rosenauer, "Determination of the chemical composition of ganas using stem haadf imaging and stem strain state analysis," *Ultramicroscopy*, 2012.

[148] A. De Backer, G. T. Martinez, A. Rosenauer, and S. Van Aert, "Atom counting in haadf stem using a statistical model-based approach: Methodology, possibilities, and inherent limitations," *Ultramicroscopy*, 2013.

[149] G. T. Martinez, A. Rosenauer, A. De Backer, J. Verbeeck, and S. Van Aert, "Quantitative composition determination at the atomic level using model-based high-angle annular dark field scanning transmission electron microscopy," *Ultramicroscopy*,

2014.

[150] L. Duschek, P. Kükelhan, A. Beyer, S. Firoozabadi, J. O. Oelerich, C. Fuchs, W. Stolz, A. Ballabio, G. Isella, and K. Volz, "Composition determination of semiconductor alloys towards atomic accuracy by haadf-stem," *Ultramicroscopy*, 2019.

[151] T. Paulauskas, F. G. Sen, C. Sun, P. Longo, Y. Zhang, S. W. Hla, M. K. Chan, M. J. Kim, and R. F. Klie, "Stabilization of a monolayer tellurene phase at cdte interfaces," *Nanoscale*, 2019.

[152] M. Pedersen, M. L. Bocquet, P. Sautet, E. Lægsgaard, I. Stensgaard, and F. Besenbacher, "Co on pt(111): Binding site assignment from the interplay between measured and calculated stm images," *Chemical Physics Letters*, 1999.

[153] F. Esch, C. Africh, G. Comelli, S. Fabris, P. Fornasiero, T. Montini, R. Rosei, and L. Zhou, "Electron localization determines defect formation on ceria substrates," *Science*, 2005.

[154] L. Li, R. Zhang, J. Vinson, E. L. Shirley, J. P. Greeley, J. R. Guest, and M. K. Chan, "Imaging catalytic activation of co2 on cu2o (110): A first-principles study," *Chemistry of Materials*, 2018.

[155] M. Gong, S. Zhao, L. Jiao, D. Tian, and S. Wang, "A novel coarse-to-fine scheme for automatic image registration based on sift and mutual information," *IEEE Transactions on Geoscience and Remote Sensing*, 2014.

[156] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1987.

[157] S. Somnath, C. R. Smith, N. Laanait, R. K. Vasudevan, and S. Jesse, "Usid and pycroscopy – open source frameworks for storing and analyzing imaging and spectroscopy data," *Microscopy and Microanalysis*, 2019.

[158] R. Bourne and R. Bourne, "Imagej," in *Fundamentals of Digital Imaging in Medicine*, 2010.

[159] G. Lowe, "Sift - the scale invariant feature transform," *International Journal*, 2004.

[160] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, 1981.

[161] M. Guizar-Sicairos, S. T. Thurman, and J. R. Fienup, "Efficient subpixel image registration algorithms," *Optics Letters*, 2008.

[162] Y. Douini, J. Riffi, A. M. Mahraz, and H. Tairi, "An image registration algorithm based on phase correlation and the classical lucas–kanade technique," *Signal, Image and Video Processing*, 2017.

[163] L. Jones, H. Yang, T. J. Pennycook, M. S. Marshall, S. Van Aert, N. D. Browning, M. R. Castell, and P. D. Nellist, "Smart align—a new tool for robust non-rigid registration of scanning microscope data," *Advanced Structural and Chemical Imaging*, 2015.

[164] J. M. Fitzpatrick, D. L. G. Hill, and C. R. Maurer Jr., "Chapter 8: Image registration," *Handbook of Medical Imaging Vol 2*, 2000.

[165] P. Thévenaz, U. E. Ruttimann, and M. Unser, "A pyramid approach to subpixel registration based on intensity," *IEEE Transactions on Image Processing*, 1998.

[166] A. B. Yankovich, B. Berkels, W. Dahmen, P. Binev, S. I. Sanchez, S. A. Bradley, A. Li, I. Szlufarska, and P. M. Voyles, "Picometre-precision analysis of scanning transmission electron microscopy images of platinum nanocatalysts," *Nature Communications*, 2014.

[167] Y. Wang, Y. Eren Suyolcu, U. Salzberger, K. Hahn, V. Srot, W. Sigle, and P. A. van Aken, "Correcting the linear and nonlinear distortions for atomically resolved stem spectrum and diffraction imaging," *Microscopy*, 2018.

[168] O. Clatz, H. Delingette, I. F. Talos, A. J. Golby, R. Kikinis, F. A. Jolesz, N. Ayache, and S. K. Warfield, "Robust nonrigid registration to capture brain shift from intraoperative mri," *IEEE Transactions on Medical Imaging*, 2005.

[169] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. a. Persson, "The materials project: A materials genome approach to accelerating materials innovation," *APL Materials*, vol. 1, no. 1, p. 11002, 2013.

[170] C. Buurma, F. G. Sen, T. Paulauskas, C. Sun, M. Kim, S. Sivananthan, R. F. Klie, and M. K. Chan, "Creation and analysis of atomic structures for cdte bi-crystal interfaces by the grain boundary genie," 2015.

[171] G. Kresse and J. Furthmüller, "Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set," *Computational Materials Science*, 1996.

[172] M. J. D. Powell, "An efficient method for finding the minimum of a function of several variables without calculating derivatives," *The Computer Journal*, 1964.

[173] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, *et al.*, "Scipy 1.0: fundamental algorithms for scientific computing in python," *Nature methods*, vol. 17, no. 3, pp. 261–272, 2020.

[174] A. Kelly and K. M. Knowles, *Crystallography and Crystal Defects: Second Edition*. 2012.

[175] Z. Sun, C. Huang, J. Guo, J. T. Dong, R. F. Klie, L. J. Lauhon, and D. N. Seidman, "Strain-energy release in bent semiconductor nanowires occurring by polygonization or nanocrack formation," *ACS nano*, vol. 13, no. 3, pp. 3730–3738, 2019.

[176] Z. Yin, T. Kanade, and M. Chen, "Understanding the phase contrast optics to restore artifact-free microscopy images for segmentation," *Medical Image Analysis*, 2012.

[177] "Our technology." `http://www.firstsolar.com/en/Modules/Our-Technology`.

[178] "Best research-cell efficiency chart." `https://www.nrel.gov/pv/cell-efficiency.html`.

[179] P. Ercius, O. Alaidi, M. J. Rames, and G. Ren, "Electron tomography: A three-dimensional analytic tool for hard and soft materials research," *Advanced Materials*,

2015.

[180] M. Welborn, W. Tang, J. Ryu, V. Petkov, and G. Henkelman, “A combined density functional and x-ray diffraction study of pt nanoparticle structure,” *Journal of Chemical Physics*, 2011.

[181] B. Meredig and C. Wolverton, “A hybrid computational-experimental approach for automated crystal structure solution,” *Nature Materials*, 2013.

[182] A. Alkauskas, M. D. McCluskey, and C. G. Van De Walle, “Tutorial: Defects in semiconductors - combining experiment and theory,” *Journal of Applied Physics*, 2016.

[183] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, “Machine learning for molecular and materials science,” *Nature*, 2018.

[184] B. Kolb, L. C. Lentz, and A. M. Kolpak, “Discovering charge density functionals and structure-property relationships with prophet: A general framework for coupling machine learning and first-principles methods,” *Scientific Reports*, 2017.

[185] O. Schütt and J. Vandevondele, “Machine learning adaptive basis sets for efficient large scale density functional theory simulation,” *Journal of Chemical Theory and Computation*, 2018.

[186] H. Chan, B. Narayanan, M. J. Cherukara, F. G. Sen, K. Sasikumar, S. K. Gray, M. K. Chan, and S. K. Sankaranarayanan, “Machine learning classical interatomic potentials for molecular dynamics from first-principles training data,” *The Journal of Physical Chemistry C*, vol. 123, no. 12, pp. 6941–6957, 2019.

[187] F. G. Sen, C. Buurma, T. Paulauskas, C. Sun, M. Kim, S. Sivananthan, R. F. Klie, and M. K. Chan, “Atomistic simulations of grain boundaries in cdte,” 2015.

[188] W. Van den Broek, X. Jiang, and C. T. Koch, “Fdes, a gpu-based multislice algorithm with increased efficiency of the computation of the projected potential,” *Ultramicroscopy*, 2015.

[189] Y. Jiang, E. Padgett, R. Hovden, and D. A. Muller, “Sampling limits for electron tomography with sparsity-exploiting reconstructions,” *Ultramicroscopy*, 2018.

[190] T. H. Choi, R. Liang, C. M. Maupin, and G. A. Voth, “Application of the scc-dftb method to hydroxide water clusters and aqueous hydroxide solutions,” *Journal of Physical Chemistry B*, 2013.

[191] C. Sun, T. Paulauskas, F. G. Sen, G. Lian, J. Wang, C. Buurma, M. K. Chan, R. F. Klie, and M. J. Kim, “Atomic and electronic structure of lomer dislocations at cdte bicrystal interface,” *Scientific reports*, vol. 6, no. 1, pp. 1–12, 2016.

[192] Z. Q. Wang, D. Stroud, and A. J. Markworth, “Monte carlo study of the liquid cdte surface,” *Physical Review B*, 1989.

[193] J. P. Perdew, K. Burke, and M. Ernzerhof, “Generalized gradient approximation made simple,” *Physical Review Letters*, 1996.

[194] J. P. Perdew, K. Burke, and M. Ernzerhof, “Erratum: Generalized gradient approximation made simple (physical review letters (1996) 77 (3865)),” *Physical Review Letters*, 1997.

[195] F. G. Sen, T. Paulauskas, C. Sun, M. Kim, R. F. Klie, and M. K. Chan, "Computational design of dopants in cdte grain boundaries for efficient photovoltaics," 2018.

[196] D. K. Ward, X. W. Zhou, B. M. Wong, F. P. Doty, and J. A. Zimmerman, "Accuracy of existing atomic potentials for the cdte semiconductor compound," *Journal of Chemical Physics*, 2011.

[197] B. Howe, P.-s. Lee, M. Grechkin, S. T. Yang, and J. D. West, "Deep mapping of the visual literature," pp. 1273–1277, 2017.

[198] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," 2014.

[199] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," 2015.

[200] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, 2016.

[201] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," 2017.

[202] J. A. Hachtel, J. C. Idrobo, and M. Chi, "Sub-Ångstrom electric field measurements on a universal detector in a scanning transmission electron microscope," *Advanced Structural and Chemical Imaging*, 2018.

[203] B. L. DeCost, M. D. Hecht, T. Francis, B. A. Webler, Y. N. Picard, and E. A. Holm, "Uhcsdb: Ultrahigh carbon steel micrograph database," *Integrating Materials and Manufacturing Innovation*, 2017.

[204] J. A. Aguiar, M. L. Gong, R. R. Unocic, T. Tasdizen, and B. D. Miller, "Decoding crystallography from high-resolution electron imaging and diffraction datasets with deep learning," *Science Advances*, 2019.

[205] R. Aversa, M. H. Modarres, S. Cozzini, R. Ciancio, and A. Chiusole, "Data descriptor: The first annotated set of scanning electron microscopy images for nanoscience," *Scientific Data*, 2018.

[206] T. Mueller, A. G. Kusne, and R. Ramprasad, "Machine learning in materials science: Recent progress and emerging applications," in *Reviews in Computational Chemistry*, 2016.

[207] V. S. Kalinin, B. G. Sumpter, and R. K. Archibald, "Big-deep-smart data in imaging for guiding materials design," *Nature Materials*, 2015.

[208] S. R. Kalidindi and M. De Graef, "Materials data science: Current status and future outlook," *Annual Review of Materials Research*, 2015.

[209] S. Jesse, M. Chi, A. Belianinov, C. Beekman, V. S. Kalinin, A. Y. Borisevich, and A. R. Lupini, "Big data analytics for scanning transmission electron microscopy ptychography," *Scientific Reports*, 2016.

[210] F. Schroff, A. Criminisi, and A. Zisserman, “Harvesting image databases from the web,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 4, pp. 754–766, 2010.

[211] L.-J. Li and L. Fei-Fei, “Optimol: automatic online picture collection via incremental model learning,” *International journal of computer vision*, vol. 88, no. 2, pp. 147–168, 2010.

[212] X.-S. Hua and J. Li, “Prajna: Towards recognizing whatever you want from images without image labeling,” 2015.

[213] Y. Yao, J. Zhang, F. Shen, L. Liu, F. Zhu, D. Zhang, and H. T. Shen, “Towards automatic construction of diverse, high-quality image datasets,” *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[214] E. Kim, K. Huang, S. Jegelka, and E. Olivetti, “Virtual screening of inorganic materials synthesis parameters with deep learning,” *npj Computational Materials*, vol. 3, no. 1, p. 53, 2017.

[215] K. T. Mukaddem, E. J. Beard, B. Yildirim, and J. M. Cole, “Imagedataextractor: A tool to extract and quantify data from microscopy images,” *Journal of chemical information and modeling*, 2019.

[216] R. F. Murphy, M. Velliste, J. Yao, and G. Porreca, “Searching online journals for fluorescence microscope images depicting protein subcellular location patterns,” pp. 119–128, 2001.

[217] P. Li, X. Jiang, C. Kambhamettu, and H. Shatkay, “Compound image segmentation of published biomedical figures,” *Bioinformatics*, 2018.

[218] M. Taschwer and O. Marques, “Automatic separation of compound figures in scientific articles,” *Multimedia Tools and Applications*, 2018.

[219] S. Tsutsui and D. J. Crandall, “A data driven approach for compound figure separation using convolutional neural networks,” 2017.

[220] X. Shi, Y. Wu, H. Cao, G. Burns, and P. Natarajan, “Layout-aware subfigure decomposition for complex figures in the biomedical literature,” 2019.

[221] A. Ahmed, A. Arnold, L. P. Coelho, J. Kangas, A.-S. Sheikh, E. Xing, W. Cohen, and R. F. Murphy, “Structured literature image finder: Parsing text and figures in biomedical literature,” *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 8, no. 2-3, pp. 151–154, 2010.

[222] G. Park, J. T. Rayz, and L. Pouchard, “Figure descriptive text extraction using ontological representation,” in *The Thirty-Third International Flairs Conference*, 2020.

[223] S. Agarwal and H. Yu, “Figsum: automatically generating structured text summaries for figures in biomedical literature,” in *AMIA Annual Symposium Proceedings*, vol. 2009, p. 6, American Medical Informatics Association, 2009.

[224] V. Venugopal, S. R. Broderick, and K. Rajan, “A picture is worth a thousand words: applying natural language processing tools for creating a quantum materials database map,” *MRS Communications*, vol. 9, no. 4, pp. 1134–1141, 2019.

[225] K. Reitz, I. Cordasco, and N. Prewitt, "Requests: Http for humans," *KennethReitz [Internet]. https://2.python-requests.org/en/master*, 2014.

[226] W. Jiang, E. Schwenker, T. Spreadbury, N. Ferrier, M. K. Chan, and O. Cossairt, "A two-stage framework for compound figure separation," 2021.

[227] K. He, X. Zhang, S. Ren, and J. Sun, "Resnet," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016.

[228] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2015.

[229] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, B. Wang, and Y. Chen, "Convolutional recurrent neural networks: Learning spatial dependencies for image representation," 2015.

[230] B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, and I. Foster, "The materials data facility: Data services to advance materials science research," *JOM*, 2016.

[231] B. Blaiszik, L. Ward, M. Schwarting, J. Gaff, R. Chard, D. Pike, K. Chard, and I. Foster, "A data ecosystem to support machine learning in materials science," *arXiv preprint arXiv:1904.10423*, 2019.

[232] E. Schwenker, W. Jiang, T. Spreadbury, N. Ferrier, O. Cossairt, and M. K. Y. Chan, "Exsclaim! validation dataset - selections from amazon mechanical turk benchmark," 2021.

[233] E. Schwenker, W. Jiang, T. Spreadbury, N. Ferrier, O. Cossairt, and M. K. Y. Chan, "Exsclaim! exploratory dataset - nanostructure images from nature journals," 2021.

[234] R. Rehurek and P. Sojka, "Gensim–python framework for vector space modelling," *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011.

Publications

Accepted or Submitted

(1) **E. Schwenker**, V.S. Chaitanya Kolluru, J. Guo, X. Hu, Q. Li, M.C. Hersam, V.P. Dravid, R.F. Kile, J.R. Guest, and M.K.Y. Chan, “Ingrained – An automated framework for fusing atomic-scale image simulations into experiments”, *submitted:* Small, 2021.

(2) **E. Schwenker**, W. Jiang, T. Spreadbury, N. Ferrier, O. Cossairt, and M.K.Y. Chan, “EXSCLAIM! – Harnessing materials science literature for self-labeled microscopy datasets”, *submitted:* Nanoscale Horiz., 2021.

(3) W. Jiang, **E. Schwenker**, T. Spreadbury, N. Ferrier, M.K.Y. Chan, and O. Cossairt, “A two-stage framework for compound figure separation”, *accepted:* Proc. 2021 IEEE Int. Conf. Image Process., 2021.

(4) Q. Li, V.S. Chaitanya Kolluru, M.S. Rahn, **E. Schwenker**, S. Li, R.G. Hennig, P. Darancet, M.K.Y. Chan, and M.C. Hersam, “Synthesis of borophane polymorphs through hydrogenation of borophene”, Science, vol. 371, pp. 1143-1148, 2021.

(5) J. Guo, A. Mannodi-Kanakkithodi, F.G. Sen, **E. Schwenker**, E.S. Barnard, A. Munshi, W. Sampath, M.K.Y. Chan, and R.F. Klie, “Effect of selenium and chlorine co-passivation in polycrystalline CdSeTe devices”, Appl. Phys. Lett., vol. 115, pp. 153901, 2019.

(6) **E. Schwenker**, F.G. Sen, C. Ophus, T. Paulauskas, J. Guo, S. Hills, R.F. Klie, and M.K.Y. Chan, “An autonomous microscopy workflow for structure determination from atomic-resolution microscopy images”, Microsc. Microanal., vol 24, pp. 510-511, 2018.

(7) **E. Schwenker**, F.G. Sen, C. Ophus, T. Paulauskas, J. Guo, S. Hills, R.F. Klie, and M.K.Y. Chan, “Leveraging first principles modeling and machine learning for microscopy data inversion”, Microsc. Microanal., vol 23, pp. 178-179, 2017.

In Preparation

(1) **E. Schwenker**, A. Mannodi-Kanakkithodi, F.G. Sen, S. Hills, J. Guo, R.F. Klie, V.P. Dravid, R.F. Klie, and M.K.Y. Chan, "Combining atomistic modeling and microscopy image simulation for automated inversion of experimental CdTe grain boundary images", *in prep*, 2021.

(2) **E. Schwenker**, C. Ophus, F.G. Sen, T. Paulauskas, J.G. Wen, C. Wolverton, and M.K.Y. Chan, "Understanding image similarity for automated comparisons in atomic-resolution electron microscopy", *in prep*, 2021.

(3) V.S. Chaitanya Kolluru, S. Hills, **E. Schwenker**, I. Malsky, F.G. Sen, A. Kinaci and M.K.Y. Chan, " FANTASTX: An Automated Experimental-Computation Approach to Determining Nanoscale Structures", *in prep*, 2021.

On arXiv

(1) **E. Schwenker**, V.S. Chaitanya Kolluru, J. Guo, X. Hu, Q. Li, M.C. Hersam, V.P. Dravid, R.F. Kile, J.R. Guest, and M.K.Y. Chan, ""Ingrained – An automated framework for fusing atomic-scale image simulations into experiments", arXiv preprint arXiv:2105.10532, 2021.

(2) **E. Schwenker**, W. Jiang, T. Spreadbury, N. Ferrier, O. Cossairt, and M.K.Y. Chan, "EXSCLAIM! – Harnessing materials science literature for self-labeled microscopy datasets", arXiv preprint arXiv:2103.10631, 2021.

(3) W. Jiang, **E. Schwenker**, T. Spreadbury, N. Ferrier, M.K.Y. Chan, and O. Cossairt, "A Two-stage Framework for Compound Figure Separation", arXiv preprint arXiv:2101.09903, 2021.