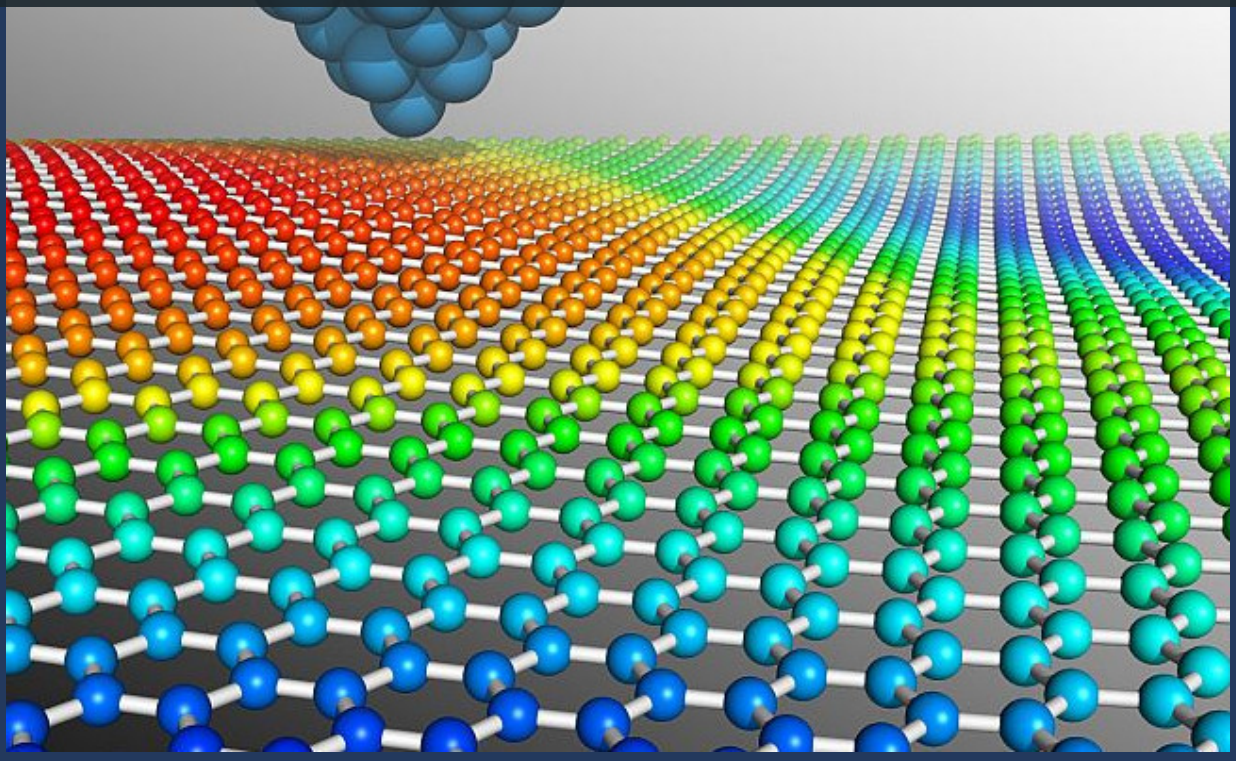
Machine Learning Accelerated Characterization of Single Adatoms Via STM Imaging

Master's End Project

Audrey Sabri



Cover photo from [1]

TU Delft

Machine Learning Accelerated Characterization of Single Adatoms Via STM Imaging

Master's End Project

Thesis Report

by

Audrey Sabri

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended publicly on May 28, 2025 at 15:00

Thesis committee:	Dr. Alexandre Artaud
	Dr. Kevin Rossi
	Dr. Sonia Conesa-Boj
Place:	Faculty of Applied Sciences, Delft
Project Duration:	October, 2024 - May, 2025
Student number:	5834538

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

Faculty of Applied Sciences · Delft University of Technology

Abstract

We demonstrate the ability of computer vision methods and unsupervised machine learning models to analyze atom-scale images acquired by scanning tunneling microscopy (STM). The imaged system of interest consists of Fe and Ti atoms adsorbed on bilayer MgO islands grown on an Ag substrate, a system studied for its quantum magnetic properties. The purpose of this analysis is to detect and classify each atom site, which includes Ti atoms, Fe atoms, defects on the surface of the sample, and unknown contaminants. Our method identifies Fe and Ti atoms from the rest in an effort to aid laboratory researchers during single-atom probing tasks. Each acquired STM image is processed with computer vision libraries to locate each atom site and to build a dataset of atom crops that highlight their features. Atom site classification and characterization tasks are performed using Gaussian mixture models (GMM), principal component analyses (PCA), and density-based clustering to classify each atom site by species and quantify structural differences and similarities between them [1].

[1] Data and code openly available at https://github.com/AudreySabri/unsupervised-ml-for-microscopy.git

Contents

Nomenclature

List of Abbreviations

Ag	Silver
Ar	Argon
Fe	Iron
GMM	Gaussian Mixture Model
IETS	Inelastic Electron Tunneling Spectroscopy
LDOS	Local Density of States
MgO	Magnesium Oxide
O	Oxygen
PCA	Principal Comopnent Analysis
RF	Radio-Frequency
STEM	Scanning Transmission Electron Microscopy
STM	Scanning Tunneling Microscopy
SVD	Singular Value Decomposition
Ti	Titanium
TM	Transition Metal
VAE	Variational Autoencoder

List of Figures

List of Tables

1 Introduction

1.1. Motivation

Electron spin manipulation can be used for information storage by encoding it in the spin state of a charge carrier, creating spintronic devices. The advantages of spintronic devices over electronic devices include greater storage capacity, lower energy consumption, and non-volatility [2]. As electronic devices become smaller and smaller, surpassing the nanoscale range, opportunities arise to combine the principles of molecule-scale electronics and spintronics. For instance, the study of single-molecule and single-atom spin manipulation has gained increasing attention by the scientific community. In fact, single spin manipulation has emerged as a solution for high-density information storage [3].

More specifically, the interest is in building devices that exploit the capture and control of the coherent quantum dynamics of a spin system. This can be done in scanning tunneling microscopy (STM) by modulating the magnetic interaction between the STM tip and a surface atom by applying a time-varying electric field in the STM junction through a radio frequency (RF) source [4]. Transition metal (TM) atoms adsorbed on ultrathin insulating layers grown on metal surfaces have emerged as prime candidates for manipulating single-atom spin dynamics thanks to their long spin coherence times, among other interesting properties [5].

In this work, we are interested in Ti atoms adsorbed on MgO bilayer surface grown on Ag. The MgO surface also adsorbed deposited Fe atoms that can be picked up by the STM tip to polarize it [6]. Naturally, these adatoms tend to prefer certain adsorption sites and have different surface interactions with the adsorbent, which are highly influential in determining the stability of the quantum spin state to manipulate [7]. However, distinguishing different atomic species from each other is a difficult task for the naked eye and requires probing each individual atom site. Thus, it is essential to develop comprehensive methods to instantaneously distinguish between different species of adatoms in STM images at different binding sites.

1.2. Objectives

Machine learning models for object detection and image classification have emerged for this purpose, mostly utilizing supervised learning models to classify single adatom sites [8], [9]. This process requires a lengthy labeling process and conformity in the STM acquisition parameters, as well as a large amount of data to adequately train the model without any overfitting.

The objective of this work is to develop a completely unsupervised framework for atom detection, image processing, and atom site classification. By using unsupervised methods, the goal is to avoid the time-consuming labeling process for model training and to build a model that does not require to be pre-trained. This results in an automated process that can be successfully deployed for a long range of acquisition parameters for each STM image of our system, and without any data augmentation requirements. Unsupervised methods can also extract hidden insights from the adatom sites by discovering the main variations between the discovered classes, leading to physical insights about the imaged system. The pipeline for our framework is presented in Figure 1.1, and will be discussed in detail in Section 3.

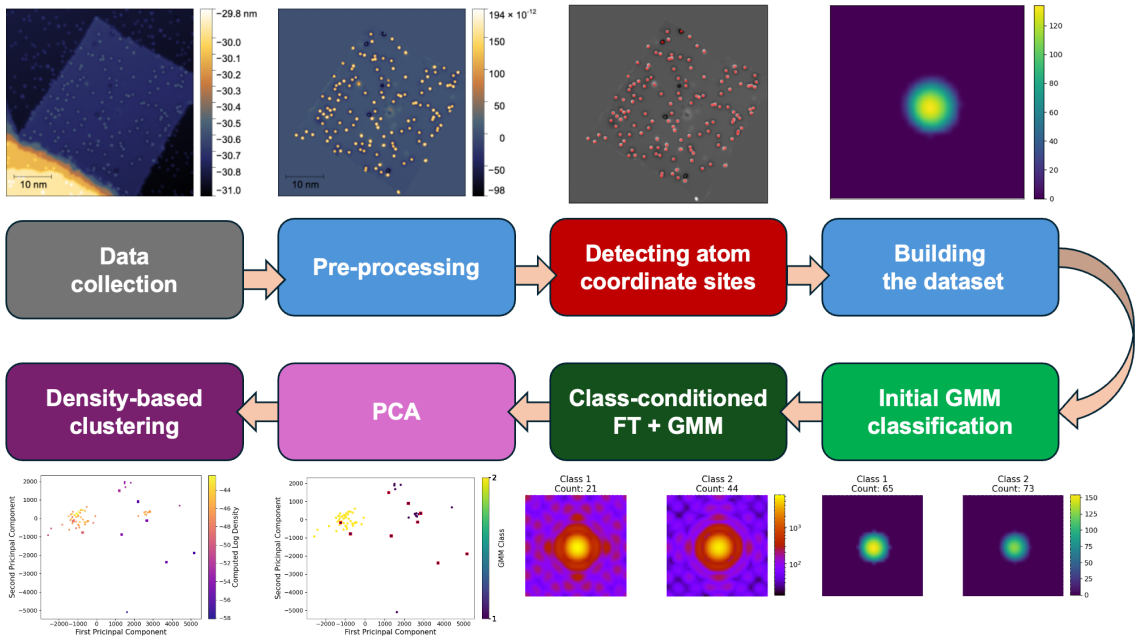


Figure 1.1: Detailed pipeline describing the framework presented in this work. We start by collecting atom-scale images of our system through Scanning Tunneling Microscopy (STM). The STM data is then pre-processed to highlight the region of interest in our images. Next, we use computer vision methods to detect all atom sites in our images and extract their coordinates. Once we have our atom sites and their coordinates, we use image processing techniques to build and prepare a dataset of single atom crops that are ready for training for each STM image. The data is now taken through the machine learning pipeline, which starts with a two-class Gaussian Mixture Model (GMM) classification that classifies data points according to their atomic species (Fe or Ti). This initial classification is refined to separate false classifications from true positive atom classifications through class-conditioned GMM, this time on the Fourier transform moduli of the Fe- and Ti-classified data. Next, the topographical differences and similarities in the Fourier transformed Fe and Ti datasets are quantified by using a Principal Component Analysis (PCA) that breaks down the data's essential features into separate components. We also use a density-based clustering method to quantify these topographical differences and similarities based on a global representation of the data. The PCA and density-based clustering methods are meant to explain and even rectify the class-conditioned GMM classifications, all while extracting topographical insight from our data.

1.3. Report Structure

In the next section, we will present a literature review to grasp the physical principles of STM operation, the outcomes of imaging TM atoms adsorbed on insulating surfaces, and the current state-of-the-art in unsupervised learning methods for microscopic imaging data.

Next, we will discuss the experimental methods of this work, as seen in Figure 1.1. We start by describing the imaged system and the STM data collection procedure, followed by an STM image preprocessing guide. We then describe the atom detection and dataset building workflows, rooted in image processing techniques such as shape detection, intensity clipping, and background removal by median filtering. Once we have described our data, we discuss the mathematical principles behind the unsupervised learning methods used in this work.

Finally, we explain the results of our models and discuss our findings. This includes a first round of Gaussian mixture model (GMM) classifications, followed by a second round of GMM subclassifications on the Fourier transformed atom site images for each detected class. The subclassification is intended to separate correctly classified Fe and Ti atoms from defects and unknown atoms in the STM image. We also run principal component analysis (PCA) and density-based clustering models on the Fourier transformed

class-conditioned data to further quantify the distinctions between the Fe and Ti atoms from these defects and unknowns in the STM image.

Part I

Theory and Methodology

2

Literature Review

2.1. Scanning Tunneling Microscopy

STM produces atomically precise (nanometer to sub-nanometer scale) material images without using optical principles. Instead, STM uses a piezoelectric mechanism to scan the sample with a voltage-biased atom-sharp conducting tip at a height of only a few angstroms while measuring the current changes induced by quantum tunneling in relation to a reference value. Based on this difference, a feedback loop is used to drive the piezoelectric mechanism in the x, y, and z directions, moving the tip along the surface. If the tunneling current is higher than the reference value, then the tip withdraws in the z direction, and vice versa. The surface topography is recorded by measuring the tip height as it scans the sample. Higher z values appear brighter, while lower z values appear darker.

The main concept of interest for STM is quantum tunneling. To explain the quantum mechanics behind STM imaging, we use the simple model of an electron with energy E and mass m moving in a potential $U(z)$ [10]

$$\frac{p_z^2}{2m} + U(z) = E \tag{2.1}$$

In classical mechanics, this electron is confined to the regions where $E > U(z)$. The motion of the electron is limited by a potential barrier set by $U(z)$, beyond which its momentum $p_z = 0$. However, in quantum mechanics, the electron state is modeled by Schrodinger's equation, which has solutions in both the classically allowed and forbidden regions of this model.

$$-\frac{\overline{h}^2}{2m}\frac{d^2}{dz^2}\Psi(z) + U(z)\Psi(z) = E\Psi(z) \tag{2.2}$$

In the classically allowed region $E > U$, the solutions are

$$\Psi(z) = \Psi(0)\,\mathsf{exp}\{\pm ikz\} \tag{2.3}$$

Here, the wave vector k is

$$k = \frac{\sqrt{2m(E-U)}}{\overline{h}} \tag{2.4}$$

In the classically forbidden region $E < U$, the solutions are

$$\Psi(z) = \Psi(0)\,\mathsf{exp}\{-\kappa z\} \tag{2.5}$$

Here, the electron's trajectory decays in the positive z direction with decay constant κ

$$\kappa = \frac{\sqrt{2m(U-E)}}{\overline{h}} \tag{2.6}$$

And the probability density of observing an electron near a point z is proportional to the squared absolute value of equation 2.3, which is nonzero in the non-classical region.

Applying this model to STM, we are interested in exciting electrons from the metal surface to vacuum, which requires a minimum energy of ϕ, a metal's work function. To model the metal-tip interaction with the metal sample, we take the vacuum level as zero point energy, making the occupied states with the highest energy, the Fermi level, have an energy $E_F = -\phi$. We assume that the tip and sample have equal work functions. Applying a bias voltage V induces a net tunneling current by exciting an electron with an energy level between $E_F - eV$ and E_F to tunnel into the tip. We assume that $\phi >> V$, meaning that all the candidate electrons for tunneling lie very close to the Fermi level, and we can approximate their energies $E_n = -\phi$. From equation 2.3, we may compute the probability that an electron in the n^{th} state tunnels through the vacuum and finds itself on the surface of the tip $z = W$

$$w \propto |\psi_n(0)|^2 \exp\{-2\kappa W\} \tag{2.7}$$

Here, $\psi_n(0)$ is the n^{th} state at the sample surface, and the decay constant in the forbidden region of a state n with energy $E_n = -\phi$ is

$$\kappa = \frac{\sqrt{2m\phi}}{\overline{h}} \tag{2.8}$$

Thus, a tunneling current is generated by a voltage bias V that excites electrons to move from the metal sample to the metal tip surface, proportional to the number of states on the sample's surface within the energy interval eV

$$I \propto \sum_{E_n=E_F-eV}^{E_F} |\psi_n(0)|^2 \exp\{-2\kappa W\} \tag{2.9}$$

This number of states is finite for metals but very small or zero for insulators. Considering that V is small enough for the electron state density to not vary significantly within the energy interval, we may write the above sum in equation 2.9 in terms of the local density of states (LDOS) at the Fermi level. The LDOS $\rho_S(z, E)$ is the number of electrons per unit volume per unit energy at a given point z and a given energy E, expressed as

$$\rho_S(z, E) = \frac{1}{\epsilon} \sum_{E_n=E-\epsilon}^{E} |\psi_n(z)|^2 \tag{2.10}$$

So, equation 2.9 can be written as

$$I \propto V \rho_S(0, E_F) \exp\{-2\kappa W\} \tag{2.11}$$

Plugging in our value for κ in equation 2.8,

$$I \propto V \rho_S(0, E_F) \exp\Big\{-1.025\sqrt{\phi} W\Big\} \tag{2.12}$$

We see that the work function can be measured by varying the tip-sample distance by writing ϕ, which also doubles as the barrier height, as the change in the logarithm of the current with respect to the tip's distance from the sample

$$\phi = \frac{\overline{h}^2}{8m}(\frac{d\log I}{dW})^2 \approx 0.95(\frac{d\log I}{dW})^2 \tag{2.13}$$

Furthermore, we may express the tunneling current according to the density of states at the tip with

$$\sum_{E_F-eV}^{E_F} |\Psi(0)|^2 \exp\{-2\kappa W\} \equiv \rho_S(W, E_F)eV \tag{2.14}$$

Which leads to (from equation 2.11)

$$I \propto \rho_S(W, E_F)V \tag{2.15}$$

Thus, we keep the tunneling current constant while scanning the sample as we map its topography by recording changes in W. In STM, we are generating a map of the LDOS, where height differences reflect how close the voltage-biased tip needs to be to the sample to excite an electron. The more the tip retracts, the higher the LDOS.

2.2. Imaging Transition Metal Atoms Adsorbed on Magnesium Oxide

We are studying a system consisting of TM atoms adsorbed on a double insulator layer grown on a metal surface. These systems stand out by their magnetic properties, such as long spin coherence and even longer spin-relaxation times, among other chemical properties that depend on the adsorption sites of single TM atoms [5]. Different TM atoms on different adsorption sites can be distinguished by the height of the STM tip in the mapped topography and by their fingerprint spin excitation spectra, measured by inelastic electron tunneling spectroscopy (IETS) [11].

The differences in tip height for different TM atoms at different adsorption sites are explained by their electronic structures that dictate the charge transfer that takes place between the adsorbent and the adsorbate. This decides the eventual orbital configuration of the adsorbed TM atom and its most favorable adsorption site. This results in different LDOS curves, characterized by the energy difference between the occupied levels of the atom and the Fermi energy E_F, as well as the hybridization between its orbital energy levels [12]. Thus, the bonding mechanism between TM atoms and the insulator layer decides the LDOS, measured by adjusting the STM tip height over the sample.

On the other hand, IETS captures the energy it takes to flip an electron's spin by measuring the energy lost by electrons to spin-flip excitations as they tunnel through the potential barrier towards the tip. This energy transfer hinges on a threshold voltage and manifests itself as a symmetric step or a gap in the conductance spectrum at zero bias, with a depth that is proportional to the applied magnetic field and a step up characterized by the threshold voltage (Figure 2.1). This threshold voltage corresponds to the Zeeman energy $\Delta = g\mu_B B$, where μ_B is the Bohr magneton, g is the Landé g-factor, and B is the applied magnetic field [6], [13].

2.3. Unsupervised Learning Methods for Microscopic Imaging Data

Unsupervised learning methods learn the features and representations of a dataset without any prior labeling. These include dimensionality reduction models that reduce the data to its most essential features. In doing so, we may extract the main sources of variation between data points in an automated way. This can reflect physical phenomena in microscopically imaged atomic systems [14]. We can find these models in deep learning frameworks that were specifically designed for electron and scanning probe microscopy as seen in the AtomAI framework [15]. This includes variational autoencoder (VAE) based models. VAE models are designed to compress unlabeled data to a latent space encoding its most descriptive features in order to reconstruct it through only these latent features. In doing so, non-linear relationships in a dataset are learned.

Applying VAE models to microscopy images first requires a preceding step to detect and extract the relevant areas of interest in the microscopic image. Such methods include semantic segmentation algorithms or supervised classification models for atom detection. The VAE will then reduce the constructed

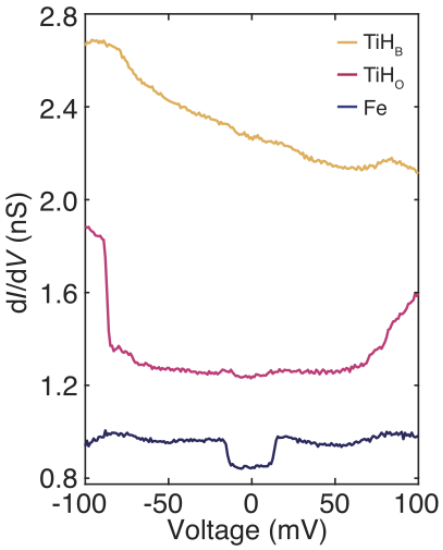


Figure 2.1: Conductance spectra of Ti and Fe single adatoms on double layer MgO, Fe adatoms are known to adsorb on O sites, while Ti adatoms may be adsorbed on either Bridge (B) sites or O sites, producing different excitation spectra. These spectra represent the atom specie's fingerprint excitation spectrum (figure from [11]).

dataset of areas of interest to its most essential features, represented as continuous variables that can be projected back onto the imaging data. Representing data points in a latent space can quantify their similarities and differences based on their most descriptive features.

This workflow has proved fruitful in extracting the main descriptors of topological structures in scanning transmission electron microscopy (STEM) images of graphene [14], classifying plasmonic particles by particle counts [16], or even identifying ferroelectric domain walls [17]. VAE-based models have also been used to detect atom species by reducing a dataset of STEM-imaged atom sites to its most essential features [8]. This leads to a representation of the dataset in its latent space, where we may use classification models to group data points by their resemblance.

VAE models can thus distinguish structurally different data points by extracting their main differences along a manifold of latent variables but require a large amount of data to effectively break it down to its essential features. Certain variations of VAE can directly undertake the classification task by modifying the model's loss function to learn a dataset's latent variables as discrete classes [16]. However, this method requires a highly expensive Bayesian optimization process. For this reason, we steer clear of deep learning models in this work despite their reported success, and stick to unsupervised machine learning methods, as we will see in the next section.

3

Experimental Methods

3.1. Data Collection

The system consists of Ti and Fe atoms adsorbed on bilayer MgO layer grown on a single-crystal Ag substrate. First, the Ag surface was thoroughly cleaned by repeated cycles of Ar-ion bombardment and annealing in ultra-high vacuum [18]. Ion bombardment erodes the surface to remove contaminants at the cost of producing a rough surface. This effect is rectified by smoothing the surface through annealing, which allows the surface Ag atoms to rearrange and lower its total energy. However, new contaminants can appear during annealing as a result of residual gases in the ultra-high vacuum chamber or because atoms separate from the bulk of the crystal as it cools. So, we repeat this process until we get clean, flat Ag terraces extending over several hundreds of nanometers. Once the substrate is prepared, MgO is grown on top in ultra-high vacuum by reactive evaporation [19]. Mg powder is evaporated with an e-beam evaporator towards the Ag surface in a pure, low-pressure O_2 atmosphere, while the sample is heated to a constant temperature. Once grown, the MgO/Ag sample is introduced into the STM for inspection. If the MgO islands are large and regular, then single Ti and Fe atoms are evaporated once again with an e-beam evaporator and deposited within the STM at a surface temperature of 10-20 K [20]. We want single-adatom sites on our surface, so we require a sufficiently low temperature to avoid surface diffusion, which would lead to cluster and island formation.

The STM images considered in this work were acquired with a Unisoku USM1300 STM with a base temperature of 300 mK in constant current mode. The temperature must be low enough for the atomic system to freeze in place. We are using an atomically sharp silver-covered tungsten tip to probe the Ti and Fe atoms adsorbed on the MgO surface. The bias voltage is set to 60 mV, the tunneling current is set between 10 and 23 pA, and the scan speed is between 8 and 40 nm/s, depending on the scan range (40 to 50 nm), the scan pixels, and the acquisition time (Table 3.1). The tip is kept slow enough to stay close to the surface in case of any obstacles. The adatoms are distinguished by their fingerprint spin excitation spectra captured by IETS and annotated by domain experts, as seen in Figure 3.1. Note that some of the STM images are not fully annotated, which presents a limitation to this work. STM measurements must be very precise, and accidental damages to the sample make measurements to the surrounding atoms unreliable. These damages are marked in black and are caused by debris dropped by the STM tip.

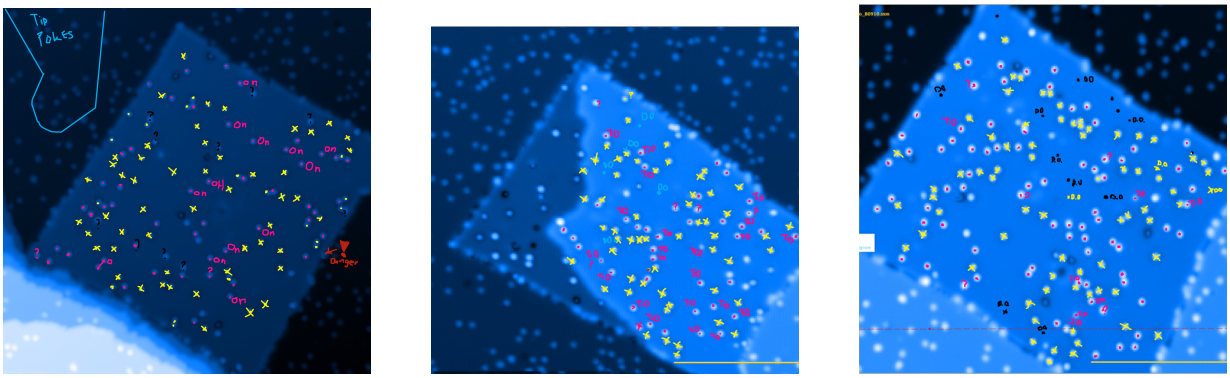
For this work, eight annotated images were provided. In our main text, we will focus on three selected images chosen to reflect the generalization of our analysis to a diversity of acquisition parameters and to the number of adatoms on the surface. In addition, we chose images with a limited number of damages to the sample to make sure that our data did not have missing annotations. The remaining images and their analysis can be found in the Appendix. The trends and findings discussed in the main text apply to these images as well.

3.2. Data Pre-processing

The first step of this project is to prepare the STM images for analysis. Once prepared, we must accurately detect each adatom site and extract their coordinates from the atomic-scale surface image by using a generalizable method across different acquisition parameters. Next, we must build a dataset of atom sites that describes each atom site by its intensity profile and, therefore, its topography.

Topography	Tunneling Current (A)	Scan Speed (m/s)	Scan Range(m)	Scan Pixels
B0917	2.27E-11	3.06E-08	4.00E-08	304
B2731	2.19E-11	1.63E-08	4.00E-08	256
B0627	2.28E-11	2.44E-08	4.50E-08	256
A001	2.00E-11	3.91E-08	5.00E-08	512
B0376	2.32E-11	3.99E-08	5.00E-08	256
A228	1.01E-11	9.96E-09	5.00E-08	784
B1544	2.24E-11	8.57E-09	4.60E-08	208
B2060	2.24E-11	9.92E-09	5.50E-08	208

Table 3.1: STM image acquisition parameters.



(a) (b) (c)

Figure 3.1: Expert annotated ground truth labels. (a) Topo A001. (b) Topo B0376. (c) Topo B0627. Titanium sites are marked in pink, iron sites are marked in yellow. Damages to the sample are marked in black, which result in unreliable measurements to the surrounding atoms, leading to unlabeled atom sites. Some images have further annotations meant for the lab personnel. The ground truth annotations were digitized to score the GMM classifications.

3.2.1. Processing and Extracting the MgO Island

First, the collected STM images are processed through Gwyddion [21], an open source software to visualize and analyze the scanning probe microscopy data.

The focus of our work is to detect TM adatom sites on the MgO island given an STM image as reported in Figure 3.2. Therefore, we want to focus on the bilayer MgO island and discard the rest of the areas in the image, composed of a single layer MgO island or the silver substrate itself. Considering that the STM operates in constant current mode, we notice that there is increased contrast in the area of interest due to a slight tilt in the sample (Figure 3.3a). We use Gwyddion's plane fitting feature to manually correct the tilt angle (Figure 3.3b).

Then, we use the two-dimensional Fourier filtering feature to remove any noise. Among the noise sources in STM imaging are environmental vibrations or electromagnetic noise [22]. We identify noise in our image as high frequency lines in the two-dimensional Fourier space found far from the center (Figures 3.4, 3.3c). As the STM scans the sample vertically, noise signatures in Fourier space will appear as vertical lines, the length of which indicates the frequency spread of the noise source. We may correct this noise safely without worrying about creating artifacts in the resulting filtered image as the vertical lines are well separated from the low spatial frequencies where the image signal is mostly located.

Once the noise has been removed (Figure 3.3d), we set the height scale to represent the bilayer MgO island as zero-point reference by shifting all values in the z-direction to set the bottom of the island as zero. The final step is to use the mask editor feature to select the area of interest and crop it, making sure to avoid step edges (Figures 3.3e, 3.3f).

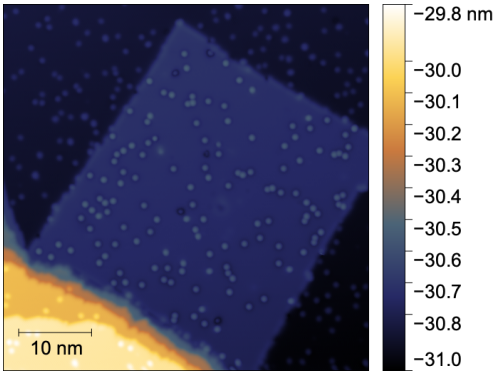


Figure 3.2: STM Image of Ti and Fe atoms adsorbed on a bilayer MgO island grown on Ag substrate. The color scale reflects the STM tip position relative to its z-piezo range.

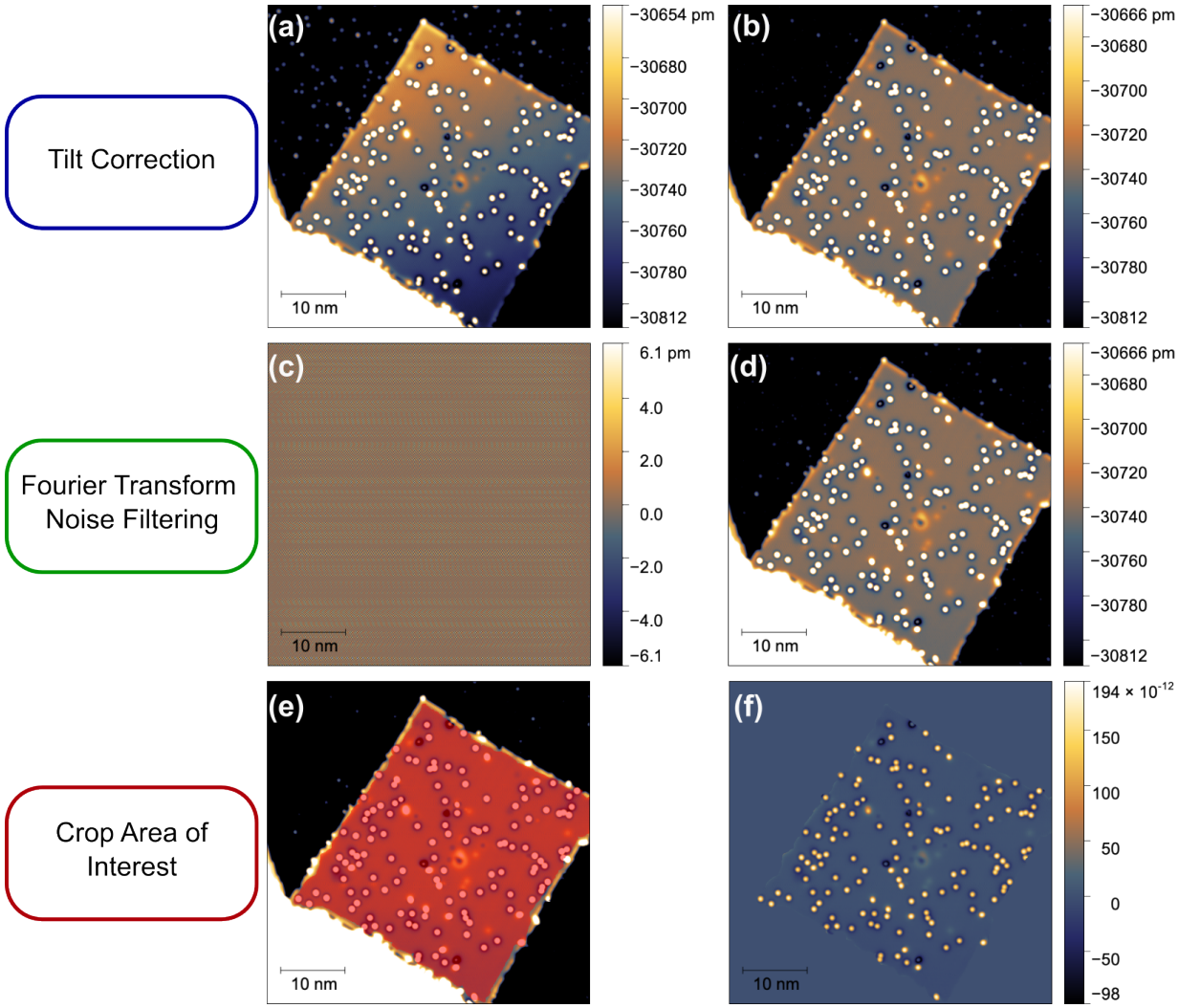


Figure 3.3: Gwyddion data pre-processing workflow. (a) Before tilt correction. (b) After tilt correction. Notice that the scale bar is now in picometers as we have narrowed the height spectrum to capture the height contrast on the MgO island. (c) Noise captured by a two-dimensional Fourier transform. (d) Filtered STM image. (e) Mask around the MgO island. (f) Cropped MgO island. Notice that there are no units on the legend, as we have shifted the height scale to set the bottom of the bilayer MgO island as zero.

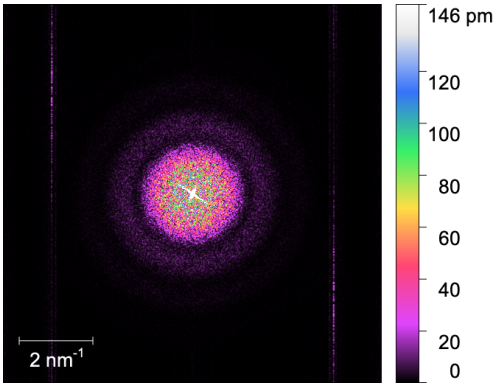
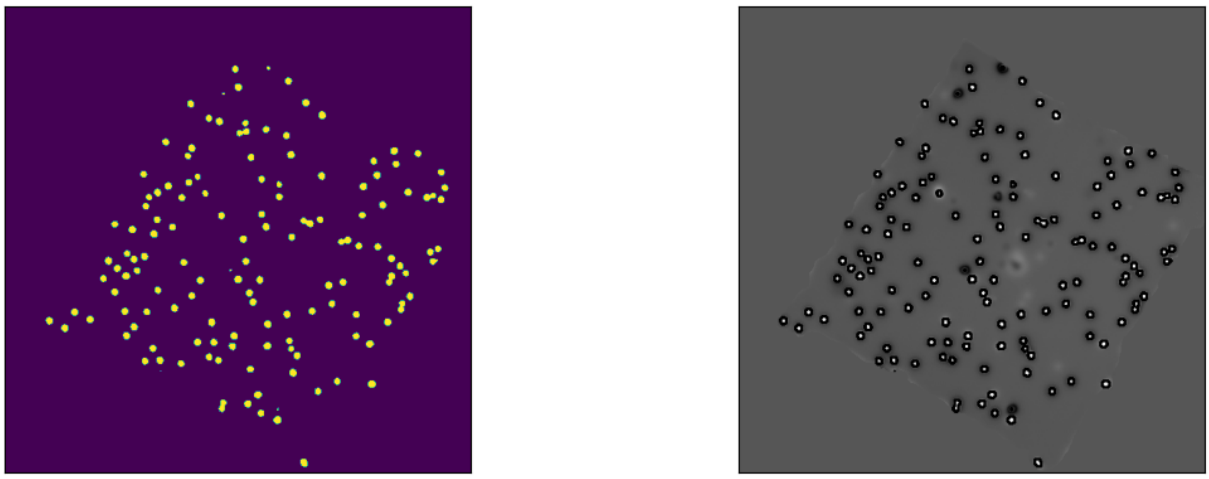


Figure 3.4: Using a two-dimensional Fourier transform to filter imaging noise. The noise is represented in the Fourier transform of the STM image as sharp vertical lines away from the center.



(a) **(b)**

Figure 3.5: Detecting and drawing contours on the STM image. (a) Binary thresholding using a Gaussian weighted sum of values in neighborhoods with a defined area minus a defined constant. (b) Contours detected by OpenCV.

3.2.2. Detecting Atom Coordinate Sites

Second, we want to detect all the atom sites and extract their coordinates. We do so using OpenCV's image processing library [23]. The first step is to apply binary thresholding to segment the shapes on top of the MgO island away from the background island itself (Figure 3.5a). These shapes may be adatoms, clusters of adatoms, defects, or unknown contaminants on our sample. Next, we use OpenCV's `findContours` function to find and draw contours around each shape (Figure 3.5b). We notice a challenge in contour definitions since some contours are drawn around the edges of the MgO island. Furthermore, dimers (polymers composed of two atoms), trimers (polymers composed of three atoms), and other close successions of atoms may be drawn as a single contour, where each atom site must be separated.

We filter out edges by applying a bounding rectangle to each contour and computing the ratio of contour area to bounding rectangle area, known as the extent. We set a threshold value to filter out any contour below a certain extent. This works because contours drawn on the edges will have a really narrow surface area, yet might span a large length (Figure 3.6a). The single atom, dimer, and trimer shapes have a higher extent ratio as their shapes completely fill the contour and may be bounded by a rectangle that closely matches their surface area (Figures 3.6b, 3.6c).

Once the edges have been filtered out, we want to separate dimers, trimers, or any contours that contain more than a single atom. We use OpenCV's `matchShapes` function to compare each contour to a

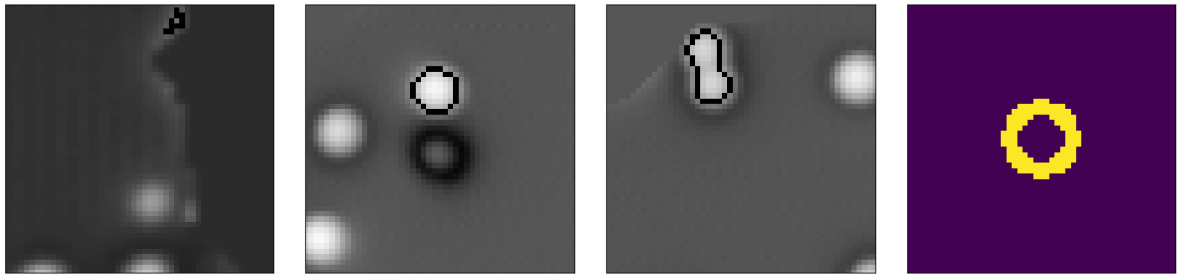


Figure 3.6: Different types of contours. (a) Contour around an island edge feature. (b) Contour around a single adatom. (c) Contour around a dimer of atoms. (d) Circle contour for comparison.

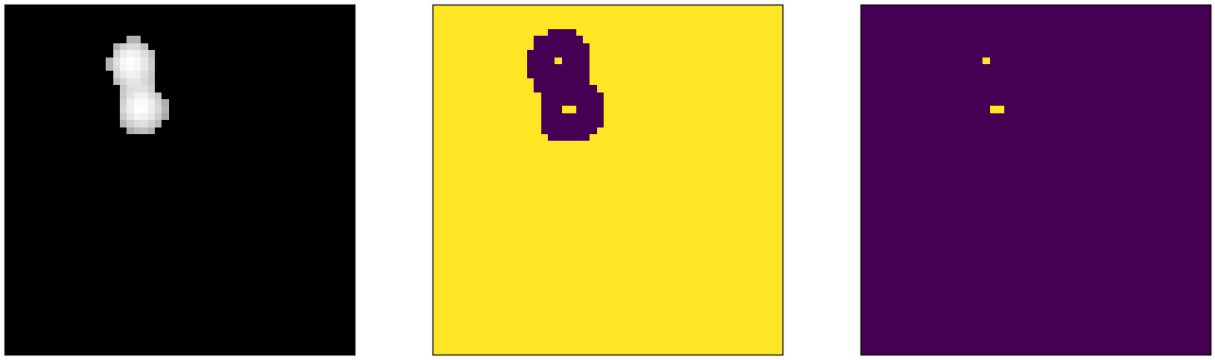
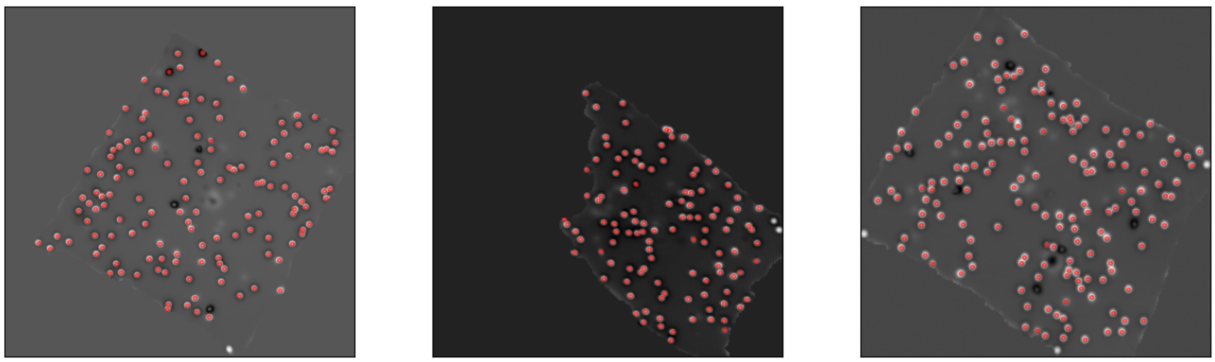


Figure 3.7: Peak detection workflow. (a) Cropped dimer contour. (b) Local maxima found in dimer contour. (c) Detected peaks in dimer contour. Notice that 2 side-by-side pixels are both detected as peaks, and one must be filtered out.

perfectly drawn circle of an appropriate radius, depending on the dimensions of each image. Since we set our images to be 512 x 512 pixels, a circle with a radius of 6 pixels is good enough to represent an atomic site (Figure 3.6d). The `matchShapes` function returns a score that shows the similarity between two shapes. The more similar they are, the closer the score is to 0. In this case, a contour that is round enough to be considered a single atom should output a score of less than 1. The center of mass for contours that are identified as single adatoms is taken as the coordinate of the atom.

In the case of dimers or trimers, the output score will be higher than 1. In this case, we search for the local intensity peaks in the contour. If more than one peak is found, we assess its validity with a threshold distance between peaks, to make sure that each peak does indeed represent a single atom. Peaks that are too close are not valid. The threshold distance should be different if more than two peaks were found. This effectively eliminates duplicate peaks for dimers, trimers, and single atoms that skipped the previous filtering step. Valid intensity peaks are taken as atomic coordinates (Figure 3.7).

The results can be seen in Figure 3.8. We notice a near-perfect atom site detection (Table 3.2). The only limitations of this method observed are in some cases where the intensity maxima (or peaks) found in dimers or trimers do not properly reflect the center of mass of an atomic site, or when atoms on the edge of the MgO island are excluded.



(a) (b) (c)

Figure 3.8: Results of our atom detection workflow on the three topographies presented in this work. The red dots correspond to detected atom coordinate sites. We see that some atoms on the edge of the bilayer MgO island are not detected. (a) Topo A001. (b) Topo B0376. (c) Topo B0627.

Topography	Number of Atoms	Number of detected sites
B0917	109	109
B2731	87	87
B0627	163	161
A001	138	138
B0376	114	112
A228	129	128
B1544	130	125
B2060	209	201

Table 3.2: Counts of true and detected atom sites.

3.2.3. Building the Dataset

The dataset for a single STM image is composed of 32 x 32 pixel window crops centered on each detected atom coordinate. These window crops are extracted with AtomAI's `get_imgstack` function [15], which extracts subimages centered at specified coordinates for a single image.

We notice that the intensity values in each atom are too spread out due to oversaturation caused by the maximum intensity values in our images (Figure 3.9a). We must thus clip (limit) the maximum intensity values to a certain degree while ensuring that no information is lost during preprocessing. So, for each subimage, we limit the maximum intensities by clipping them to the 99.9^{th} quantile of the intensity distribution. Thus, we prevent higher intensities from skewing the intensity profile, creating a more concentrated intensity distribution (Figure 3.9b).

The next step is to remove the background. Thanks to intensity clipping, the application of a median filter now accurately extracts the MgO island as the background because its range of values is more narrowly distributed and therefore more distinguishable from any other elements in the image. So, we subtract from each subimage the result of its median filter (Figures 3.10b, 3.10c).

Finally, we make sure to remove any neighboring atoms from the subimages by taking a circular crop of a set radius centered on each detected atom coordinate. We tried our best to optimize this radius by trial and error to optimize the single atom crop in dimers or trimers (Figure 3.10d).

3.3. Data Analysis

The goal of this project is to classify adatoms by species in atom-scale surface images. The main challenges are the limited number of data, the topographical similarities between different species, and the difference in

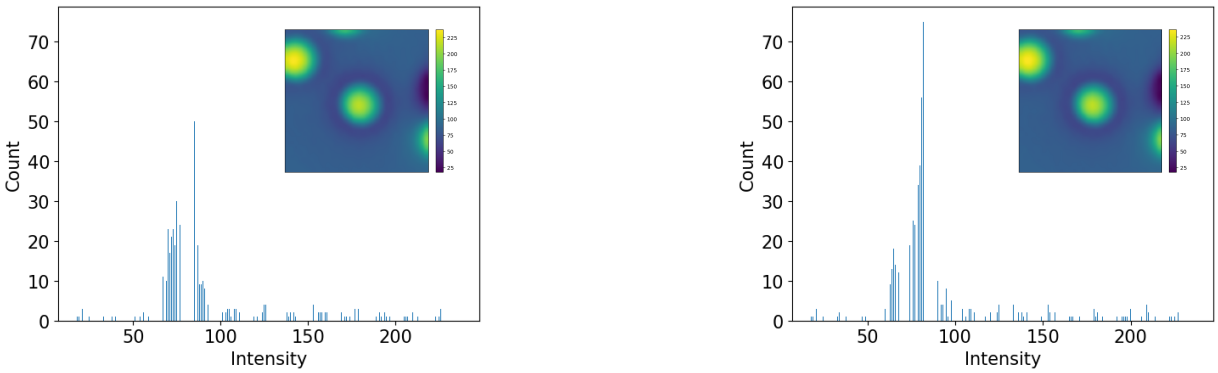


Figure 3.9: Effects of clipping on intensity distributions, the median intensity values become concentrated within the same range. (a) Intensity profile of the window crop without clipping. (b) Intensity profile of the window crop with clipping.

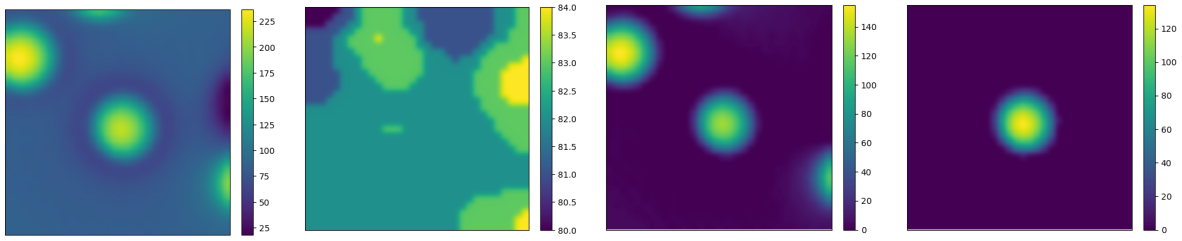


Figure 3.10: Dataset building workflow. (a) Intensity clipped window crop. (b) Background extracted by median filter, facilitated by intensity clipping. (c) Window crop with background removed. (d) Cropped atom image.

acquisition parameters between images. We exploit a number of unsupervised machine learning methods to overcome these challenges. This section covers their mathematical foundation. The result section describes how these were adapted and validated in the context of identifying adatom species.

3.3.1. Gaussian Mixture Model

A GMM is a probabilistic model that models the probability density function of classifying a set of data vectors in one of M components (or classes) as a mixture of a finite number of Gaussian distributions [24], [25],

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} w_i g(\vec{x}|\vec{\mu_i}, \vec{\Sigma}_i) \tag{3.1}$$

where $i = 1, ..., M$, $\vec{x}$ is a N-dimensional data vector, w_i are the mixture weights that sum up to 1, and $g(\vec{x}|\vec{\mu_i}, \vec{\Sigma}_i)$ are the component Gaussian distributions,

$$g(\vec{x}|\vec{\mu_i}, \vec{\Sigma}_i) = \frac{1}{(2\pi)^{M/2}|\vec{\Sigma}_i|} \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu_i})'\vec{\Sigma}_i^{-1}(\vec{x} - \vec{\mu_i})\right] \tag{3.2}$$

with mean vector $\vec{\mu_i}$ and diagonal covariance matrix $\vec{\Sigma}_i$. Thus, the GMM is parameterized by λ such that

$$\lambda = \{w_i, \vec{\mu_i}, \vec{\Sigma}_i\}, i = 1, ..., M \tag{3.3}$$

with the goal of optimizing these parameters to produce the most accurate classification based on the latent variables of the data. To do this, we use an expectation-maximization algorithm. Depending on the initialized M number of components, the same number of clusters is taken as priors based on a k-means clustering algorithm [26]. Based on this designation, the probability that a data point is located in any of the M clusters is computed. This likelihood is maximized through an iterative process in which the parameters λ are updated until the algorithm converges to an optimum. Mathematically, we want to maximize the probability of the GMM fit given T data points such that our data set is represented as $X = \{\vec{x_1}, ..., \vec{x_T}\}$. This probability can be written as

$$p(X|\lambda) = \prod_{t=1}^{T} p(\vec{x_t}|\lambda) \tag{3.4}$$

where we begin with an initial λ model that gets new parameters at every iteration until $p(X|\lambda)$ is maximized. The updated $\{w_i, \vec{\mu_i}, \vec{\Sigma}_i\}$ are computed as

$$w_i = \frac{1}{T}\sum_{t=1}^{T} Pr(i|\vec{x_t}, \lambda) \tag{3.5}$$

$$\vec{\mu_i} = \frac{\sum_{t=1}^{T} Pr(i|\vec{x_t}, \lambda)\vec{x_t}}{\sum_{t=1}^{T} Pr(i|\vec{x_t}, \lambda)} \tag{3.6}$$

$$\vec{\Sigma}_i = \frac{\sum_{t=1}^{T} Pr(i|\vec{x_t}, \lambda)\vec{x_t}^2}{\sum_{t=1}^{T} Pr(i|\vec{x_t}, \lambda)} - \vec{\mu_i}^2 \tag{3.7}$$

with the posterior probability for component i,

$$Pr(i|\vec{x_t}, \lambda) = \frac{w_i g(\vec{x_t}|\vec{\mu_i}, \vec{\Sigma}_i)}{\sum_{k=1}^{M} w_k g(\vec{x_t}|\vec{\mu_k}, \vec{\Sigma}_k)} \tag{3.8}$$

For each STM image, we fit the GMM to the generated dataset of pre-processed cropped atom images described in the last section (Figure 3.10d). The purpose is to find the topographical similarities and differences between atom sites depending on their species. Since we have two species of adatoms, we set the number of components M to represent the number of atomic species that we want to detect (two). We hypothesize that the latent variables discovered by the GMM during the optimization of the parameters λ will differentiate Fe sites from Ti sites based on their intensity distributions.

Considering that the model is initialized through randomized k-means clustering, a random seed must be fed to the model to ensure reproducibility. We repeat the GMM classification on 100 different random seeds and choose the most common results. This ensures that the random initialization of the λ parameters does not skew our results.

Once we have a categorical separation between the two main species of adatoms in our system, we want to further distinguish true positive Fe and true positive Ti sites from false positives. These false positives may be defects or contaminants in the sample that might have occurred during sample preparation, for example, sample damages from e-beam evaporation during TM deposition or contaminants grown during the MgO growth. We also want to rectify any false classifications by the model (Fe classified as Ti or vice versa). To this end, we apply a second round of GMM classifications to each of the two classes discovered by the GMM with the same random seed initialization process.

However, to avoid learning the same latent variables during the second round of classifications, we must classify our atom sites based on their topographical similarities and differences. Thus, we apply two-dimensional Fourier transformations on each data point to extract their topographical representation as a sum of periodic sinusoidal components. This allows us to represent, classify, and compare the topographies of each atom site by their spatial frequencies in Fourier space.

For each of the two subclassifications, we set the number of components M to two, one for true positives and one for false positives. We hypothesize that the GMM will fit true positives according to the similarities in their Fourier space representations, discovered during the λ optimization. The second component is reserved for outliers who do not share similar topographical features with the true positives or potentially with each other. We are leveraging the fact that we know that part of the atom sites belong to the species that we deposited, and they are expected to have similar topographies. The defects and contaminants have unknown compositions, so they are expected to stand out among topographically similar atom sites.

3.3.2. Two-dimensional Fourier Transform

A two-dimensional Fourier transformation transforms an image to its spatial frequencies

$$F(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,\mathrm{exp}[-j2\pi(ux+vy)]dxdy \tag{3.9}$$

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(u,v)\,\mathrm{exp}[j2\pi(ux+vy)]dudv \tag{3.10}$$

where x and y are the image coordinates, u and v are spatial frequencies. From Euler's equation,

$$\mathrm{exp}[j2\pi(ux+vy)] = \cos 2\pi(ux+vy) + j\sin 2\pi(ux+vy) \tag{3.11}$$

a Fourier transformation decomposes a given image function $f(x,y)$ into a weighted sum of two-dimensional orthogonal sinusoidal basis functions to describe its intensities, shapes, edges, rotations, and other features such as periodic components. We take the Fourier transform of a sine function to illustrate this, as it will have a simple Fourier transform (Figure 3.11). In Fourier space, the center contains the average value of the image and is the center of symmetry such that the first and third quadrants are symmetric, similarly for the second and fourth quadrants [27].

In the left panel of Figure 3.11, the horizontal periodic component of the sine function is represented as bright symmetric dots in Fourier space. These dots are spaced away from the center in the horizontal direction with a spacing that is based on the period of the sine function. In the right panel of Figure 3.11, the same function is rotated diagonally and produces diagonally oriented bright spots in Fourier space.

However, a cross-shaped component appears in the horizontal and vertical directions. This is because the Fourier transform considers the image in question to be part of an array of identical images that are periodically repeated horizontally and vertically to infinity, creating edges between neighboring images. The cross-shaped component is the Fourier transform of the edges generated by repeating the original image.

Since $F(u, v)$ is complex, we take its modulus to analyze the resulting Fourier transform. In the context of atom images such as in Figure 3.10d, our aim is to extract the topography of each atom by a mathematical description of their features. Since an atom site can be modeled by a Gaussian function, the resulting Fourier transforms are also Gaussian functions with the highest values in the center to describe the mean intensity and expanding outer rings of lower values to describe the Gaussian intensity distribution of the original image (Figure 3.12). The more intricate details of the image, such as shapes and atom orientation, are represented as lower values in the high frequencies away from the center. By deploying our unsupervised methods on the Fourier transformed atom images, we may classify them based on their topographical similarities and differences.

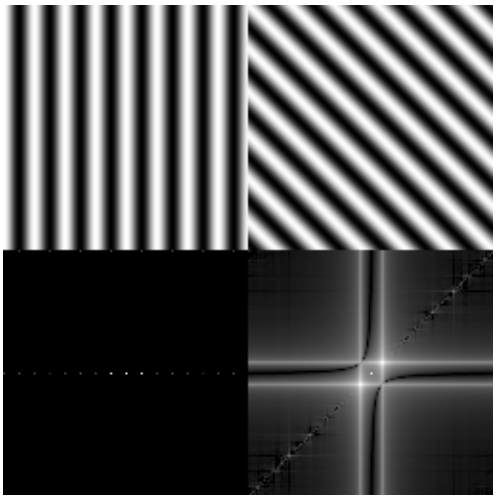


Figure 3.11: Sinusoidal functions (top) and their respective Fourier transforms (bottom). In the left panel, we have a horizontally periodic sine function of eight cycles. In the right panel, we have the same function rotated diagonally (figure from [27]).

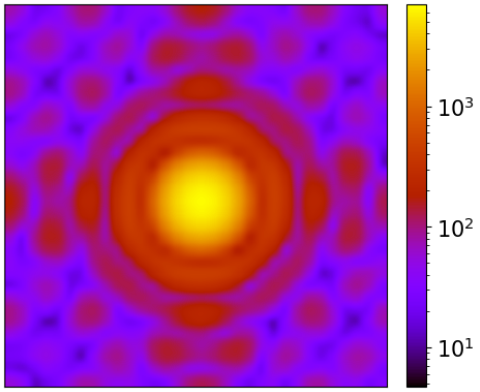


Figure 3.12: A Fourier transformed atom crop plotted in log-scale.

3.3.3. Principal Component Analysis

A PCA is a linear transformation that projects a high-dimensional dataset onto a lower-dimensional space through feature extraction. The feature extraction is done by singular value decomposition (SVD),

$$A = U\Sigma V^T \tag{3.12}$$

where A is a high-rank matrix of size m x n, U is an m x m orthogonal matrix with columns $\vec{u_1}...\vec{u_m}$, V is a n x n orthogonal matrix with columns $\vec{v_1}...\vec{v_n}$, and Σ is an m x n matrix with only positive diagonal entries in descending order, its entries being the singular values of A where

$$A\vec{v_i} = \sigma_i \vec{u_i} \tag{3.13}$$

with σ_i as the singular values and $\vec{v_i}$, $\vec{u_i}$ as the components of the new orthonormal bases $\vec{v}$, $\vec{u}$. Thus, A is reduced to a two-basis diagonal matrix [28].

This represents the linear decomposition of our dataset as a two-basis principal component space that captures its main uncorrelated variations. The vector components $\vec{v_i}$, $\vec{u_i}$ are the principal components in this case.

For our purposes, we choose a large enough number of principal components so that the variance explained by each component converges to 0. Considering that the singular values σ_i are in descending order, the first principal component best explains the variance in the data, followed by the second, and then the third, etc..

We apply the PCA on the two-dimensional Fourier transformations of our data points as a more quantitative classification after the class-conditioned GMM, as previously discussed. A PCA on the two-dimensional Fourier transformations of our data points quantifies the differences and similarities between the topographical features of our data points based on their extracted principal components. We plot our data points in the principal component space composed of the most significant components to see how the distribution of data points forms. The distances between data points along each of the principal component axes quantify their differences and similarities based on the relevant component. Thus, we look for clusters to signify data points that represent similar species and outliers to signify defects, contaminants, or misclassifications.

3.3.4. Density-based Clustering

PCA has its shortcomings. A dimensionality reduction method learns the variability in our data by separating it into components, rather than learning it from a global representation of the data. To do so, we choose a density-based clustering algorithm, where the cluster centers are determined by the density profile of the dataset [29]. We choose this method so that the number of clusters appears without initialization, while still identifying outliers, by choosing cluster centers as the points with the maximum local density at a maximum distance to the points with higher density [30].

The probability density is computed as

$$\rho_i = \frac{1}{N}\frac{k}{V_{i,k}} \tag{3.14}$$

where k is the number of nearest neighbors considered, and $V_{i,k}$ is the volume they occupy. The volume is computed as $V_{i,k} = \omega_{ID} d_{i,k}^{ID}$ where ω_{ID} is the unit-sphere volume in $\mathbf{R}^{ID}$ and $d_{i,k}$ is the distance between point i and its k^{th} nearest neighbor [29]. ID represents the intrinsic dimension of the dataset, i.e., the minimum number of variables needed to represent the dataset.

The minimum distance between the discovered probability density peak i and any other point with a higher density j is computed as

$$\delta_i = \min_{j|\rho_j > \rho_i} d_{ij} \tag{3.15}$$

where we are looking for both a very large ρ_i and a very large δ_i. Thus, cluster centers are found to be clear outliers in the distribution of ρ_i versus δ_i. After finding the cluster centers, the remaining points are assigned to the cluster corresponding to their nearest neighbor of higher density [29].

In order to visualize the results of the density-based clustering algorithm, we need a dimensionally reduced representation of the data. By using the previously discovered PCA space, we can compare the methods used in this analysis. This serves to validate our findings through a number of machine learning methods. In addition, by plotting the density-based clustering results in PCA space, we can extract further insight from its discovered data distribution. Points that have a higher density should have a higher certainty of belonging to the atomic specie in focus, and outliers have a higher chance of representing

contaminants, defects, or misclassifications depending on their distance from a cluster center. If a cluster of outliers forms, then a nearby low-density point is more likely to represent the atom specie in focus. Despite the low-density point's proximity to the cluster of outliers, its density is still computed relative to the cluster of true positive atoms, indicating their resemblance in the intrinsic dimension of the data.

3.3.5. Scoring Methods

We use a number of scoring methods to assess the performance of our GMM. For the first round of GMM predictions, we only score the model classification based on the atom sites that were annotated by domain experts (Figure 3.1). This means that contaminants, defects, and non-annotated sites are omitted from the first round of scoring. This was done to demonstrate that the model successfully distinguishes between Fe and Ti adatoms. During the second round of GMM predictions, for both Fe and Ti subclassifications, we score the model classifications based on whether or not Fe and Ti sites were distinguished from contaminants, defects, non-annotated sites, and misclassified Ti or Fe adatoms by the first GMM. Thus, for the first GMM, we are scoring on two classes: Fe or Ti. For the second GMM, we are scoring on two classes: Fe/Ti or outlier.

Confusion Matrix

A confusion matrix visualizes the performance of a classification in a table consisting of rows that represent the ground truth classification and columns that represent the predicted classification. The diagonals in the table represent correct classifications. We use the digitized expert-annotated data as ground truth and the GMM results as predictions.

Precision

The precision scores the model's ability to avoid misclassifications. The best value is 1 and the worst value is 0.

$$precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{3.16}$$

As we want to compute the precision for both classes, we compute a weighted average of each label's precision.

$$\frac{1}{\sum_{l \in L} |y_l|} \sum_{l \in L} |y_l| Precision \tag{3.17}$$

where L is the set of labels, y is the set of ground truth (sample, label) pairs, and y_l is the subset of y with label l [24].

Recall

The recall scores the model's ability to correctly classify all atom sites. The best value is 1 and the worst value is 0.

$$recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{3.18}$$

As we want to compute the recall for both classes, we compute a weighted average of each label's recall.

$$\frac{1}{\sum_{l \in L} |y_l|} \sum_{l \in L} |y_l| Recall \tag{3.19}$$

where L is the set of labels, y is the set of ground truth (sample, label) pairs, and y_l is the subset of y with label l [24].

F_1 Score
The F_1 score is a weighted mean of the precision and recall scores. The best value is 1 and the worst value is 0.

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.20}$$

As we want to compute the F_1 score for both classes, we compute a weighted average of each label's F_1 score.

$$\frac{1}{\sum_{l \in L} |y_l|} \sum_{l \in L} |y_l| F_1 \tag{3.21}$$

where L is the set of labels, y is the set of ground truth (sample, label) pairs, and y_l is the subset of y with label l [24].

Part II

Results and Discussion

4

Results and Discussion

In this chapter, we present the results of the analysis as described in Figure 1.1. We start with a two-class GMM to classify atoms based on their atomic species, Fe or Ti. We analyze our findings to understand the main variations between the two classes and plot them for further comparison with the ground-truth labeling. Class-conditioned classifications are then performed by taking the Fourier transforms of the atom crop data within each learned class (Fe and Ti) and performing a second round of GMM classifications. The intent is to separate outliers from true positives within each learned class. We follow this subclassification with a PCA and a density-based clustering approach performed on the class-conditioned Fourier transformed data to quantify the differences between the true positive data points and the outliers.

4.1. Classifying Atoms by Species

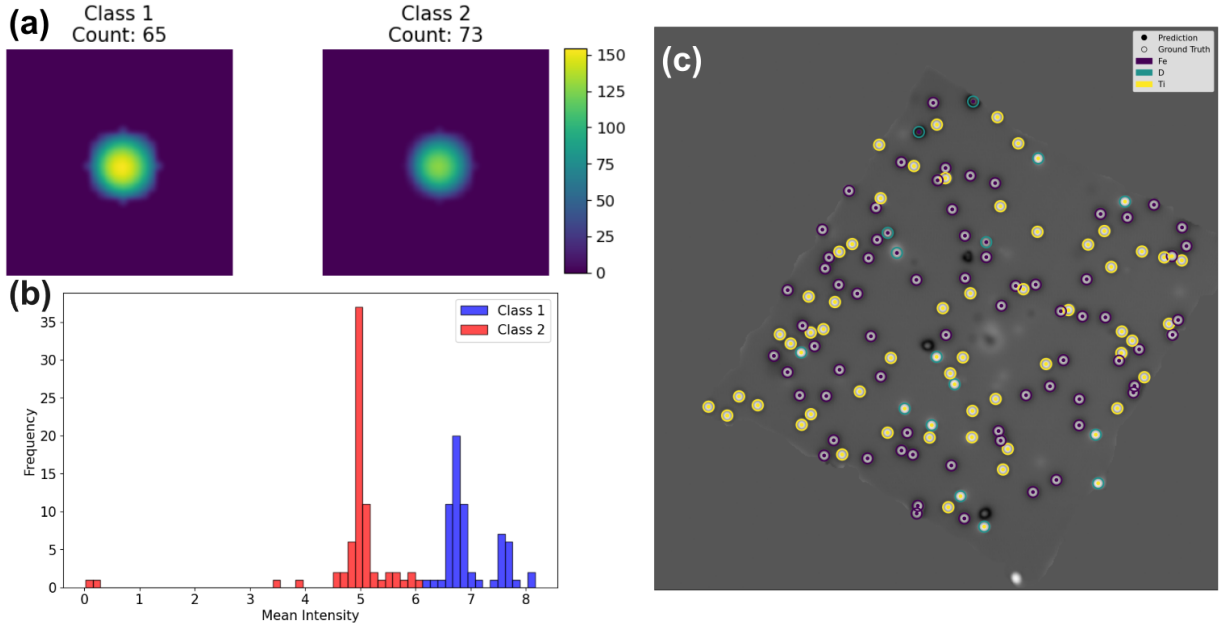


Figure 4.1: Topo A001 first GMM analysis. (a) Mean intensity image within each learned class. (b) Mean intensity distribution by class. (c) Plot of the GMM results with an overlay of the ground truth labels. The GMM classification is represented by the dots at each coordinate site. The ground truth labels are represented by the circles around each coordinate site. Fe is for iron atoms, D is for defects and unknowns (unlabeled sites), Ti is for titanium atoms.

The first step of this analysis is to feed the constructed dataset to a GMM to classify each atomic site into one of two classes (Fe or Ti). The data representation learned by the GMM can be described by the mean intensity image within each learned class (Figures 4.1a, 4.2a, 4.3a). We notice from the class distinction produced by the GMM that the mean intensity image within one class has a higher range of

values than that of the other.

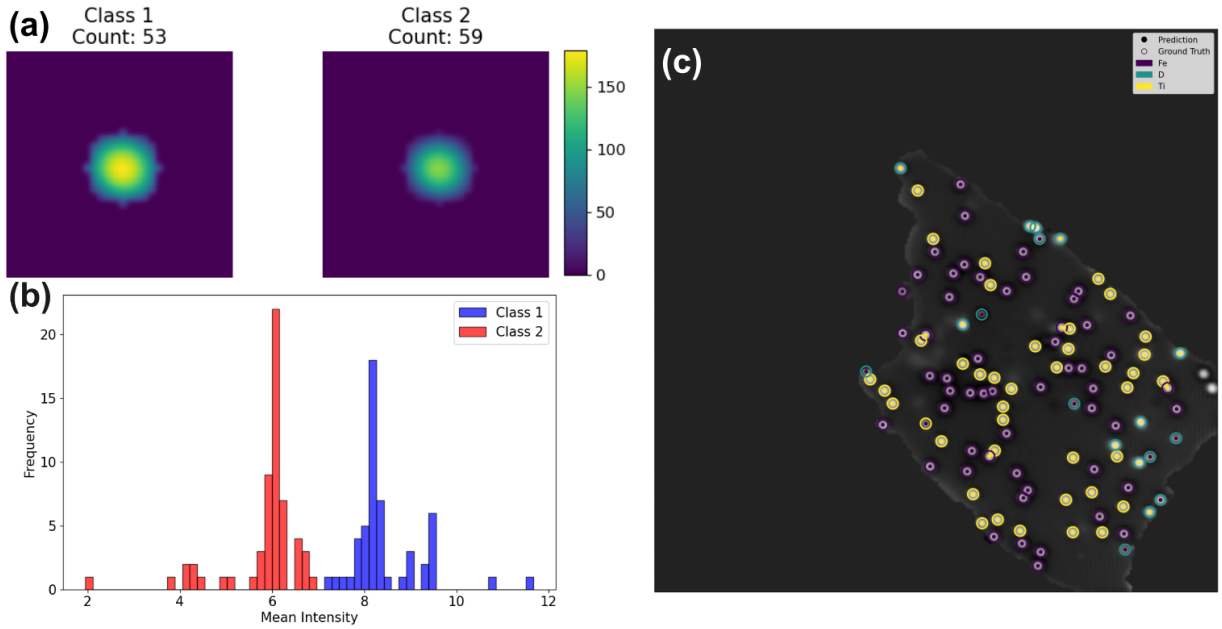


Figure 4.2: Topo B0376 first GMM analysis. (a) Mean intensity image within each learned class. (b) Mean intensity distribution by class. (c) Plot of the GMM results with an overlay of the ground truth labels.

To better understand the classification, we plot the mean intensity distribution of the data points within each class (Figures 4.1b, 4.2b, 4.3b). We notice two distributions that are clearly separated by their minima and maxima. The class with the higher value range in its mean intensity image has both higher minima and maxima in its distribution of mean intensities. Thus, the main variation between coordinate sites discovered by the GMM is quantified by their intensities. This ties into the STM imaging process, as the measured LDOS of one class of atoms should be clearly distinguishable from the other, giving rise to protrusions of higher or lower apparent height.

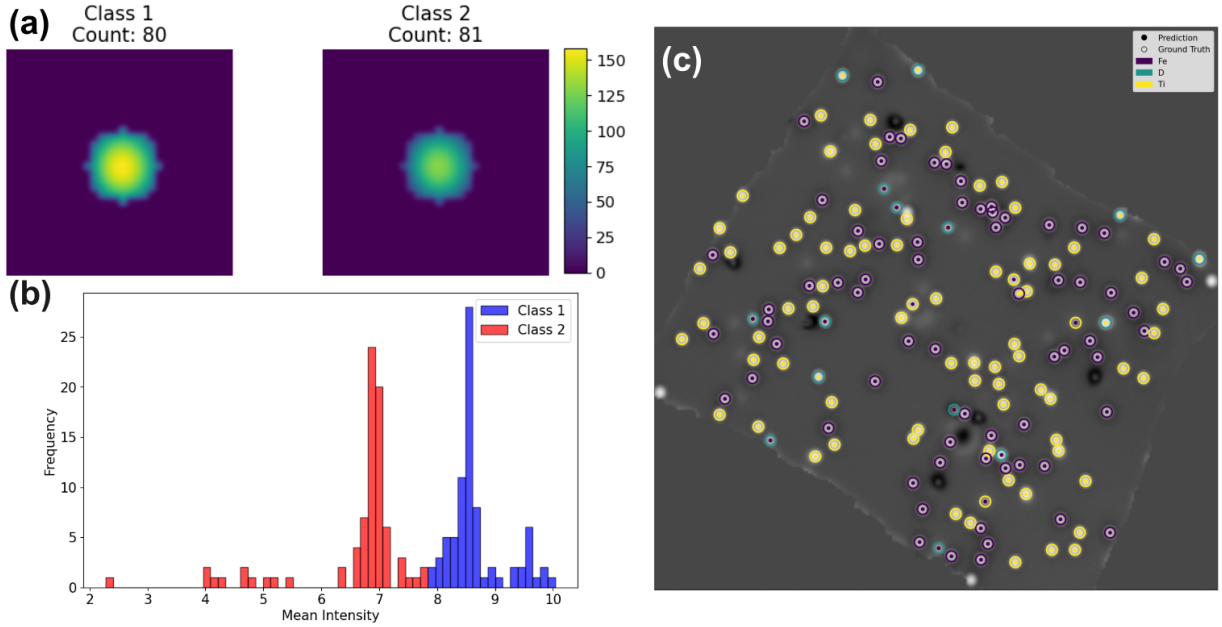


Figure 4.3: Topo B0627 first GMM analysis. (a) Mean intensity image within each learned class. (b) Mean intensity distribution by class. (c) Plot of the GMM results with an overlay of the ground truth labels.

We visualize the results and evaluate them by comparing them with domain expert annotations (Figures 4.1c, 4.2c, 4.3c). We conclude that Ti on bilayer MgO has a higher LDOS than Fe, since the class with the highest mean intensity spectrum belongs to Ti-annotated sites.

For the first round of classification, we want to evaluate how well the model performs in detecting true positive Ti and Fe atoms, omitting defects and unknowns that have not been annotated. We present our results in the form of confusion matrices for each image, showing near-perfect classification of Fe and Ti atoms (Figures 4.4a, 4.4b, 4.4c). The model's capabilities are quantified by using F1, recall, and precision scoring, where we obtain almost perfect results (superior to 0.9) for seven out of eight of the STM images provided for this work (Figure 4.4d). The model thus adequately differentiates between the Fe and Ti adatoms. The next step is to remove outliers, unknowns, and defects from our classification.

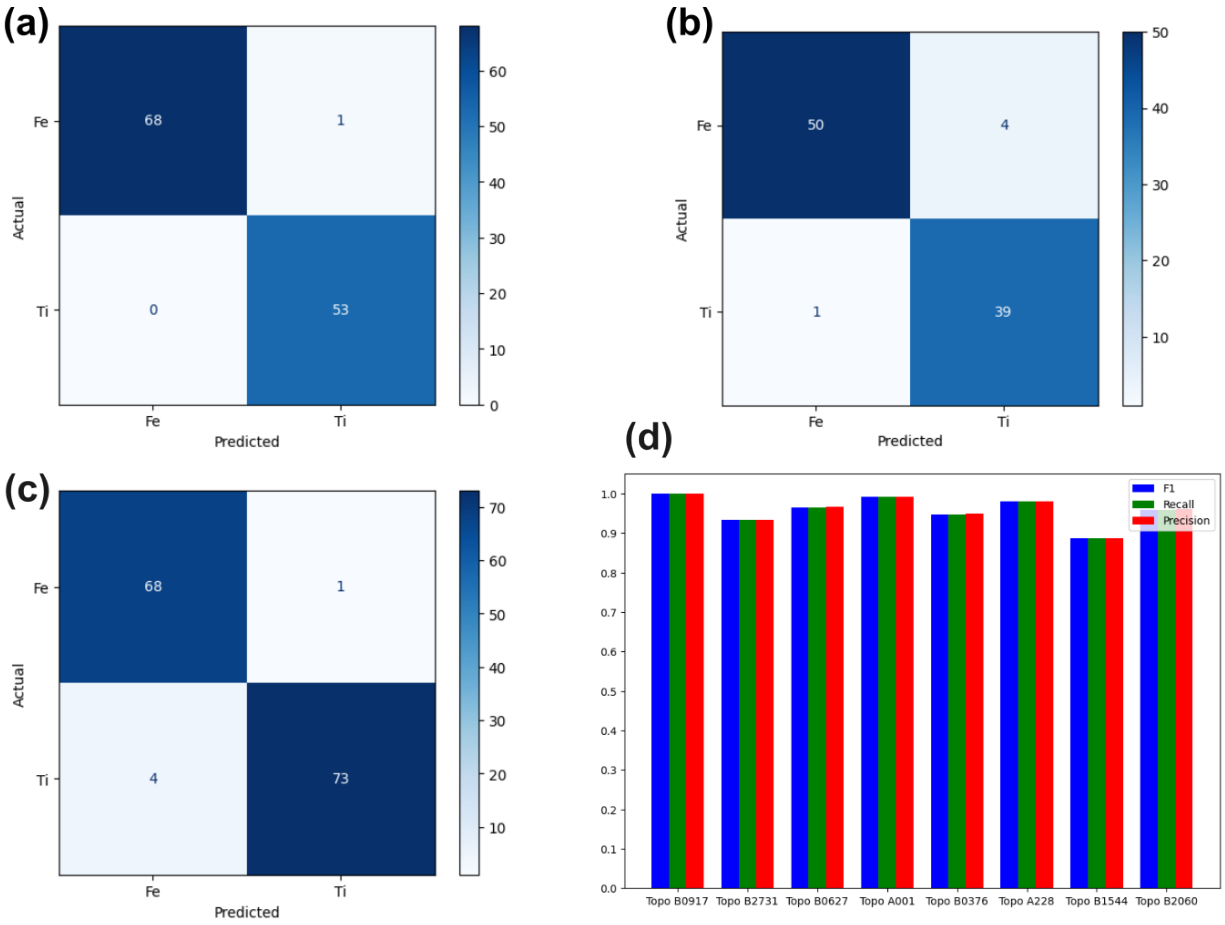


Figure 4.4: Confusion matrices of the GMM results, defects and unknowns omitted. (a) Topo A001. (b) Topo B0376. (c) Topo B0627. (d) Precision, recall, and F1 scores of the initial GMM classification for 8 different topographies, omitting defects and unknowns.

4.2. Assessing Unknowns, Defects, and Misclassifications

4.2.1. Class-Conditioned Subclassification on Fe-classified Sites

To further tune the detected atom site classification, we leverage the discrete Fourier transform's ability to extract the topographical features of an image (mean intensity, shapes, transitions between intensities) to distinguish between true positives within each learned class and defects, unknowns, or misclassifications. With this in mind, there is a risk of distinguishing atoms corresponding to dimers from single atoms because their atom crops can be topographically different due to interference in their cropped images by their nearest neighbor.

In Fourier space, dominant features lie at the center and have the highest values, intensity transitions

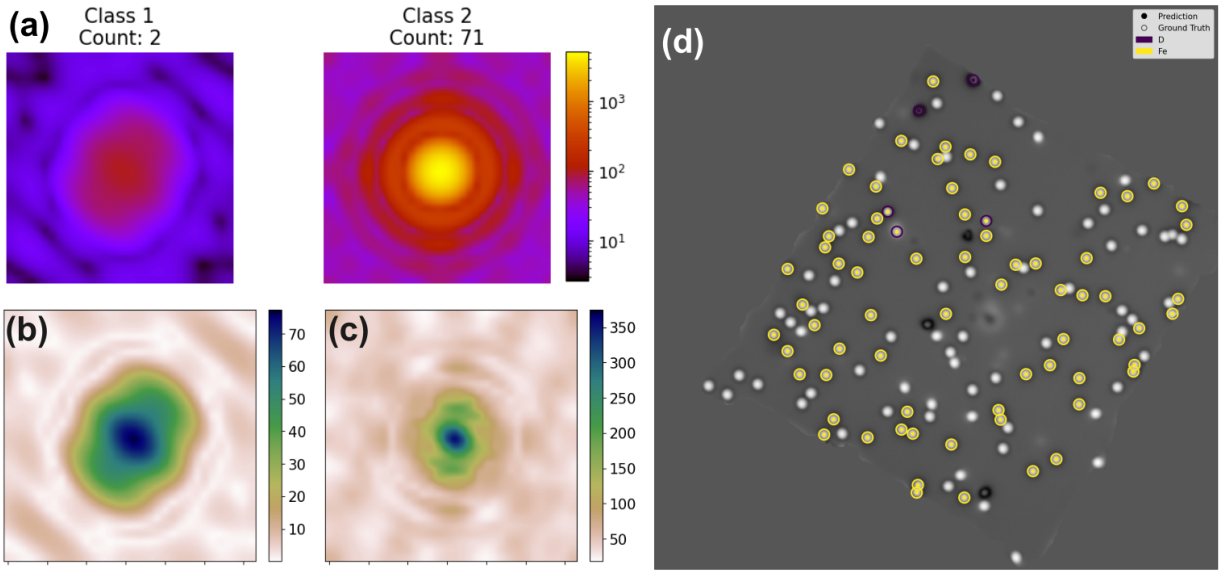


Figure 4.5: Topo A001 class-conditioned GMM analysis on Fourier transformed Fe-classified sites. (a) Mean Fourier modulus of the atom-cropped image within each learned class. (b and c) Standard deviation in Fourier space within each learned class. (b) Class 1 (outliers). (c) Class 2 (Fe atoms). The outliers in this case skew the classification. (d) Plot of the GMM results with an overlay of the ground truth labels. The GMM classification is represented by the dots at each coordinate site. The ground truth labels are represented by the circles around each coordinate site. Fe is for iron atoms, D is for defects, unknowns (unlabeled sites), and misclassified Ti atoms.

are represented by circular rings around the center, and more intricate features (such as edges and shapes) are represented as low values at high frequencies away from the center (Figure 3.12).

We perform a second round of classifications on the discrete Fourier transform moduli of our dataset within each class learned by the initial GMM. In this classification task, we wish to extract the most accurate representation of a Fe atom, so we set the GMM to distinguish between two classes: topographically similar Fe-classified sites and outliers. The learned data representation is plotted as the mean Fourier transform modulus within each class (Figures 4.5a, 4.6a, 4.7a).

To interpret our results, we plot the standard deviation within each learned class (Figures 4.5b, 4.5c, 4.6b, 4.6c, 4.7b, 4.7c). By analyzing the variation from the mean at each pixel in Fourier space for each class, we get to understand how topographically different the atom crops in each class are. We expect the class with the most topographical similarity between data points to represent single Fe atoms more accurately and thus to contain our true positive classifications. Meanwhile, the other class should contain defects, unknowns, and misclassifications.

We look at the range of values in our standard deviation plot and notice that the GMM fits the data by minimizing the standard deviation within one of the two classes. A lower range of values in the standard deviation indicates a higher topographical similarity between atom crops in each class and vice versa. The exception is when the defects detected by the GMM are so significantly different from the other coordinate sites, but similar to each other, that they skew the classification, such as in Topography A001 (Figure 4.5a).

In addition, we notice that the finer details in the atom topography appear as sources of variation in the standard deviation plot of class 2, which exhibits high topographical similarity (Figures 4.6c, 4.7c). These are represented by shapes in the surrounding area away from the central frequencies. This indicates that the differences between the data points in this class are not concentrated on their essential features. These differences may arise from features such as the atom's shape and orientation (Figure C.2). Meanwhile, the standard deviation in the other class (Figures 4.6b, 4.7b) is concentrated in those central frequencies, indicating topographical differences between its data points in their most essential features (Figure C.1). These essential features include the intensity values in the image and thus the captured coordinate site's LDOS, which translates to lower or higher apparent height.

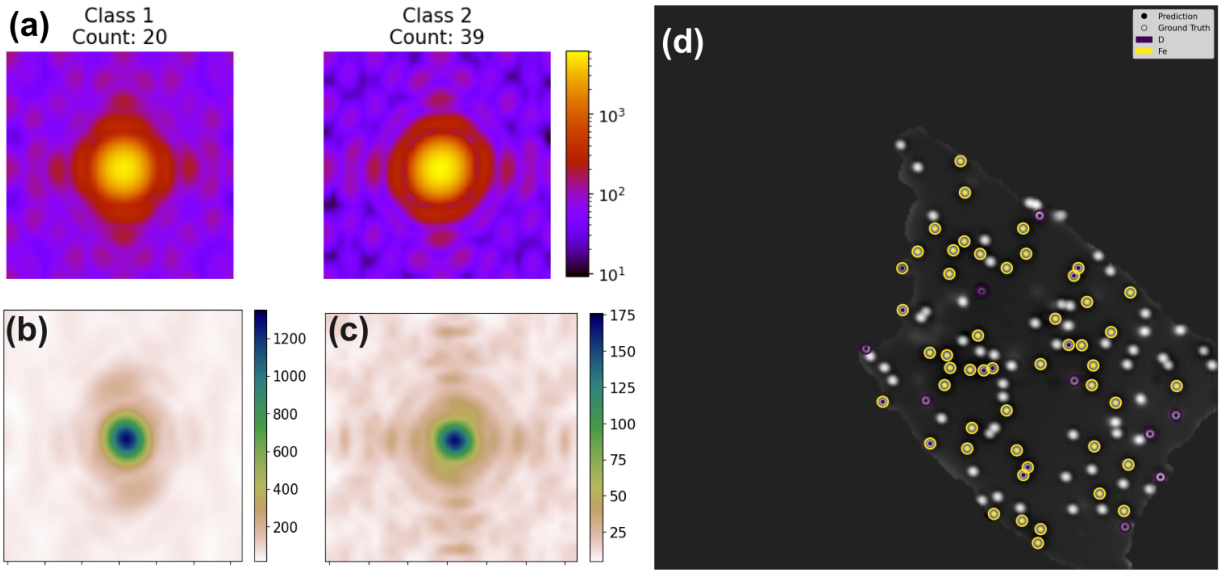


Figure 4.6: Topo B0376 class-conditioned GMM analysis on Fourier transformed Fe-classified sites. (a) Mean Fourier modulus of the atom-cropped image within each learned class. (b and c) Standard deviation in Fourier space within each learned class. (b) Class 1 (outliers). (c) Class 2 (Fe atoms). Notice that the value ranges are widely different for each class. (d) Plot of the GMM results with an overlay of the ground truth labels.

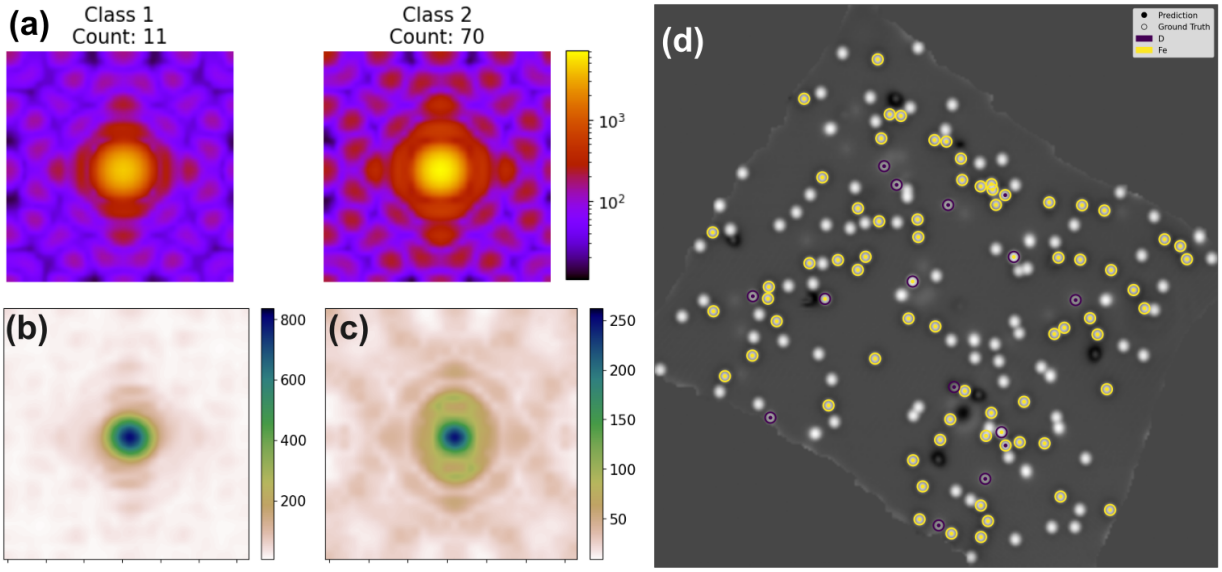


Figure 4.7: Topo B0627 class-conditioned GMM analysis on Fourier transformed Fe-classified sites. (a) Mean Fourier modulus of the atom-cropped image within each learned class. (b and c) Standard deviation in Fourier space within each learned class. (b) Class 1 (outliers). (c) Class 2 (Fe atoms). Notice that the value ranges are widely different for each class. (d) Plot of the GMM results with an overlay of the ground truth labels.

Thus, the GMM classifies atom crops based on their topographical similarities and differences, extracting the atom sites that resemble each other the most while separating outliers. We may validate this classification as we have done in the previous section, with confusion matrices and F1, recall, and precision scoring (Figure 4.8). We note that most data points are correctly classified as true positive Fe sites, with some true positives falsely classified as outliers, obtaining scores at around 0.8 or higher for all of our topographies. We notice that the precision score tends to be the highest as it indicates the model's

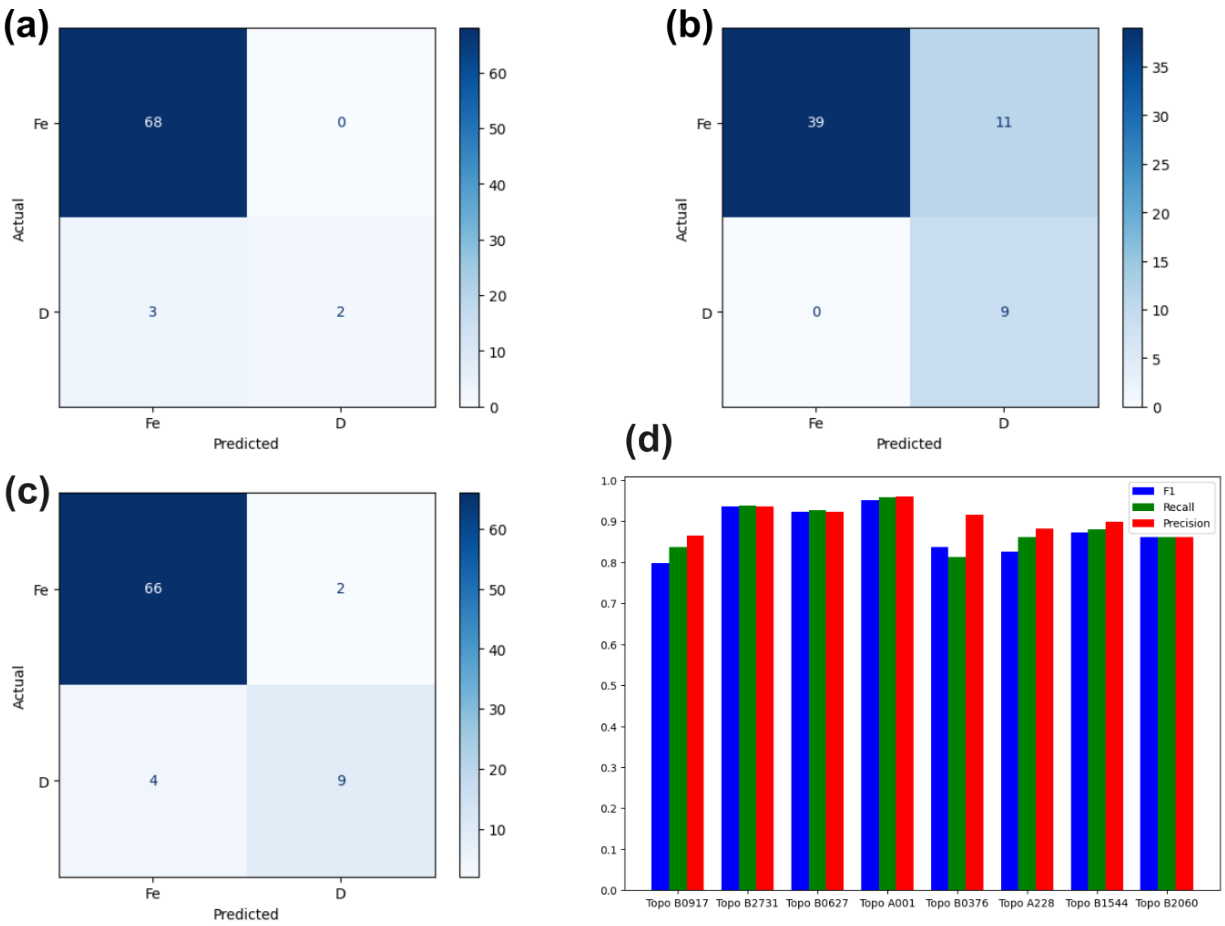


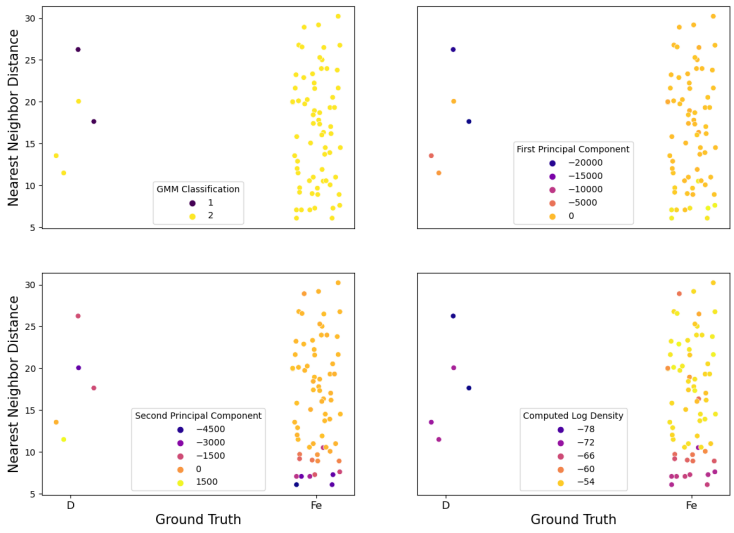
Figure 4.8: Confusion matrices of results from the Fe class-conditioned GMM classification of Fourier transforms. (a) Topo A001. (b) Topo B0376. (c) Topo B0627. (d) Precision, recall, and F1 scores of the Fe class-conditioned GMM classification of Fourier transforms for 8 different topographies.

ability to not misclassify a data point. Meanwhile, the recall is lower if there is a reason that the model cannot identify all the points belonging to a given class.

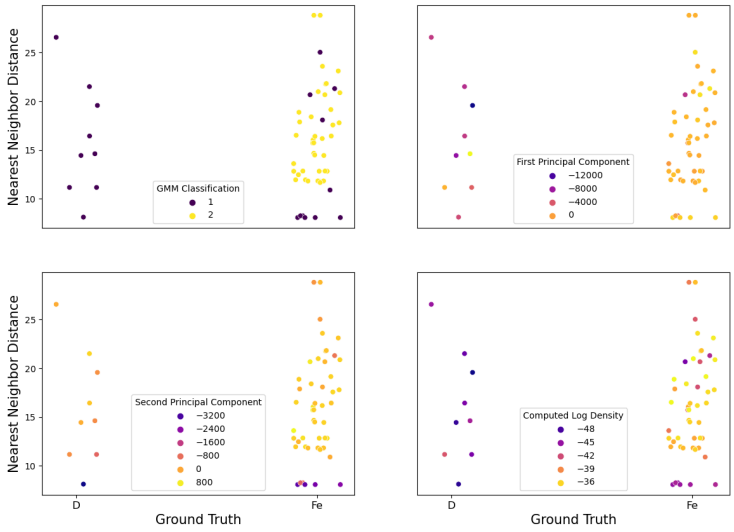
We plot our results and compare them to the ground truth labeling (Figures 4.5d, 4.6d, 4.7d) to understand the model's limitations. We notice that most false negative classifications are due to atoms that correspond to dimers having topographical differences in their atom crops compared to single atoms because of near-neighbor interference. This leads to lower recall scores as not all Fe adatoms are correctly identified.

We verify whether dimers are the cause of false negative classifications, and therefore lower recall scores, by plotting the coordinate site's distance to its nearest neighbor against its ground truth labeling, colorized by its GMM classification (Figure 4.9). We see that the Fe atoms classified as outliers lie in the same range of nearest-neighbor distances. Precisely, for Topography B0376, seven out of eleven of the falsely classified Fe atoms belong to dimers. The rest are atoms lying on the edge of the island. For Topography B0627, the two falsely classified Fe atoms are in dimers. However, many Fe atoms with close neighbors are correctly classified for Topography B0627, leading to scores higher than 0.9. Thus, atoms with near-neighbors that form dimers are the main cause of false negative misclassifications.

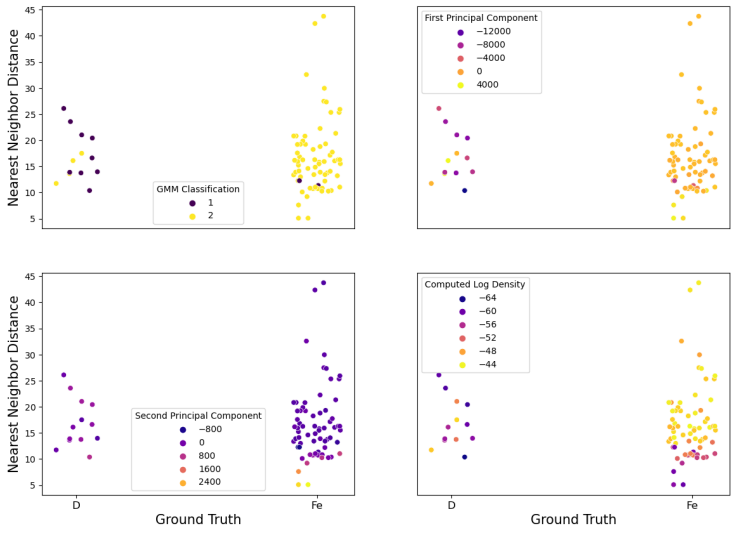
The model appears limited, as a two-component classification is too restrictive. We utilize PCA, a method that can transform our data to a new coordinate space such that the data classification is more nuanced. Analyzing our data by a coordinate space transformation can represent differences and similarities between data points more quantitatively.



(a)



(b)



(c)

Figure 4.9: Plotting the Fe-classified coordinate site distances to their nearest neighbor against the ground truth labeling, colorized by GMM classification, first PCA component, second PCA component, and computed log density. (a) Topo A001. (b) Topo B0376. (c) Topo B0627.

4.2.2. Quantifying the Variation in Fe-classified Sites

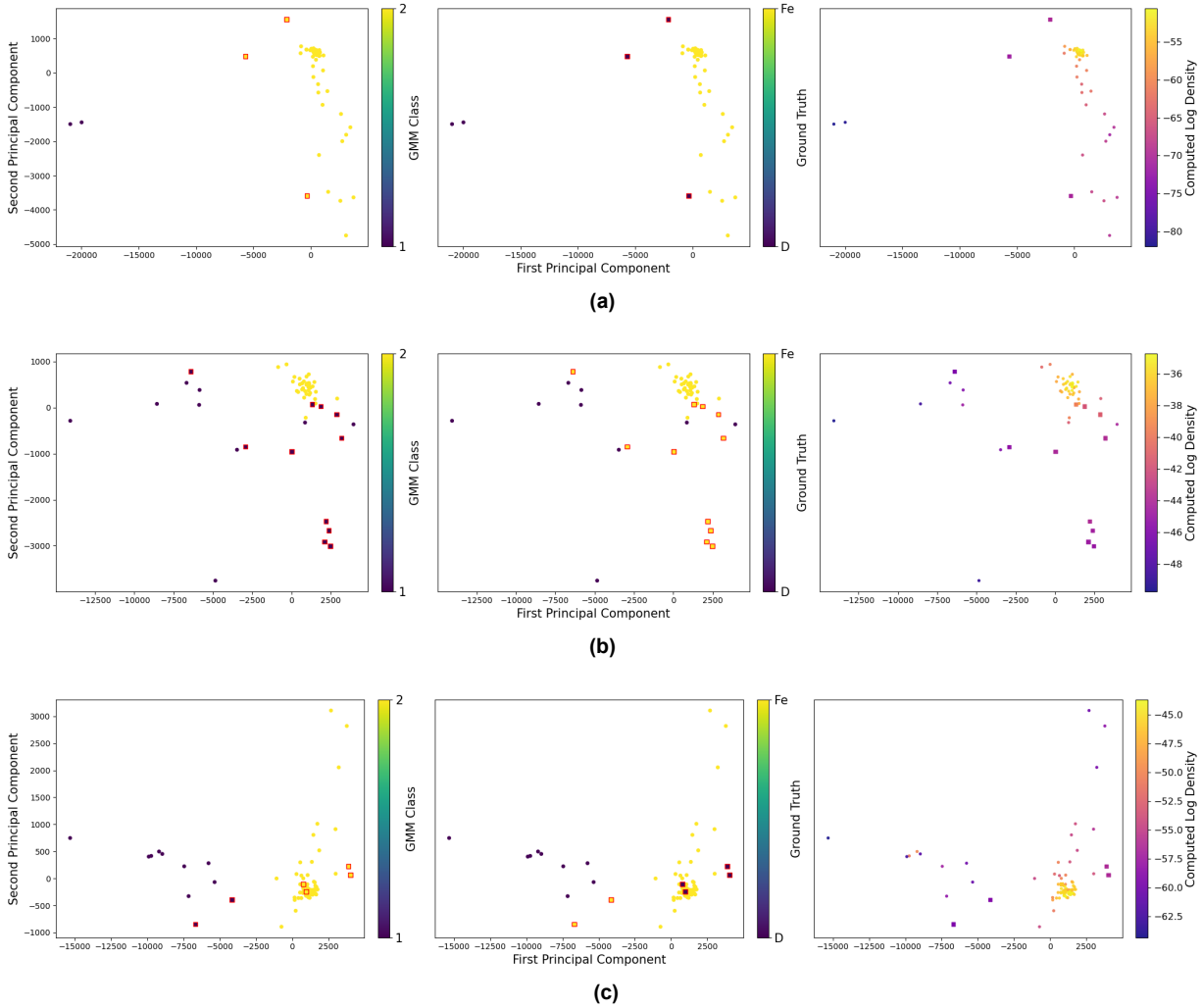


Figure 4.10: PCA space plot of our Fe data points, colorized by GMM classification, ground truth, and density-based clustering results. Adequately classified sites are represented by circles. Misclassified sites are represented by red squares. (a) Topo A001. (b) Topo B0376. (c) Topo B0627.

The next step is to perform a PCA on the Fourier transformed Fe-classified data. The goal is to utilize a linear transformation of the data points to a two-dimensional principal component space that quantifies the largest variations in the data beyond a binary classification. As mentioned in Section 3.3.3, we must plot the explained variance by component to see how much of the variation in the data is explained by the first two learned components (Figure D.1). For our three cases, we see that the main variation in the data is learned by the first two components in an average 91/9 split, with the first component strongly dominating, and the remaining 9% explained by the next few components in descending order.

We plot the two-dimensional PCA space composed of the first two, and most significant, components and colorize it according to both the GMM subclass predictions and the ground truth (Figure 4.10). We see that a main cluster of points forms in our three cases, where we find true positive Fe sites for the majority of points. These are the same points that were classified together by the GMM based on their topographical similarity. Plotting each individual principal component (Figure D.2), we see that the first principal component has its highest values in the central Fourier frequencies, where the most essential features are represented. This means that the most essential topographical features are the biggest source of variation in the first principal component and therefore in the dataset. This complements our interpretation of the GMM classification: in one class, we find the most likely topography of an Fe atom site, because its data points are more similar in their essential features - their intensity values - than in the

other class. Therefore, they lie in the same range of values on the first principal component axis.

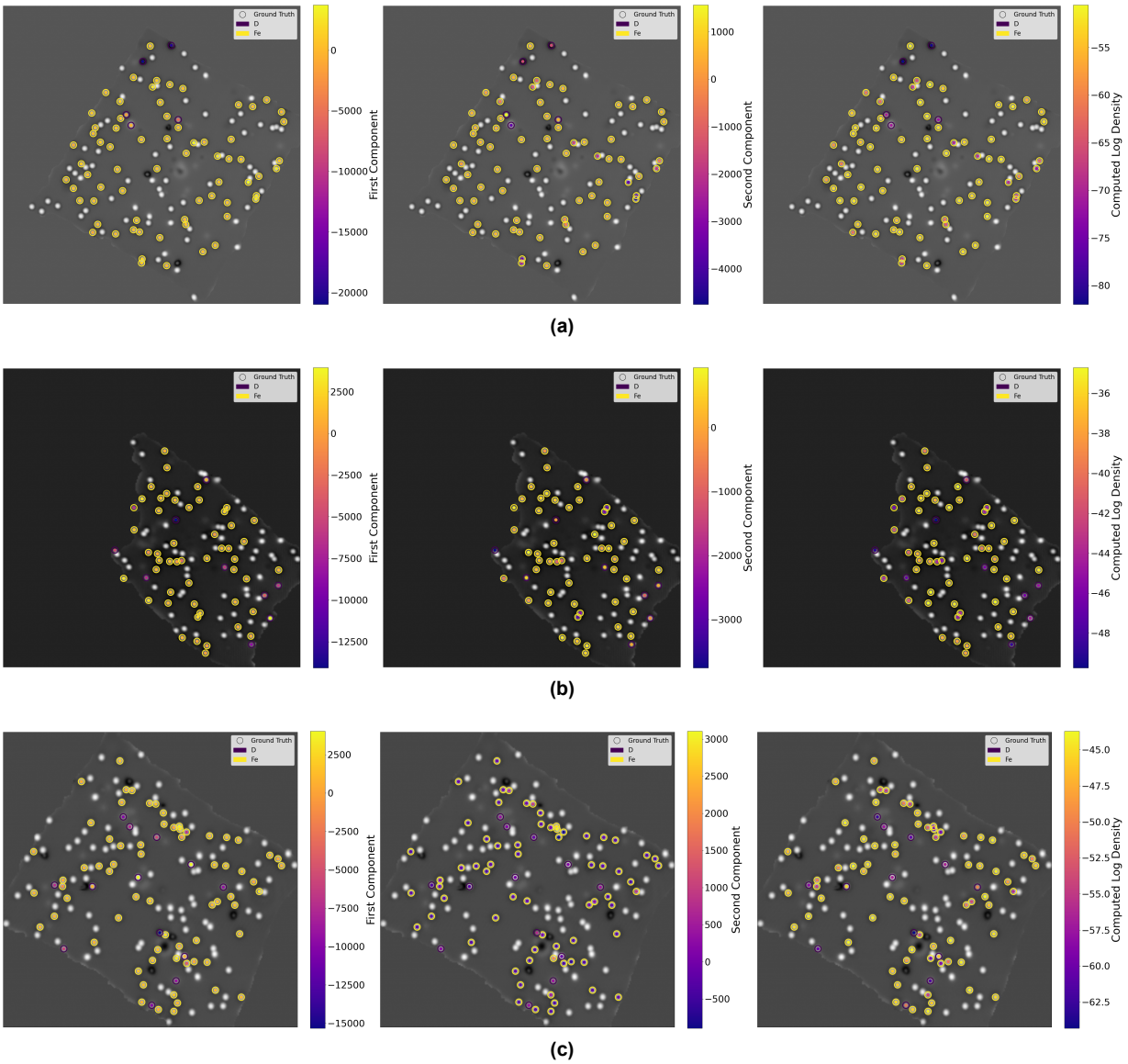


Figure 4.11: Projecting the PCA and density-based clustering results onto our Fe-classified coordinate sites. (a) Topo A001. (b) Topo B0376. (c) Topo B0627.

In addition, this means that the outliers on the first principal component axis are distinguishable by their essential features. This gives them a high probability of being defects, unknowns, or misclassified Ti sites. In this case, comparing the GMM class and ground truth color-coded PCA plots (Figure 4.10) shows that the first principal component is more accurate in grouping data points by their essential features than the GMM. In our plots, misclassified sites are highlighted as red squares. As we can see, points in the same range on the first principal component axis as the main cluster of points are more likely to be Fe atoms. Some of these are in fact misclassified by the GMM (Figure 4.10b).

These misclassified sites are only considered as outliers along the second principal component, shedding light on the shortcomings of the GMM. This could mean that the second principal component outlines atoms that correspond to dimers. We verify this by plotting the coordinate site's distance to its nearest neighbor against its ground truth labeling, and colorize the plots by its value on both the first and second principal component axes (Figure 4.9). Through these two plots, we see that by combining the insight from both principal components, we may rectify the GMM classifications. We may also project the principal components onto our coordinate sites to obtain a better visualization of the quantifiable differences

between atom sites (Figure 4.11). We can visually confirm that the second principal component delineates coordinate sites with close nearest neighbors. Thus, through PCA, we have managed to properly identify some of the Fe sites that correspond to dimers, thanks to the first and second principal components identifying Fe atoms that correspond to dimers.

We also apply a density-based clustering approach to assess the topographical similarity between data points as an alternative to the GMM. By discovering cluster centers in the dataset, we extract the ideal representations of Fe atoms. The data points are then clustered based on their proximity to a nearest neighbor with higher density, as mentioned in Section 3.3.4. This process scores the data points based on their similarity to the cluster centers, and thus based on their likelihood of being Fe atoms. To visualize the clustering results, we need a two-dimensional reduction of our data, so we use the same PCA space for comparison (Figure 4.10). We see that the clusters found by PCA are the same as the ones found by density-based clustering, further proving the topographical similarity between the Fe atoms subclassified as true positives.

In view of supporting researchers in their STM measurements, this method provides a certainty measure of each atom site's composition in a straightforward way. In contrast to PCA, a single metric is used to describe topographical similarity between data points rather than a number of components. By projecting the clustering results onto our data, we get a visual assessment of the likelihood that a coordinate site represents a Fe adatom (Figure 4.11). Once again, there is a visible limitation when it comes to assessing atoms in dimers due to their nearest neighbor interfering with their atom crops. However, when we colorize the nearest neighbor distance against ground truth labeling plot by computed density (Figure 4.9), we see that density-based clustering is clearly differentiating Fe atoms in dimers as outliers.

4.2.3. Class-Conditioned Subclassification on Ti-classified Sites

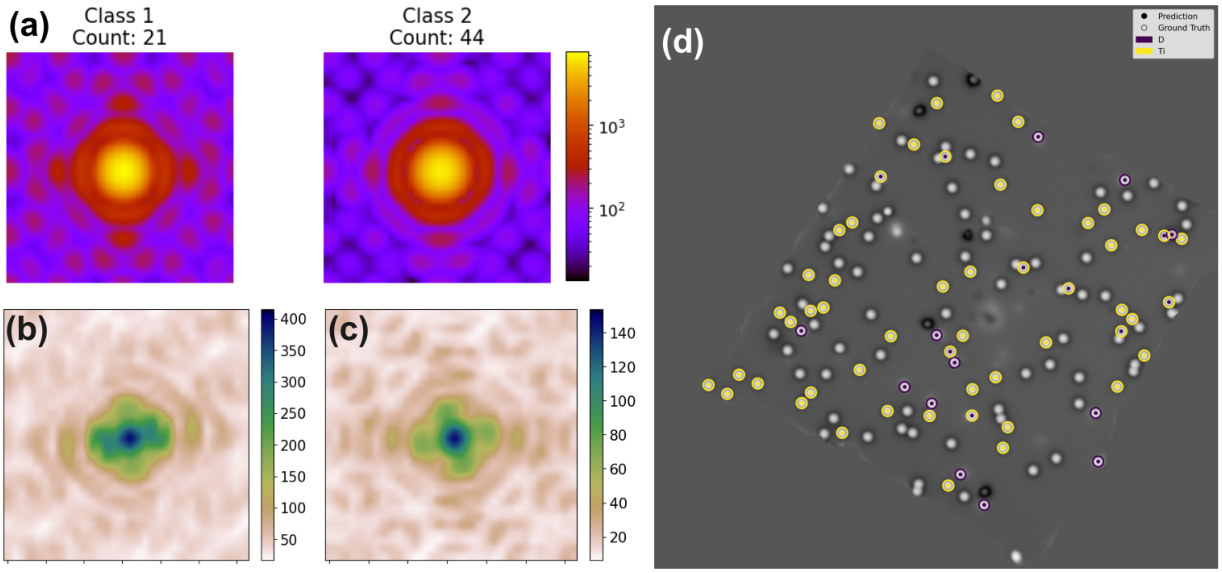


Figure 4.12: Topo A001 class-conditioned GMM analysis on Fourier transformed Ti-classified sites. (a) Mean Fourier modulus of the atom-cropped image within each learned class. (b and c) Standard deviation in Fourier space within each learned class. (b) Class 1 (outliers). (c) Class 2 (Ti atoms). Notice that the value ranges are widely different for each class. (d) Plot of the GMM results with an overlay of the ground truth labels. The GMM classification is represented by the dots at each coordinate site. The ground truth labels are represented by the circles around each coordinate site. Ti is for titanium sites, D is for defects, unknowns (unlabeled sites), and misclassified Fe sites.

Once again, we leverage the discrete Fourier transform's ability to extract the topographical features of an image to distinguish between true positive Ti sites found by the initial GMM classification and defects, unknowns, or misclassifications. Similarly to the Fe case, we expect the topography of an atom site to be represented in Fourier space such that the dominant features lie at the center where the Fourier modulus is the highest, the intensity transitions are represented by rings around the center, and the more intricate

features (such as edges and shapes) are represented as lower value shapes in the high frequencies away from the center, oriented vertically, horizontally, or diagonally depending on their orientation in the original image (Figure 3.12).

By using an unsupervised classification method like GMM on Fourier transformed image data, we may classify atom sites based on their topographical similarities with the aim of finding a subclass of topographically ideal true positive Ti-sites. Thus, we set the GMM to distinguish between two classes: topographically similar Ti-classified sites and outliers.

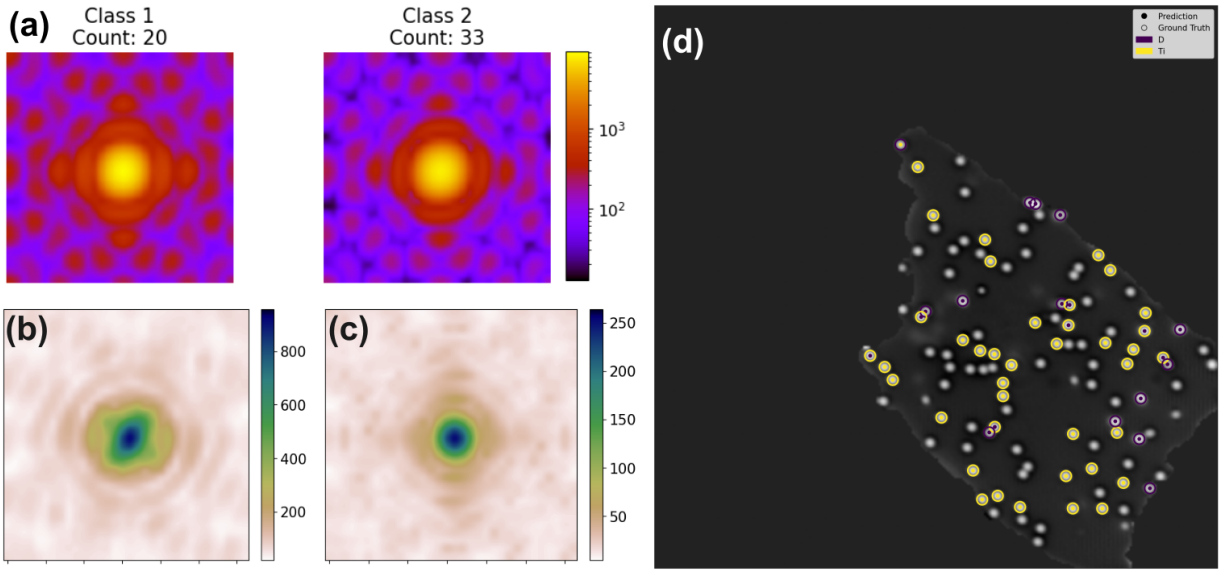


Figure 4.13: Topo B0376 class-conditioned GMM analysis on Fourier transformed Ti-classified sites. (a) Mean Fourier modulus of the atom-cropped image within each learned class. (b and c) Standard deviation in Fourier space within each learned class. (b) Class 1 (outliers). (c) Class 2 (Ti atoms). Notice that the value ranges are widely different for each class. (d) Plot of the GMM results with an overlay of the ground truth labels.

The GMM outputs learned representations based on a best-fit classification of the data according to its latent variables. In this case, the latent variables characterize two separate classes of topographically similar data points and outliers. We represent these learned representations by the mean frequency image contained within each class (Figures 4.12a, 4.13a, 4.14a). As we have previously discussed, if the Fourier transform captures the atom's topographical features, then topographical similarities between data points can be measured by assessing the standard deviation between data points in Fourier space.

We verify this statement by plotting the standard deviation between data points within each GMM subclass (Figures 4.12b, 4.12c, 4.13b, 4.13c, 4.14b, 4.14c). The first thing to notice is the stark difference in the range of standard deviation values for the two learned classes. This indicates that the class with the substantially lower variances from the mean contains topographically similar points, while the other class contains highly dissimilar topographies.

To further understand the model's classification, we examine the trends that appear in our standard deviation plots. We see that the standard deviation in the central frequencies is much higher in class 1 than in class 2, leading to a skewed distribution of the standard deviation values (Figures 4.12b, 4.13b, 4.14b). This implies that although the surrounding spatial frequencies are still relevant to distinguish between data points, the central spatial frequencies dominate. Therefore, the data points in class 1 are significantly different in their essential features (Figure E.1). They can thus be labeled as outliers belonging to different species than the data points in class 2, or they may have different intensity images due to a near neighbor.

When observing the trends in the class 2 standard deviation plot for the same topographies (Figures 4.12c, 4.13c, 4.14c), we see that the standard deviation in the surrounding frequencies are much lower, especially compared to class 1. This indicates a strong topographical resemblance between data points

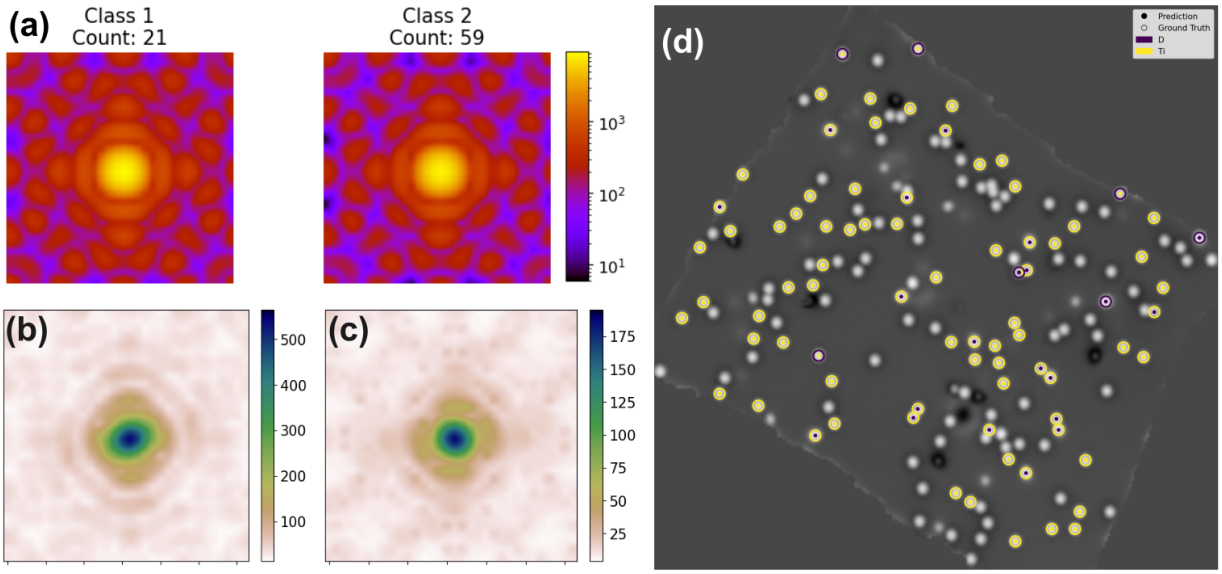


Figure 4.14: Topo B0627 class-conditioned GMM analysis on Fourier transformed Ti-classified sites. (a) Mean Fourier modulus of the atom-cropped image within each learned class. (b and c) Standard deviation in Fourier space within each learned class. (b) Class 1 (outliers). (c) Class 2 (Ti atoms). Notice that the value ranges are widely different for each class. (d) Plot of the GMM results with an overlay of the ground truth labels.

in their finer features (Figure E.2). The standard deviation in the central frequencies is also much lower, indicating a higher resemblance in the essential features of class 2 atoms than in class 1.

We validate the results of the GMM subclassification with confusion matrices and precision, recall, and F1 scoring methods (Figure 4.15). We see that most true positive Ti sites are in fact classified in the same class, with most defects, unknowns, and misclassifications classified in the other. This is reflected in the precision scores achieved for two of our three topographies, that have values around or just under 0.9. These are the cases of Topographies A001 and B0376, where the model is only slightly prone to misclassifying true positive sites and is successful in separating outliers.

The misclassified sites are for the most part Ti atoms in dimers or trimers, as we can see by projecting the GMM results onto our coordinate sites and comparing them with the ground truth labeling (Figures 4.12d, 4.13d). We may further confirm this by plotting the coordinate site's distance to its nearest neighbor against its ground truth labeling, colorized by its GMM classification (Figure 4.16). We see that eight out of the nine misclassified Ti sites are in dimers for Topography A001. For Topography B0376, six out of the seven misclassified Ti sites are in dimers. This leads to a slightly lower recall score as the model is underperforming in finding all true positive Ti atoms.

In the case of Topography B0627, there is a large number of misclassified true positive Ti sites, and a much smaller number of outliers to detect. Due to the low number of outliers and their similarity to true positive Ti atoms, the model is forced to separate true positive Ti atoms to complete its classification. These separated Ti atoms seem to have higher intensity values (Figure E.1c). In this case, a second round of classification is redundant. Two classes of Ti atoms are discovered, one of which contains the majority, while also separating half of the outliers. The majority of correct classifications despite the distinction between Ti atoms is reflected in the high precision score compared to the much lower recall score.

We must once again project the GMM results onto our coordinate sites to make sense of the falsely classified Ti atoms (Figure 4.14d). We notice that Topography B0627 contains a large number of dimers composed of Ti atoms, as seen in the ground truth labeling. Through our nearest neighbor plot (Figure 4.16c), we can distinguish eleven out of eighteen Ti atoms in dimers misclassified as outliers. This shows a clear limitation of our model. To counter this, we use more quantifiable methods to refine this subclassification, such as PCA or density-based clustering.

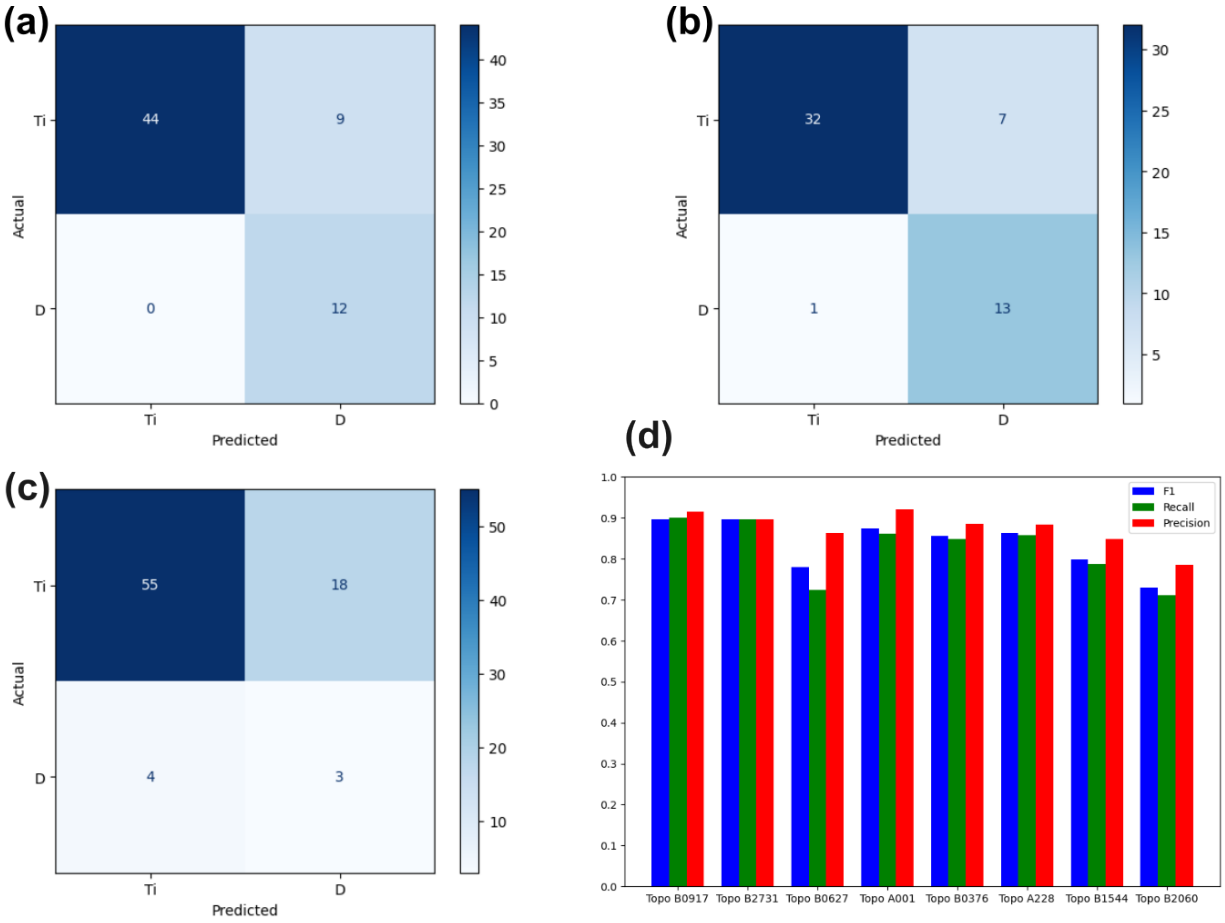
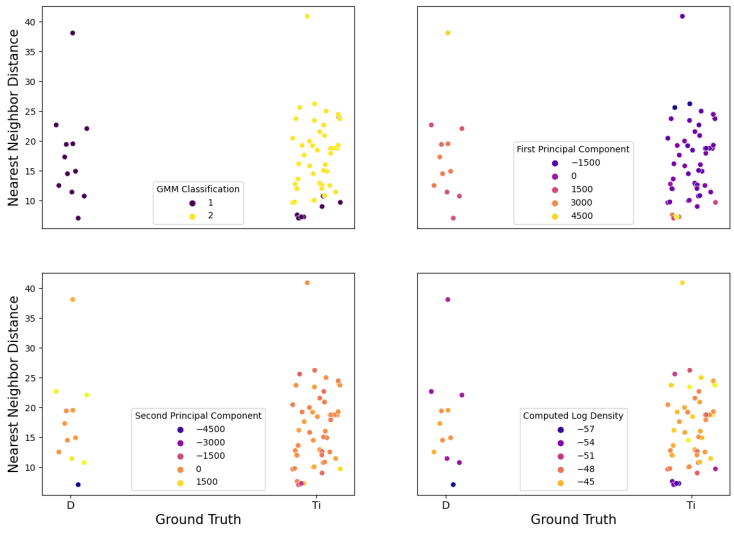
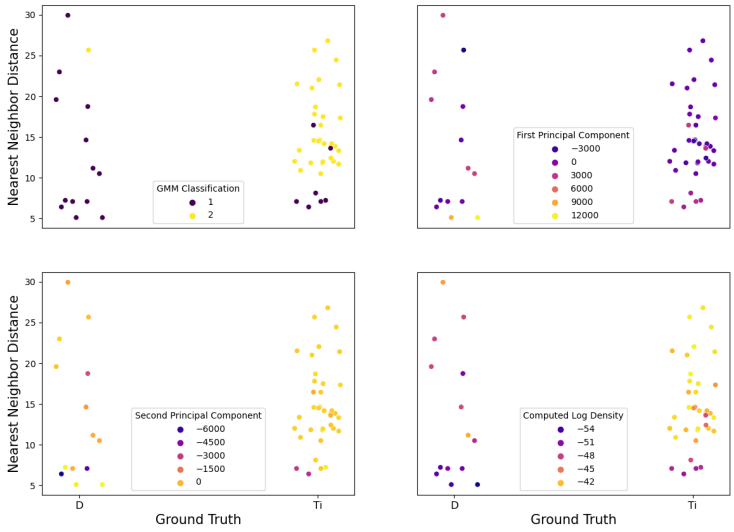


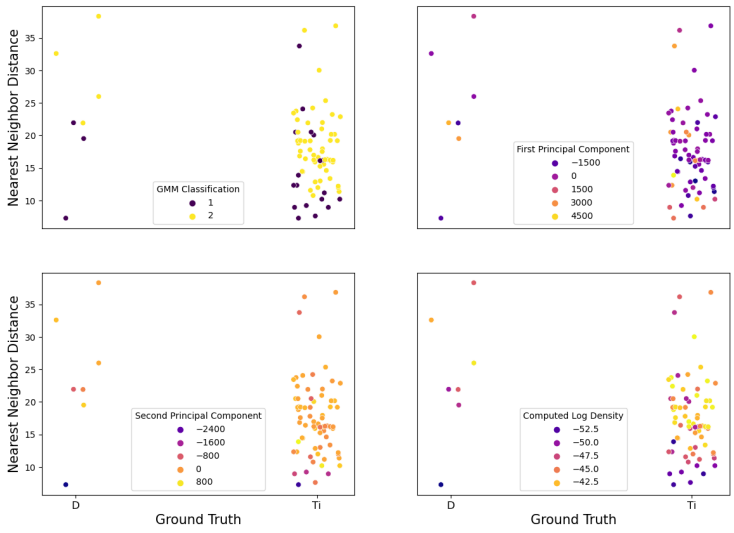
Figure 4.15: Confusion matrices of results from the Ti class-conditioned GMM classification of Fourier transforms. (a) Topo A001. (b) Topo B0376. (c) Topo B0627. (d) Precision, recall, and F1 scores of the Ti class-conditioned GMM classification of Fourier transforms for 8 different topographies.



(a)



(b)



(c)

Figure 4.16: Plotting the Ti-classified coordinate site distances to their nearest neighbor against the ground truth labeling, colorized by GMM classification, first PCA component, second PCA component, and computed log density. (a) Topo A001. (b) Topo B0376. (c) Topo B0627.

4.2.4. Quantifying the Variation in Ti-classified Sites

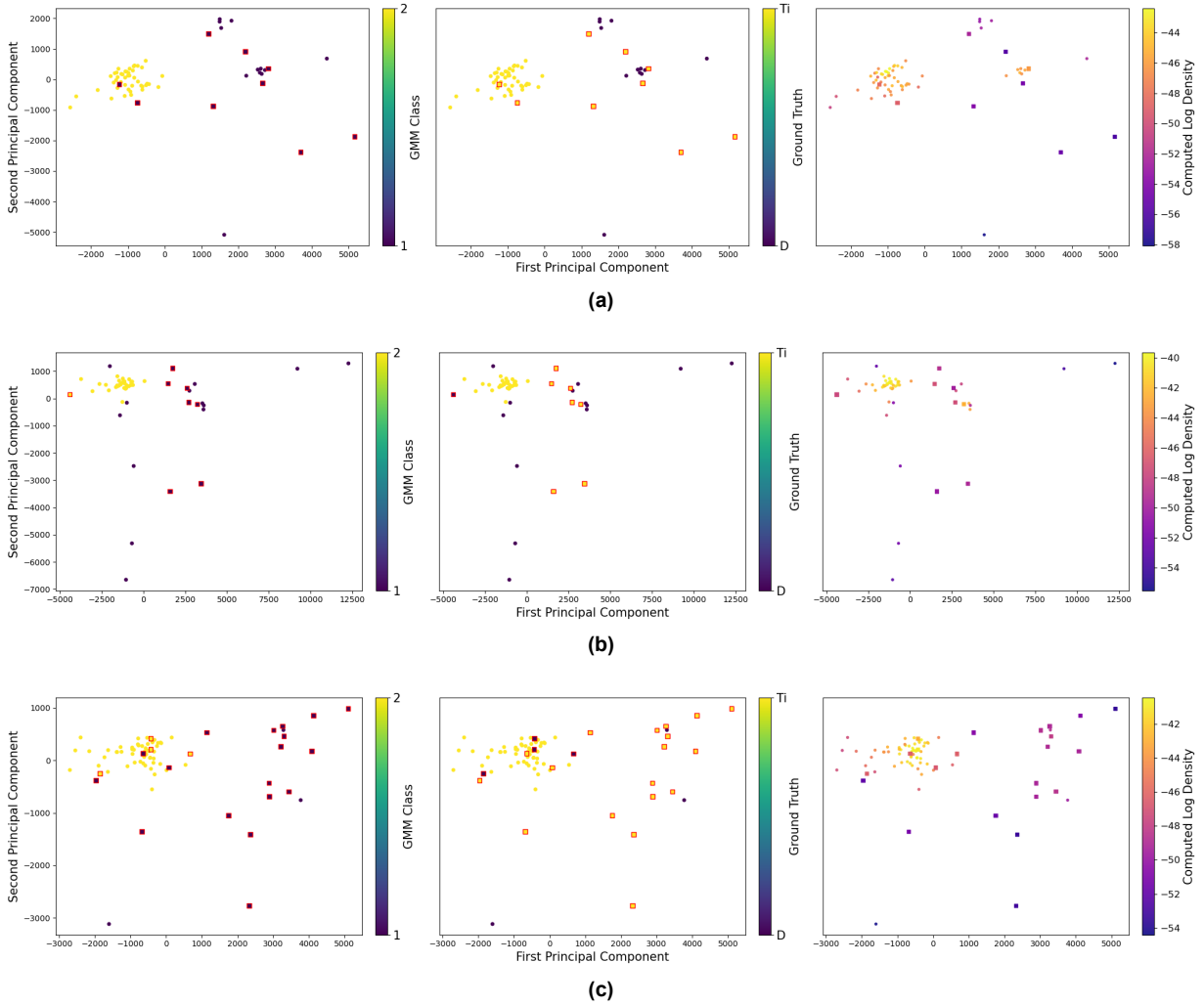


Figure 4.17: PCA space plot of our Ti data points, colorized by GMM classification, ground truth, and density-based clustering results. Adequately classified sites are represented by circles. Misclassified sites are represented by red squares. (a) Topo A001. (b) Topo B0376. (c) Topo B0627.

Once again, we wish to transform our Ti subclassified data to a two-dimensional reduced space. We utilize PCA to extract the main latent components in the data and quantify the topographical differences between the data points. We must first verify that the first two principal components represent the largest variation in the data. To this end, we plot the explained variance by principal component (Figure F.1). We see an average 83/17 split in explained variance between the first two components and the rest. For Topography B0627, the decrease in explained variance between the first component and the second is much greater, urging us to look at the learned components in more detail (Figure F.2).

For all three cases, the first principal component has its highest value at the center of the two-dimensional Fourier transform, where the Fourier modulus is the largest. It also seems to take lower Fourier modulus shapes in the surrounding frequencies as strong sources of variation in the dataset. This indicates that the first component encompasses both the essential and the finer topographical details of our data. On the other hand, the second component learns the same feature for all three cases represented by high values in two neighboring shapes that form slightly away from the center. We must plot our data in PCA space to further make sense of this.

We plot the Ti-classified data points against the first and second principal components and colorize them by their GMM subclassification and their ground truth assignments (Figure 4.17). For our three cases, a main cluster of points appears, where the GMM correctly assigns true positive Ti atoms, bar a few

misclassifications represented as red squares in the plot. However, unlike for the Fe-classified data points, we have a much more scattered distribution of points outside of this cluster. This is especially relevant for the first principal component axis, as it is supposed to encode the topographical features that distinguish atomic species from each other.

In the cases of Topographies A001 and B0376 (Figures 4.17a, 4.17b), some of the falsely classified Ti atoms can be assessed based on their distance to the main cluster on the first principal component. However, other falsely classified Ti atoms belong to outlier clusters that form away from the main cluster on the first principal component. As we discussed earlier, the majority of false negative Ti atoms belong to dimers, where the mean intensity is skewed higher due to interference from a near-neighbor. This leads to a lack of resemblance in the central Fourier frequencies with the true positive Ti atoms, where the first principal component is highest. Unfortunately, it is unclear whether the combined insight from the subsequent principal components along with the first could allow us to make the distinction between Ti atoms in dimers and the rest of the outliers.

In the case of Topography B0627 (Figure 4.17c), we see scattered Ti data points on the first principal component axis with a small outlier cluster forming. The dispersion of these data points coupled with the low explained variability for the second learned component of Topography B0627 shows that the Ti dataset in this case has lower variation between its data points due to the low number of defects and unknowns.

We can further analyze the data distribution in PCA space by plotting the atom site distance to its nearest neighbor against the ground truth, colorized by the first and second principal components (Figure 4.16). We realize that Ti atoms that were misclassified due to near-neighbor interference in their cropped images cannot be identified by combining both components for Topographies A001 and B0376 (Figures 4.16a, 4.16b). In the case of Topography B0627, the data points outside of the main cluster are too scattered to draw any conclusions. The small number of defects and unknowns in the Ti dataset makes the model differentiate between Ti atoms.

We may verify our results by projecting the two principal components on our coordinate sites and comparing with the ground truth labeling (Figure 4.18). For Topography A001 (Figure 4.18a), we see that the first principal component differentiates between true positive and defect single-atom sites very accurately. In addition, for Topography B0376 (Figure 4.18b), data points located away from the main cluster are for the most part correctly identified as outliers or corresponding to dimers. However, both principal component plots must be combined to arrive at these conclusions. To better analyze data points corresponding to dimers, we require a method that utilizes a global representation of our data to quantify topographical differences between data points instead of multiple components, such as density-based clustering.

We need a two-dimensional space to visualize the density-based clustering results, so we use the PCA space for the sake of comparison (Figure 4.17). Interestingly, the clusters discovered by the PCA are the same ones discovered by the density-based clustering method. We see that smaller clusters are formed away from the main cluster of points, where a separate high-density cluster center is found. Since this is a density-based method, we must take into account both the computed density for each point and the density peak it is relative to, made clear by the point's distance to nearby peaks.

When assessing the main cluster, the computed density of the surrounding points gives an accurate measurement of the probability that the data point is a true positive Ti atom. Low-density points near the main cluster are in fact outliers, as we can see by comparing to the ground-truth PCA plots for Topographies A001 and B0376 (Figures 4.17a, 4.17b). The smaller clusters are mainly made up of defects, unknowns, and misclassifications.

To rectify the misclassified Ti atoms by the GMM, we see that low-density points closer to the smaller cluster of outliers are more likely to be Ti atoms. Despite being closer in distance to the cluster of outliers, their densities are still being computed relative to the main cluster of Ti points. By plotting the nearest-neighbor distance against the ground truth label and colorizing by the computed density of each data point (Figures 4.16a, 4.16b), we see that these misclassified low-density Ti atoms correspond to dimers.

This shows that the density-based clustering algorithm takes into account the resemblance between the Ti atoms corresponding to dimers and single Ti atoms, while separating defects and unknowns into their own cluster. We can identify these points by projecting the density-based clustering results on our data (Figures 4.18a, 4.18b). Thus, this method corrects some of the dimer atom misclassifications.

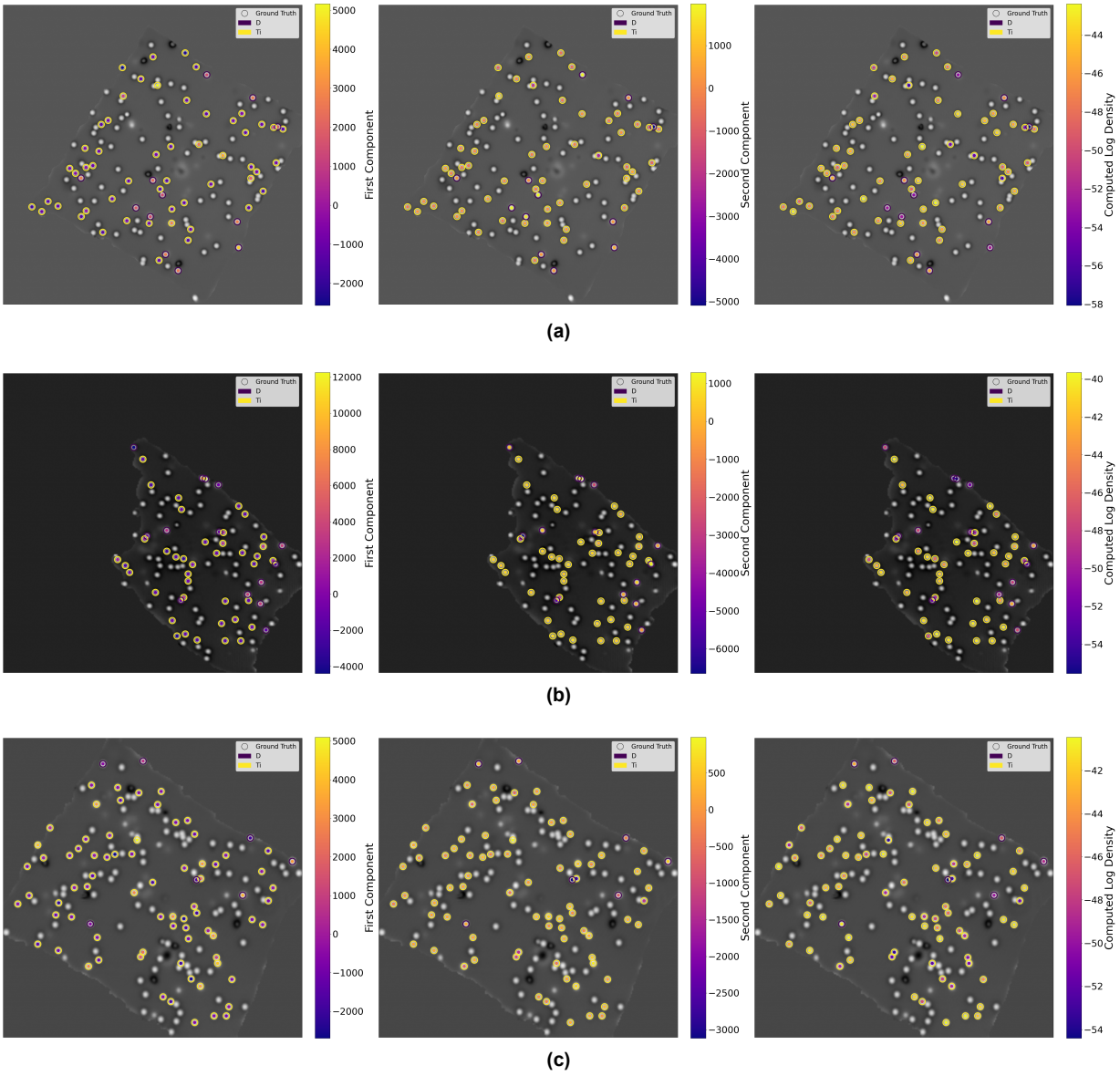


Figure 4.18: Projecting the PCA and density-based clustering results onto our Ti-classified coordinate sites. (a) Topo A001. (b) Topo B0376. (c) Topo B0627.

In the case of Topography B0627 (Figure 4.17c), the data is more scattered and requires closer inspection. As mentioned previously, this dataset has a small number of defects and unknowns that are not significantly different from the Ti atoms, which leads to the model finding topographical differences between Ti atoms.

Similarly to the other two cases, the density-based clustering algorithm appears to detect a smaller cluster of outliers. However, in this case, the outlier cluster center has a much lower density because of the close distance between the cluster centers in the intrinsic dimension of the dataset. This is indicative of the similarity between the outliers and the points in the main cluster, since they mostly correspond to Ti atoms. Therefore, the second round of GMM classification on the Topography B0627 Ti dataset is redundant.

4.3. Parameter and Data Processing Choices

In this section, we justify the parameter and data processing choices made in Sections 4.1 and 4.2. This includes discussions on the number of classes chosen for the GMM classification of atom species in Section 4.1 and the use of Fourier transforms.

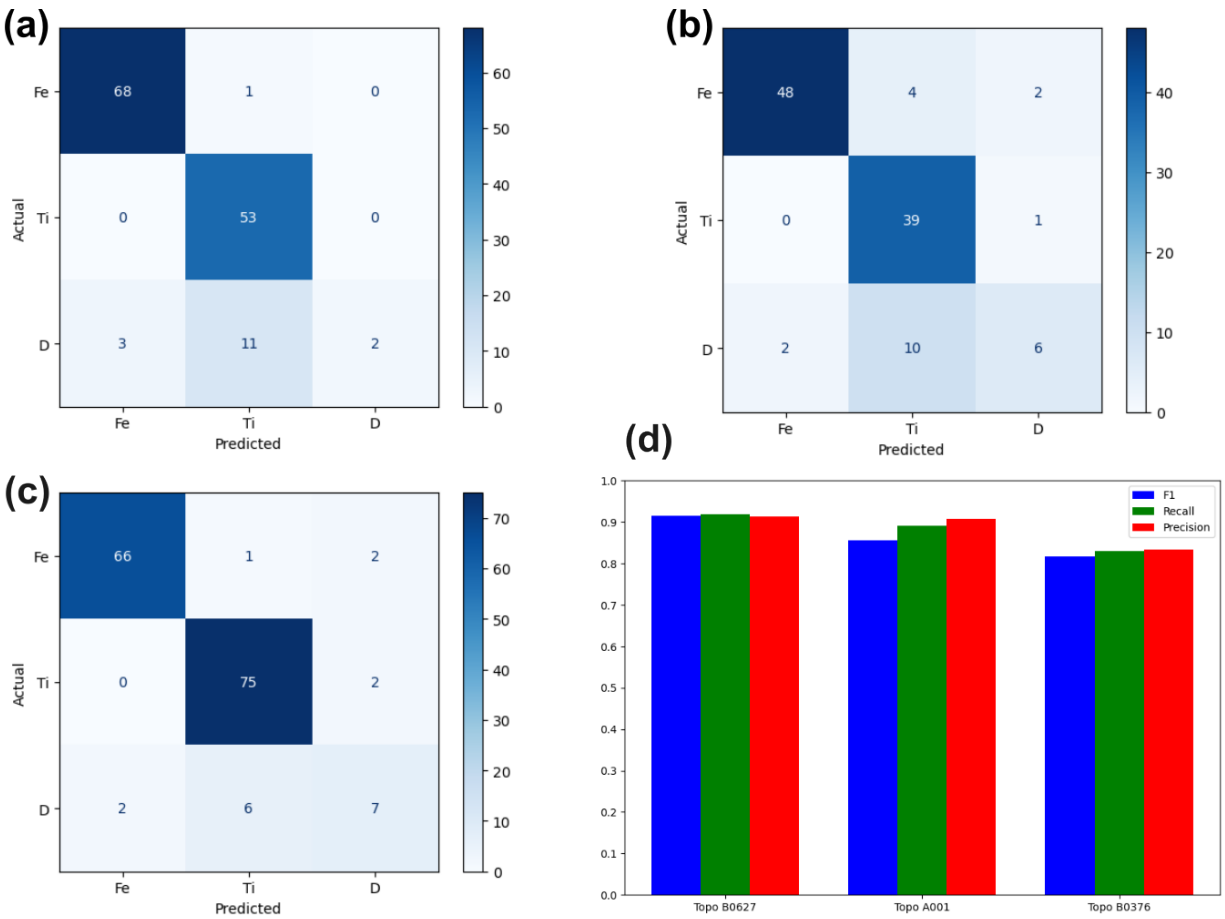


Figure 4.19: Confusion matrices of the GMM results with three classes. (a) Topo A001. (b) Topo B0376. (c) Topo B0627. (d) Precision, recall, and F1 scores of the three-class GMM classification for the three different topographies presented in this work.

4.3.1. Parameter Choices for the GMM Atom Species Classification

In this section, we justify the parameter choices made during the initial GMM classification. The decision to train the GMM on two classes and subsequently filter outliers, as seen in Section 4.1, is justified by looking at the results produced by training the GMM on three classes (Figure 4.19).

As we can see, the model emulates the two-class classification seen in Figure 4.4 by finding the true positive Fe and Ti atoms. However, adding a third component to classify defects and unknowns is not very useful as it struggles to properly classify them. This results in a significant misclassification of defects as Ti atoms. This is reflected in the lower scores achieved for recall, precision, and F1 (Figure 4.19d). Proceeding with the analysis to further refine the initial classification with subsequent classifications is a redundant task, as it only produces results similar to or worse than the two-class model.

4.3.2. Data Processing Choices for the GMM Atom Species Classification

Secondly,we justify avoiding the use of a Fourier transform for the GMM classification of atom species in Section 4.1 as we did for the class-conditioned GMM classifications in Section 4.2. As seen in Figure 4.20, transforming the data to perform the atom species classification produces a large number of misclassified Fe atoms for the three topographies, reducing all three of our classification scores.

However, if we proceed with this atom species classification on the Fourier transformed data, a subsequent PCA of the Ti-classified sites can separate the misclassified Fe atoms seen in Figure 4.20. The PCA is more useful in this context than a GMM in separating the Ti atoms from the rest because it produces a more descriptive distribution of the data points, as can be seen in Figure 4.21. We see that for the three topographies, a clear separation is made along the first principal component axis to distinguish

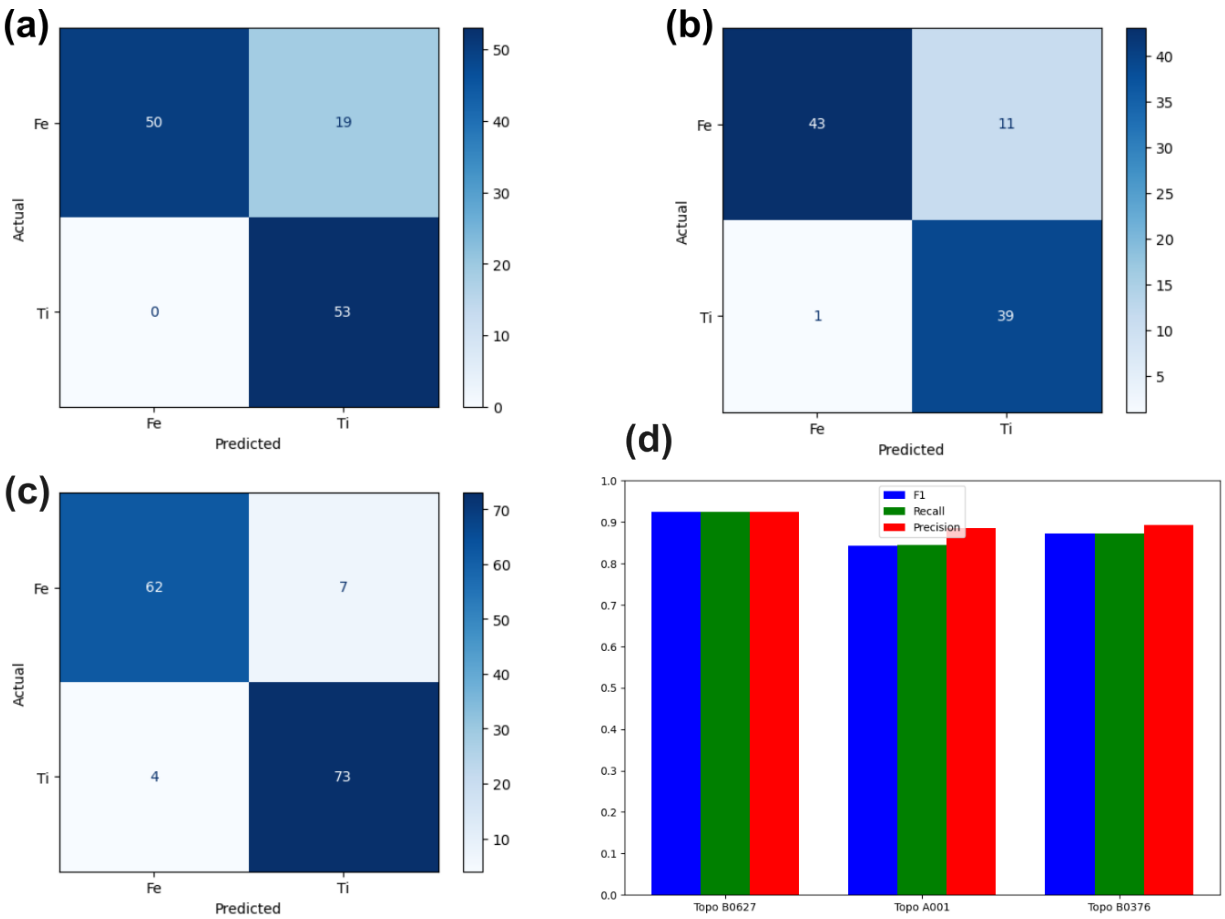


Figure 4.20: Confusion matrices of the GMM results on the Fourier transform moduli of our data, defects and unknowns omitted. (a) Topo A001. (b) Topo B0376. (c) Topo B0627. (d) Precision, recall, and F1 scores of the Fourier space GMM classification for the three different topographies presented in this work, omitting defects and unknowns.

Fe atoms, Ti atoms, and defects or unknowns. In the case of Topography B0627 (Figure 4.21c), we see that the PCA emulates our findings in Section 4.2.4 that a distinction is made along the first principal component axis between Ti atoms due to the low number of defects in this dataset.

Furthermore, by performing a density-based clustering analysis (Figure 4.21), we see that the Fe atoms and defects can form their own clusters away from the denser Ti cluster. Data points that do not belong to any cluster can be assessed by their computed density and proximity to a nearest cluster to distinguish their atom species. In the case of Topography B0627, in the cluster of separated Ti atoms from the main cluster, the computed densities are lower than in the other two images where outlier clusters also form. This once again shows the topographical similarity of these points with the main cluster of Ti atoms, indicating that they belong to the same species.

4.3.3. Data Processing Choices for the Class-Conditioned Subclassifications on Fe- and Ti-classified Sites

In this section, we justify the use of Fourier transforms to separate true positive Fe and Ti atoms from defects, unknowns, and misclassifications as seen in Section 4.2. Upon examining the results of the Fe-class-conditioned GMM classification without using a Fourier transform, we see that defects and unknowns are adequately separated (Figure 4.22). In the case of Topography B0376 (Figure 4.22b), the model even performs better with a smaller number of misclassified Fe atoms than in Figure 4.8b, where Fe atoms corresponding to dimers were misclassified.

However, examining the results of the Ti-class-conditioned GMM classification without using a Fourier

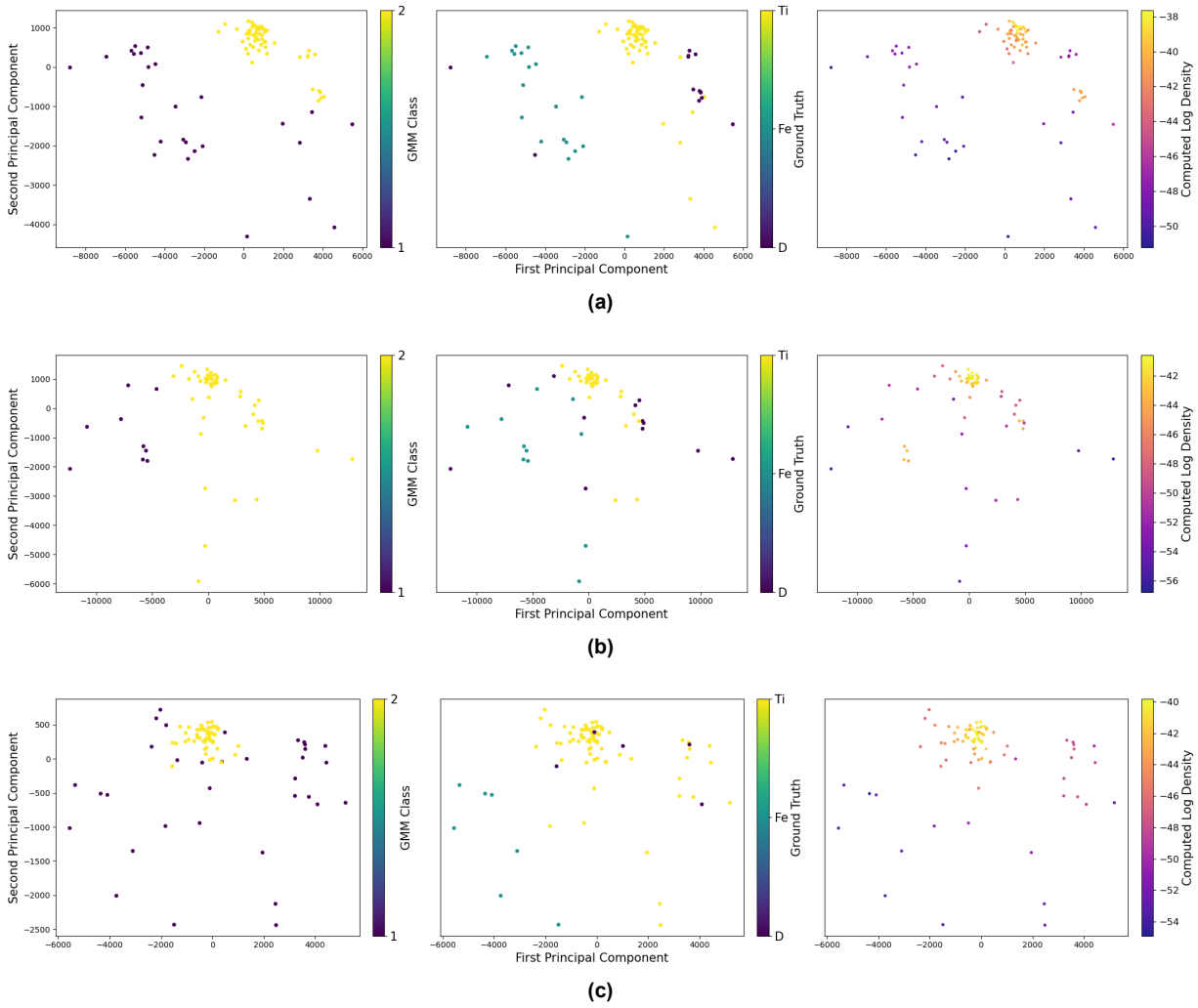


Figure 4.21: Performing a PCA and density-based clustering analysis on the data points classified as Ti-sites by a GMM trained on Fourier transformed data. (a) Topo A001. (b) Topo B0376. (c) Topo B0627.

transform (Figure 4.23), we notice substantial decreases in all our scores for all our metrics in Figure 4.23d compared to Figure 4.15d. As we saw in Section 4.2, the differences between Ti-classified data points are more intricate than for Fe-classified data, where they could be distinguished by the first PCA component. This was made clear by the on average lower explained variance for the first PCA component in the Ti case (Figures D.1, F.1) and the higher scattering of data points in the Ti PCA space (Figures 4.10, 4.17).

Without using Fourier transforms to represent the topographical differences between the Ti-classified data points, the number of misclassified sites is high enough to discredit the model. The shortcomings of the classification on Fourier transformed Ti-classified data points were explainable because they stemmed from atoms corresponding to dimers. In this case, the topographical features are indistinguishable and the classification is uninterpretable.

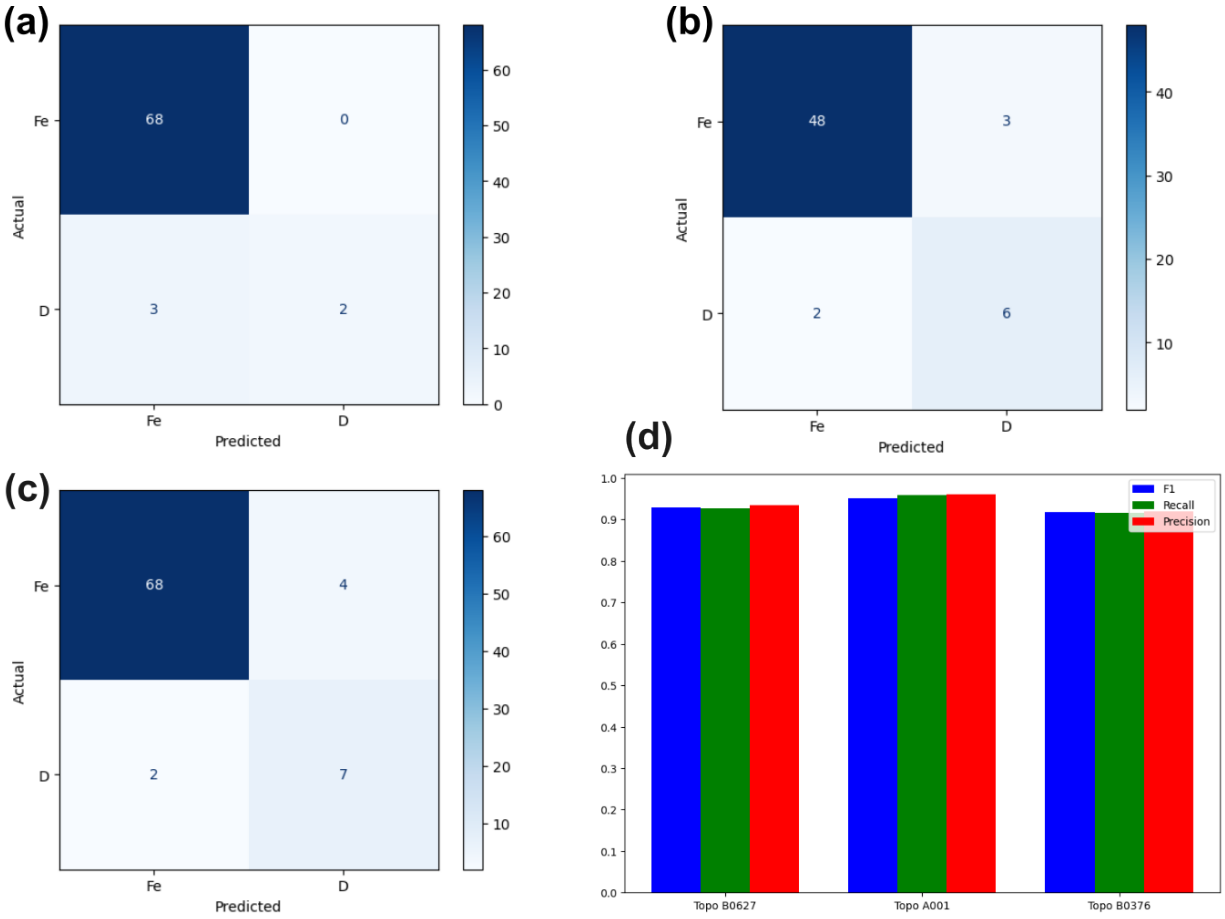


Figure 4.22: Confusion matrices of results from the Fe class-conditioned GMM classification without using Fourier transforms (a) Topo A001. (b) Topo B0376. (c) Topo B0627. (d) Precision, recall, and F1 scores for the Fe class-conditioned GMM classification without using Fourier transforms for the three different topographies presented in this work.

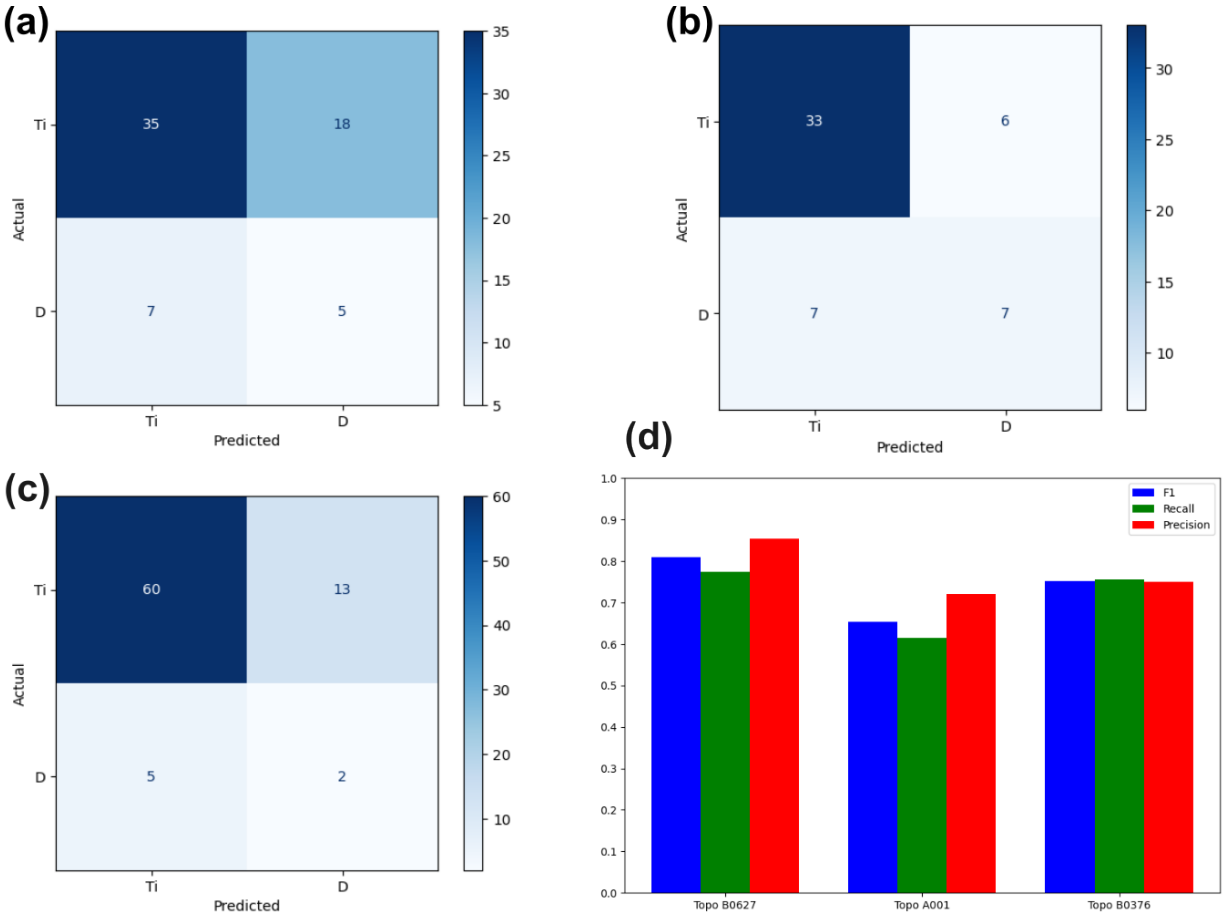


Figure 4.23: Confusion matrices of results from the Ti class-conditioned GMM classification without using Fourier transforms (a) Topo A001. (b) Topo B0376. (c) Topo B0627. (d) Precision, recall, and F1 scores for the Ti class-conditioned GMM classification without using Fourier transforms for the three different topographies presented in this work.

Part III

Closure

5

Conclusion

In this work, we analyzed STM images collected with different acquisition parameters of Ti and Fe adatoms on bilayer MgO islands grown on an Ag substrate. The purpose of the analysis was to detect and classify adatom sites on the imaged surface. We performed our analysis on eight different images, three of them being the focus of this work. The three samples were chosen to showcase a diversity of cases in terms of acquisition parameters, number of adatoms, and challenges. As we have seen, among these challenges are dimer formations and defects in the sample creating clear outliers.

Image processing techniques were used to detect adatom sites and prepare them for machine learning classification. First, we used contour detection and geometric analysis tools provided by the extensive OpenCV library to locate adatom and defect sites in the sample with their coordinates. Next, we built a dataset of these coordinate sites for each STM image by taking window crops centered at each coordinate site and processing them to extract their most pertinent features. This includes intensity clipping, background subtraction, and taking a circular crop around each coordinate site.

Once we had constructed our datasets for each STM image, we used unsupervised machine learning methods to classify coordinate sites by species and to quantify the topographical differences between them. The process began with an initial GMM classification to classify Fe and Ti atoms. A subsequent class-conditioned GMM classification was performed on the Fourier transform modulus of each data point to separate erroneous classifications from the true positive atoms in each class. These erroneous classifications included defects, unlabeled atom sites, or misclassifications. Since class-conditioned GMM classifications produced imperfect results, we proceeded with class-conditioned PCA and density-based clustering analyzes to quantify the topographical differences and similarities between our coordinate sites.

The initial GMM classification was programmed to produce clear distinctions between two atom classes: Fe and Ti. This distinction was successful with near-perfect F1, precision, and recall scores greater than 0.9 in classifying the Fe and Ti sites into two separate classes for seven out of eight topographies (Figure 4.4d). The Ti and Fe classes were distinguished on the basis of the intensity values in their atom images. We found that Ti images have a higher mean intensity, indicating that Ti atoms have a higher LDOS on bilayer MgO.

The subsequent class-conditioned GMM subclassifications required Fourier transforms to represent the topographical features (shapes, edges, orientations, intensity values) of our coordinate sites to successfully distinguish between true positive Fe or Ti sites and outliers. In this subclassification task, the model was tasked with classifying the class-conditioned data points into one of two classes: topographically similar atoms and outliers. The classification results became evident thanks to the standard deviation plots of each subclass. These indicated that the class-conditioned GMM subclass that contained true positive atoms had strong topographical similarities, while the data points in the outlier class exhibited much more variation in their topographies.

The main shortcomings of the class-conditioned GMM were its inability to distinguish between atom species in dimers and its mishandling of outliers - whether they resembled true positive atoms or skewed the classification due to how different they were. Despite these shortcomings, the class-conditioned GMM classifications still managed F1, precision, and recall scores greater than 0.8 in distinguishing between true positive Fe/Ti sites and defects, unknowns, or misclassifications in most cases (Figures 4.8d, 4.15d).

Topographies that did not score well simply had a greater number of dimer formations or unlabeled coordinate sites due to human errors during the acquisition process (Figure A.1b). This is reflected in the precision scores being higher than the recall scores, as the precision scores the model's ability not to mislabel a given data point, while the recall scores the model's ability to find all points within a class. Thus, high precision scores show that single atoms are being classified correctly.

In response to the class-conditioned GMM's misclassifications, we used a more quantifiable approach starting with a PCA (Figure 4.10). In the case of Fe-classified atoms, the PCA discovered that the two main components that explained the variation in the dataset were the topographical features found at the central frequencies with the highest modulus values in the Fourier transforms of the atom images and the presence of a near-neighbor with the former far more dominant (Figures D.1, D.2).

Since essential topographical features lie at the center of the Fourier transform, the first principal component managed to group misclassified Fe atoms in the same range as correctly classified Fe atoms. The second principal component showed that these misclassified atoms corresponded to dimers. It should also be noted that the coordinate sites classified as topographically similar by the GMM formed a dense cluster of true positive Fe sites in PCA space (Figure 4.10). The density-based clustering approach in this case allowed us to score each data point's resemblance to this cluster of Fe points. By projecting the clustering results onto our data, we obtained a visual assessment of the likelihood that a coordinate site represents a Fe adatom (Figure 4.11).

In the case of Ti-classified atoms, the PCA discovered that the first principal component that explained the variation in the dataset were the topographical features found both at the central and in the surrounding frequencies in Fourier space. However, the second component was less interpretable, but more pertinent than in the Fe case with a higher explained variance (Figures F.1, F.2). In this case, PCA produced a main cluster of points that represented the true positive single Ti atoms, which were also properly classified by the GMM (Figure 4.17).

Furthermore, a smaller clusters of outliers were formed that contained both defects/unknowns and false negative Ti atoms. Unfortunately, PCA was inconclusive in identifying Ti atoms corresponding to dimers apart from the other outlier. Unlike the Fe case, where we could for the most part find Fe atoms corresponding to dimers on the same range in the first principal component as single atoms of the same species, the Ti data points were much more scattered in PCA space (Figures 4.10, 4.17). In addition, the first PCA component had a lower explained variance for the Ti atom dataset than the Fe atom dataset (Figures D.1, F.1). This led us to believe that the Ti dataset is topographically more complex. Thus, we required an approach that computes a single value based on the global representation of our data to quantify its topographical similarities.

Therefore, we used the density-based clustering algorithm to assess the topographies of our Ti atoms (Figure 4.17). We found that the false negative Ti atoms could be detected by assessing their computed density, their distance to the nearest cluster, and their relative cluster center. This meant that low-density points closer to the true positive cluster were more likely to be defects/unknowns, whereas low-density points closer to the outlier cluster were more likely to be false negative Ti atoms mostly corresponding to dimers. This indicated that despite their proximity to the outlier cluster, their densities were computed relative to the main cluster of true positive Ti atoms. Thus, they shared a higher topographical similarity with atoms of the same species.

6

Recommendations

Further steps can be taken to validate the general applicability of our model by deploying it on a variety of STM imaged systems. In addition, combining physical insights with the findings of this model could facilitate tasks such as lattice fitting. For instance, knowing that Fe sites only deposit on O sites [5], we can utilize the Fe site detection and classification in this work to develop a lattice fitting algorithm for TM atoms adsorbed on bilayer MgO islands. We may also explore other unsupervised methods for the classification of different atom species in STM images. This includes using VAE with different loss functions [16], or using transfer learning [31] to overcome the small dataset bottleneck encountered for VAE. In this case, we would transfer the learned latent variables from training on simulated STM images to improve the classification in our target task.

References

[1] H. Greenside. *Modern physics, an introduction to special relativity and quantum mechanics*. 2015. URL: https://webhome.phy.duke.edu/~hsg/264L/.

[2] B. Chen et al. “Spintronic devices for high-density memory and neuromorphic computing – A Review”. In: *Materials Today* 70 (Nov. 2023), pp. 193–217. DOI: 10.1016/j.mattod.2023.10.004.

[3] L. Bogani et al. “Molecular spintronics using single-molecule magnets”. In: *Nature Materials* 7 (Mar. 2008), pp. 179–186. DOI: 10.1038/nmat2133.

[4] K. Yang et al. “Coherent spin manipulation of individual atoms on a surface”. In: *Science* 366 (Oct. 2019), pp. 509–512. DOI: 10.1126/science.aay6779.

[5] E. Fernandes et al. “Adsorption sites of individual metal atoms on ultrathin MgO(100) films”. In: *Physical Review B* 96 (July 2017), p. 045419. DOI: 10.1103/physrevb.96.045419.

[6] S. Loth et al. “Spin-polarized spin excitation spectroscopy”. In: *New Journal of Physics* 12 (Dec. 2010), p. 125021. DOI: 10.1088/1367-2630/12/12/125021.

[7] C. Hübner et al. “Symmetry effects on the spin switching of adatoms”. In: *Physical Review B* 90 (Oct. 2014), p. 155134. DOI: 10.1103/physrevb.90.155134.

[8] K. Rossi et al. “Quantitative description of metal center organization and interactions in single atom catalysts”. In: *Advanced Materials* 36 (Dec. 2023), p. 2307991. DOI: 10.1002/adma.202307991.

[9] A. Lafleur et al. *Automated classification of individual atoms on surfaces using machine learning*. Oct. 2024. DOI: 10.48550/arXiv.2410.13711.

[10] C. J. Chen. *Introduction to scanning tunneling microscopy*. Oxford University Press, 1993.

[11] J. Hwang et al. “Development of a scanning tunneling microscope for variable temperature electron spin resonance”. In: *Review of Scientific Instruments* 93 (Sept. 2022), p. 093703. DOI: 10.1063/5.0096081.

[12] S. Sicolo et al. “Adsorption of late transition metal atoms on MgO/Mo(100) and MgO/Ag(100) ultrathin films: A comparative DFT study”. In: *The Journal of Physical Chemistry C* 113 (Sept. 2009), pp. 16694–16701. DOI: 10.1021/jp905592c.

[13] A. J. Heinrich et al. “Single-atom spin-flip spectroscopy”. In: *Science* 306 (Oct. 2004), pp. 466–469. DOI: 10.1126/science.1101077.

[14] M. Valleti et al. “Physics and chemistry from parsimonious representations: Image analysis via invariant variational autoencoders”. In: *npj Computational Materials* 10 (Aug. 2024). DOI: 10.1038/s41524-024-01250-5.

[15] M. Ziatdinov et al. “AtomAI framework for deep learning analysis of image and spectroscopy data in electron and scanning probe microscopy”. In: *Nature Machine Intelligence* 4 (Dec. 2022), pp. 1101–1112. DOI: 10.1038/s42256-022-00555-8.

[16] A. Biswas et al. “Optimizing training trajectories in variational autoencoders via latent Bayesian optimization approach”. In: *Machine Learning: Science and Technology* 4 (Feb. 2023), p. 015011. DOI: 10.1088/2632-2153/acb316.

[17] N. Creange et al. “Towards automating structural discovery in scanning transmission electron microscopy”. In: *Microscopy and Microanalysis* 27 (July 2021), pp. 2770–2772. DOI: 10.1017/s1431927621009727.

[18] C. Becker. “UHV surface preparation methods”. In: *Encyclopedia of Interfacial Chemistry* (2018), pp. 580–590. DOI: 10.1016/b978-0-12-409547-2.11050-9.

[19] S. Prakash et al. “Superconducting films grown by activated reactive evaporation for high frequency device applications”. In: *Superconductor Science and Technology* 3 (Nov. 1990), pp. 543–545. DOI: 10.1088/0953-2048/3/11/005.

[20] Z. Wang et al. “Electron beam evaporation deposition”. In: *Advanced Nano Deposition Methods* (Sept. 2016), pp. 33–58. DOI: 10.1002/9783527696406.ch2.

[21] D. Nečas et al. “Gwyddion: An open-source software for SPM data analysis”. In: *Open Physics* 10 (Dec. 2011), pp. 181–188. DOI: 10.2478/s11534-011-0096-2.

[22] J. Ge et al. “Achieving low noise in scanning tunneling spectroscopy”. In: *Review of Scientific Instruments* 90 (Oct. 2019), p. 101401. DOI: 10.1063/1.5111989.

[23] G. Bradski. “The OpenCV library”. In: *Dr. Dobb’s Journal of Software Tools* 25 (Nov. 2000), pp. 120–125.

[24] F. Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830. DOI: 10.5555/1953048.2078195.

[25] D. Reynolds. “Gaussian mixture models”. In: *Encyclopedia of Biometrics*. Boston, MA: Springer US, 2009, pp. 659–663. DOI: 10.1007/978-0-387-73003-5_196.

[26] X. Jin et al. “K-means clustering”. In: *Encyclopedia of Machine Learning* (2011), pp. 563–564. DOI: 10.1007/978-0-387-30164-8_425.

[27] J. M. Brayer. *Introduction to Fourier transforms for image processing*. URL: https://www.cs.unm.edu/~brayer/vision/fourier.html.

[28] A. Edelman. *The singular value decomposition (SVD)*. 2016. URL: https://math.mit.edu/classes/18.095/2016IAP/lec2/SVD_Notes.pdf.

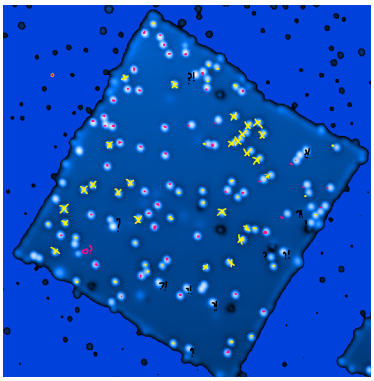
[29] A. Glielmo et al. “DADApy: Distance-based analysis of data-manifolds in Python”. In: *Patterns* 3 (Oct. 2022), p. 100589. DOI: https://doi.org/10.1016/j.patter.2022.100589.

[30] A. Rodriguez et al. “Clustering by fast search and find of density peaks”. In: *Science* 344 (June 2014), pp. 1492–1496. DOI: 10.1126/science.1242072.

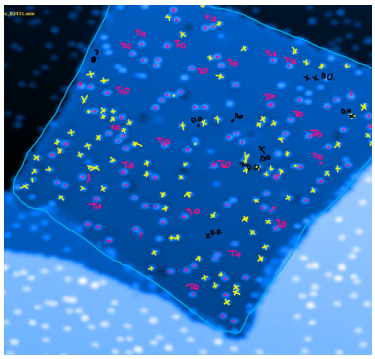
[31] L. Bonheme et al. *How good are variational autoencoders at transfer learning?* Apr. 2023. DOI: 10.48550/arXiv.2304.10767.

A

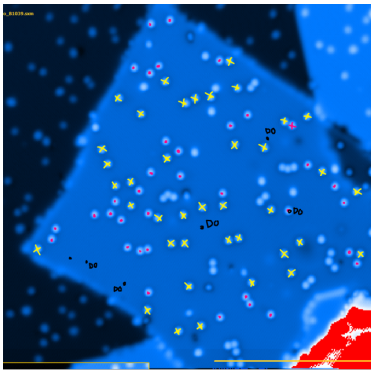
Supplementary Materials A: Experimental Methods



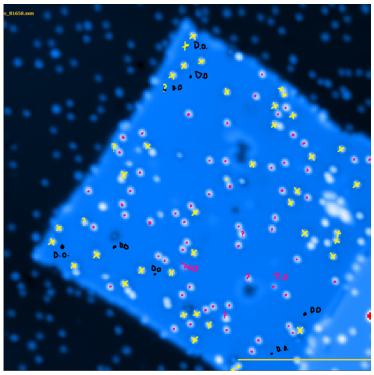
(a) Topo A228.



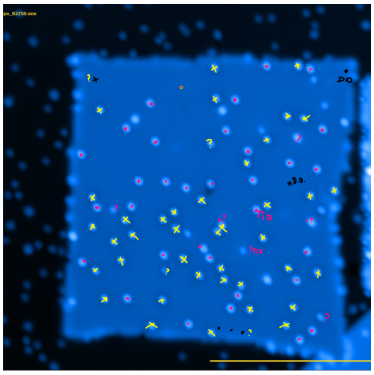
(b) Topo B2060.



(c) Topo B0917.

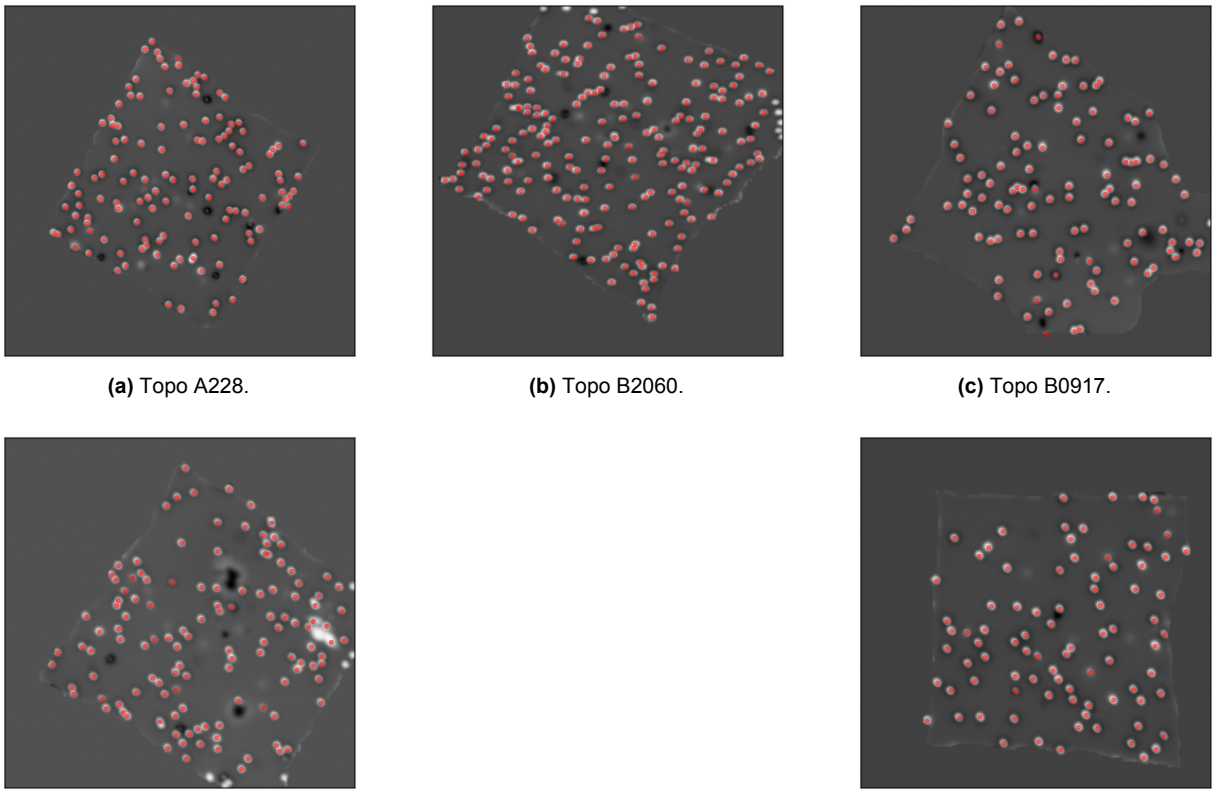


(d) Topo B1544.



(e) Topo B2731.

Figure A.1: Expert annotated ground truth labels. Titanium sites are marked in pink, iron sites are marked in yellow. Damages to the sample are marked in black, which result in unreliable measurements to the surrounding atoms, leading to unlabeled atom sites. Some images have further annotations meant for the lab personnel. These annotations were digitized to score the GMM classification.



(a) Topo A228. **(b)** Topo B2060. **(c)** Topo B0917.

(d) Topo B1544. **(e)** Topo B2731.

Figure A.2: Results of our atom detection workflow on five different topographies. The red dots correspond to detected atom coordinate sites. We see that some atoms on the edge of the bilayer MgO island are not detected.

B

Supplementary Materials B: Classifying Atoms by Species

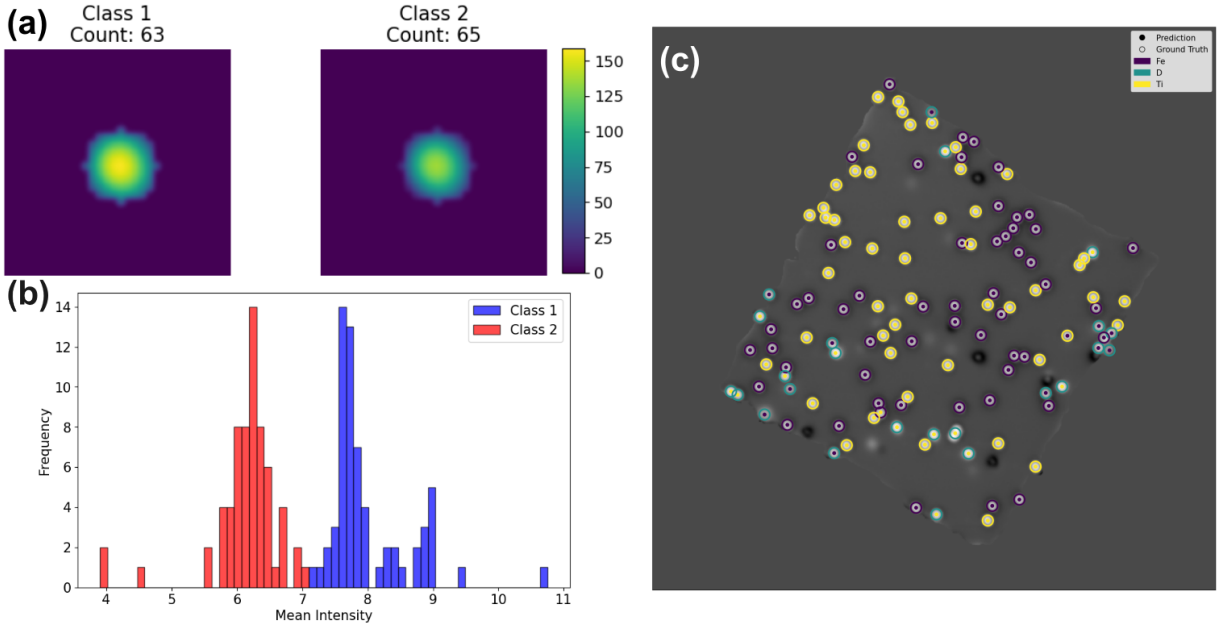


Figure B.1: Topo A228 first GMM analysis. (a) Mean intensity image within each learned class. (b) Mean intensity distribution by class. (c) Plot of the GMM results with an overlay of the ground truth labels.

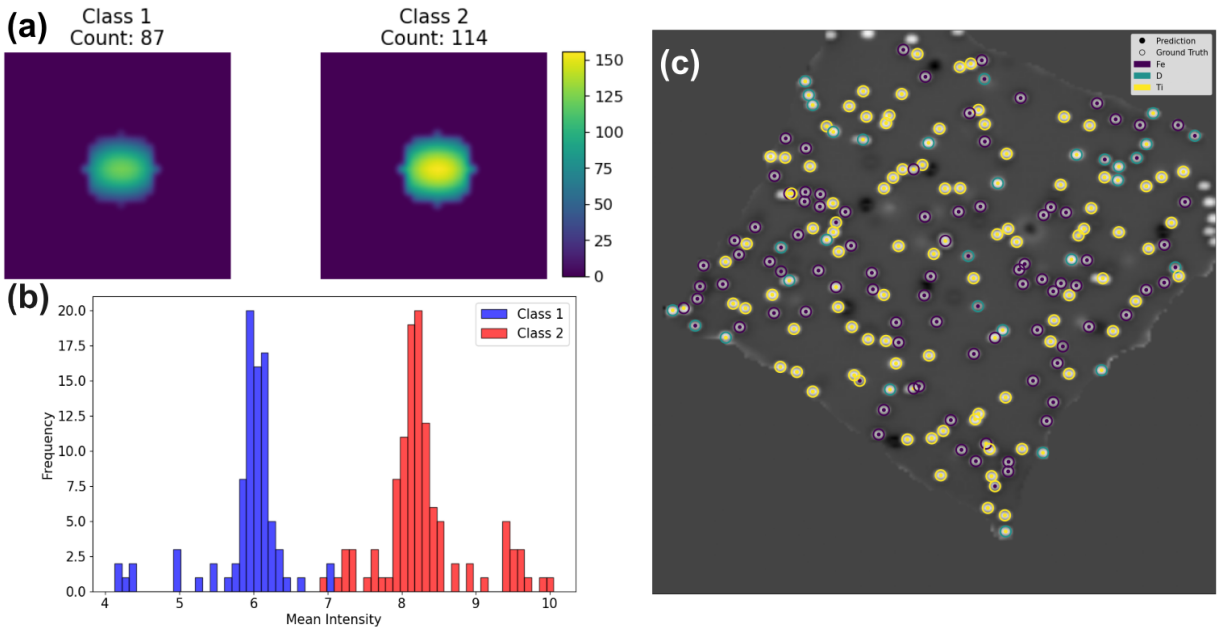


Figure B.2: Topo B2060 first GMM analysis. (a) Mean intensity image within each learned class. (b) Mean intensity distribution by class. (c) Plot of the GMM results with an overlay of the ground truth labels.

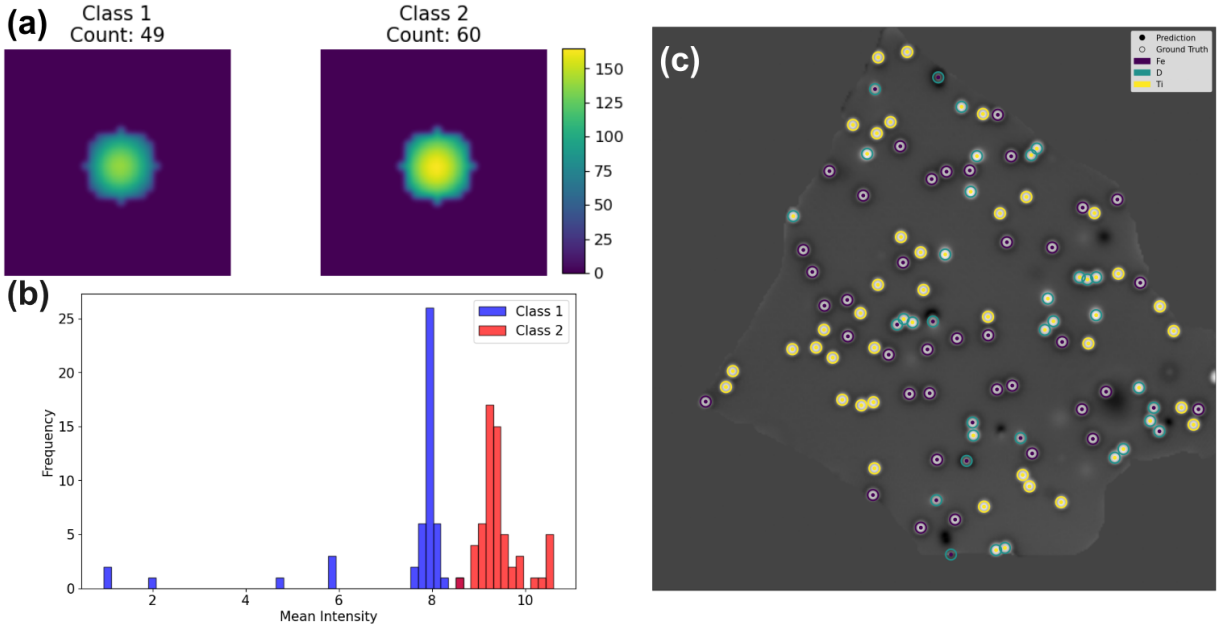


Figure B.3: Topo B0917 first GMM analysis. (a) Mean intensity image within each learned class. (b) Mean intensity distribution by class. (c) Plot of the GMM results with an overlay of the ground truth labels.

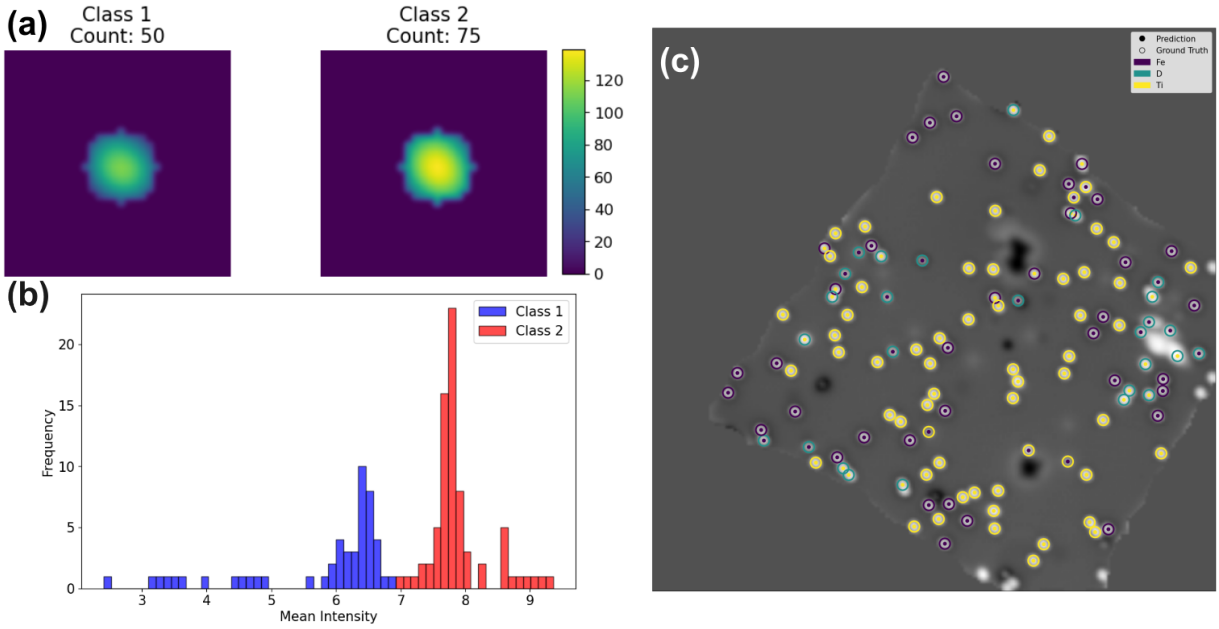


Figure B.4: Topo B1544 first GMM analysis. (a) Mean intensity image within each learned class. (b) Mean intensity distribution by class. (c) Plot of the GMM results with an overlay of the ground truth labels.

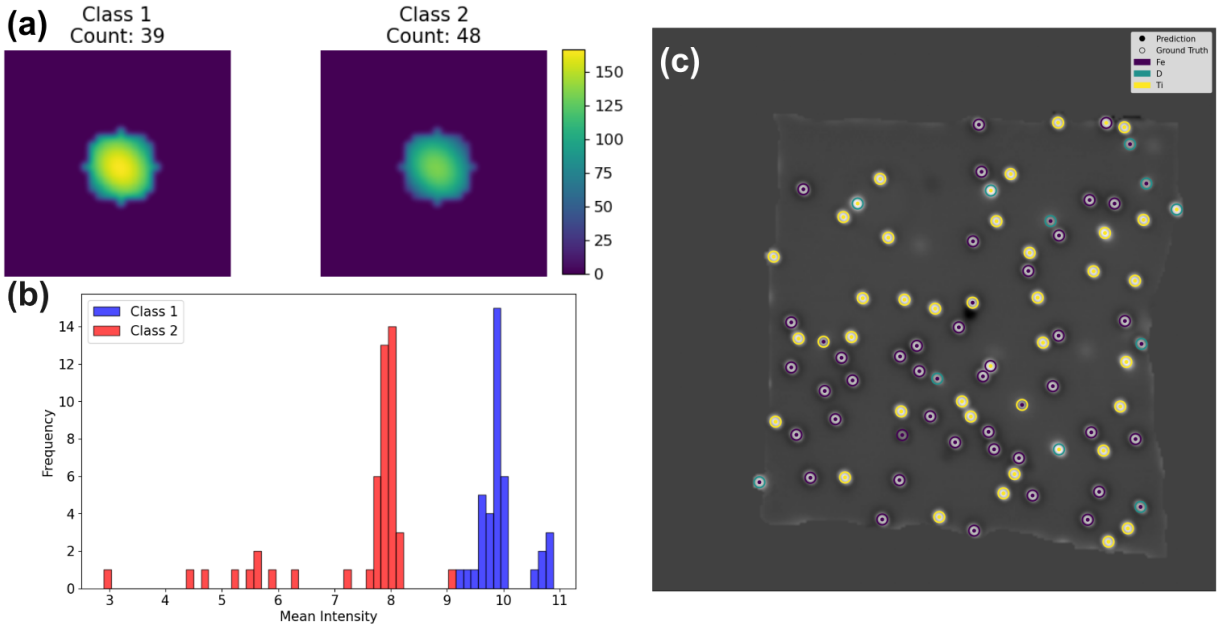
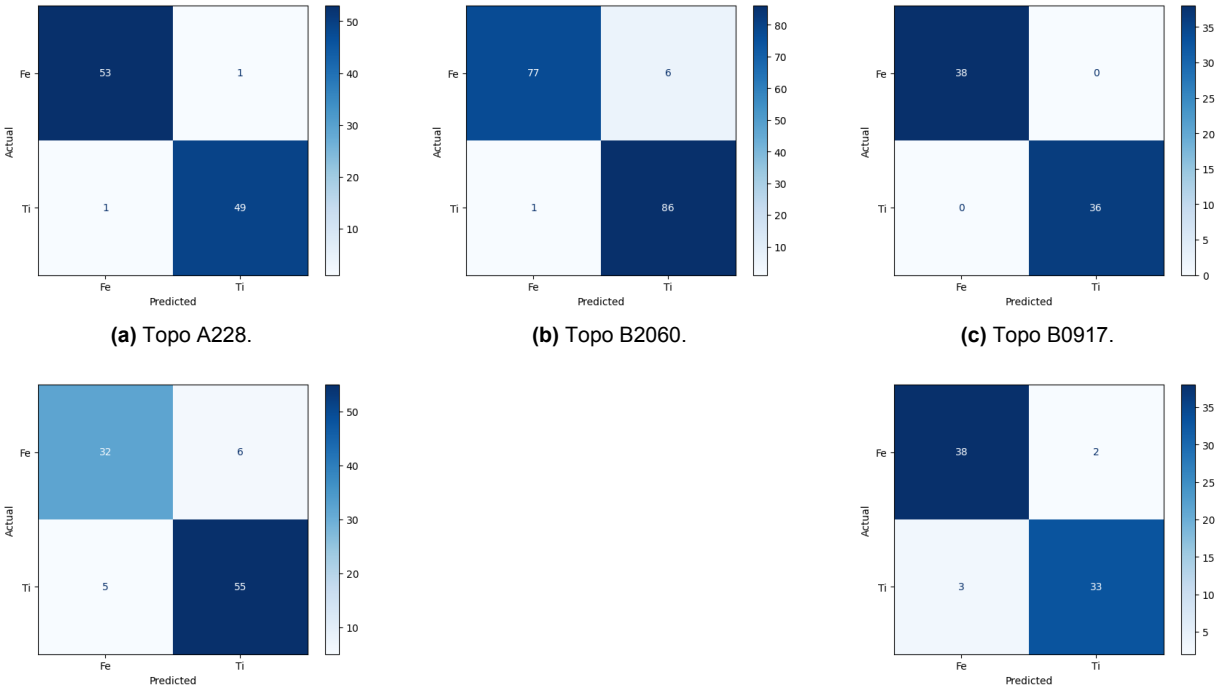


Figure B.5: Topo B2731 first GMM analysis. (a) Mean intensity image within each learned class. (b) Mean intensity distribution by class. (c) Plot of the GMM results with an overlay of the ground truth labels.



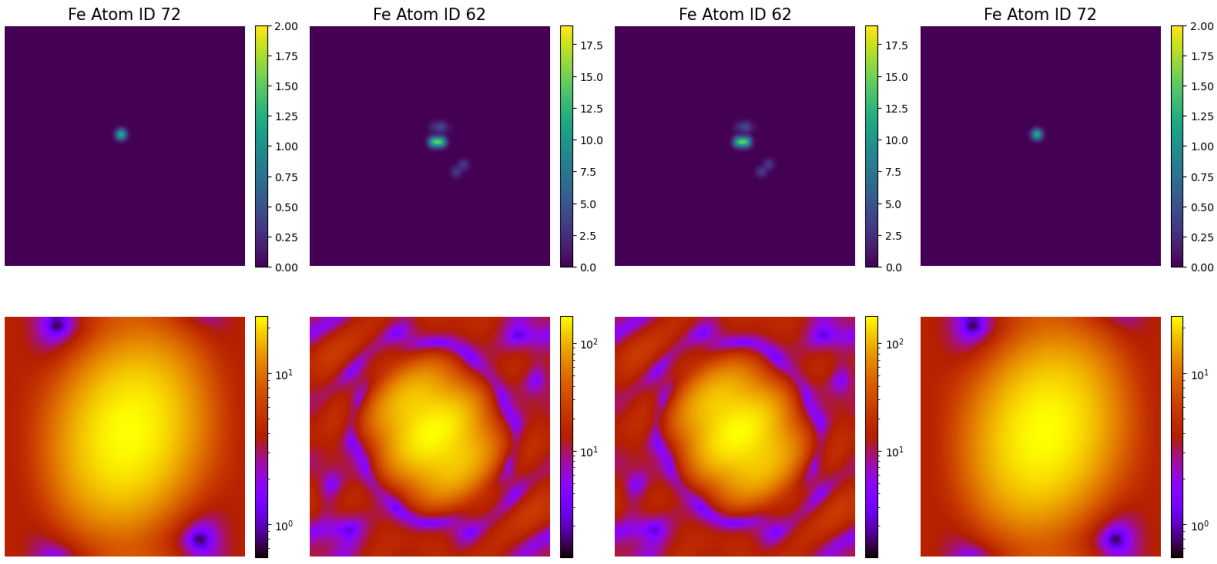
(a) Topo A228. **(b)** Topo B2060. **(c)** Topo B0917.

(d) Topo B1544. **(e)** Topo B2731.

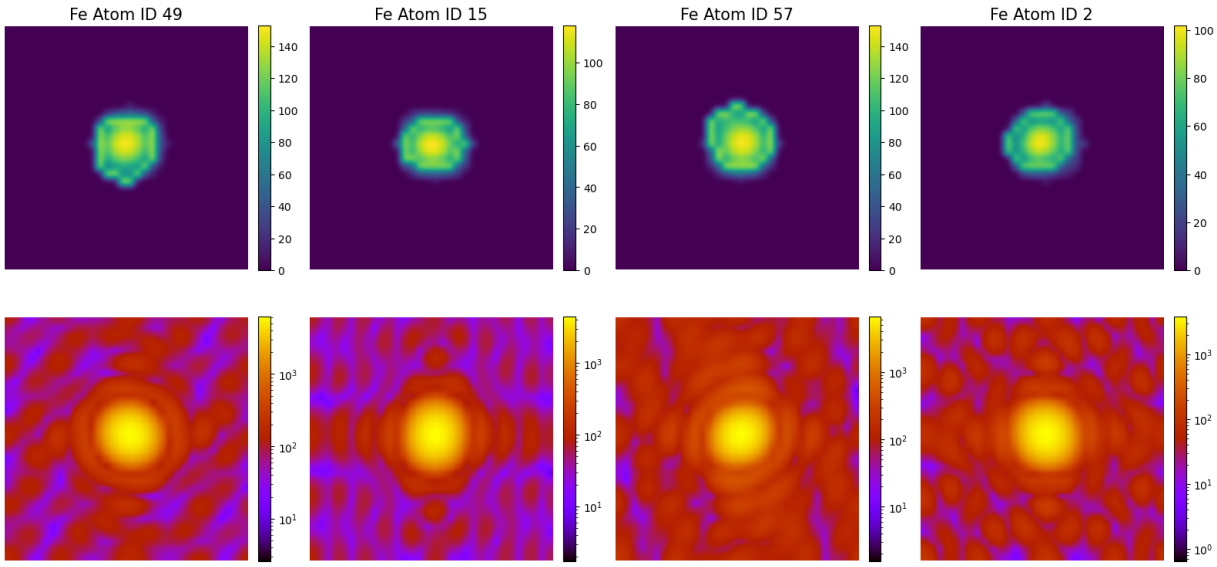
Figure B.6: Confusion matrices of the GMM results, defects and unknowns omitted.

C

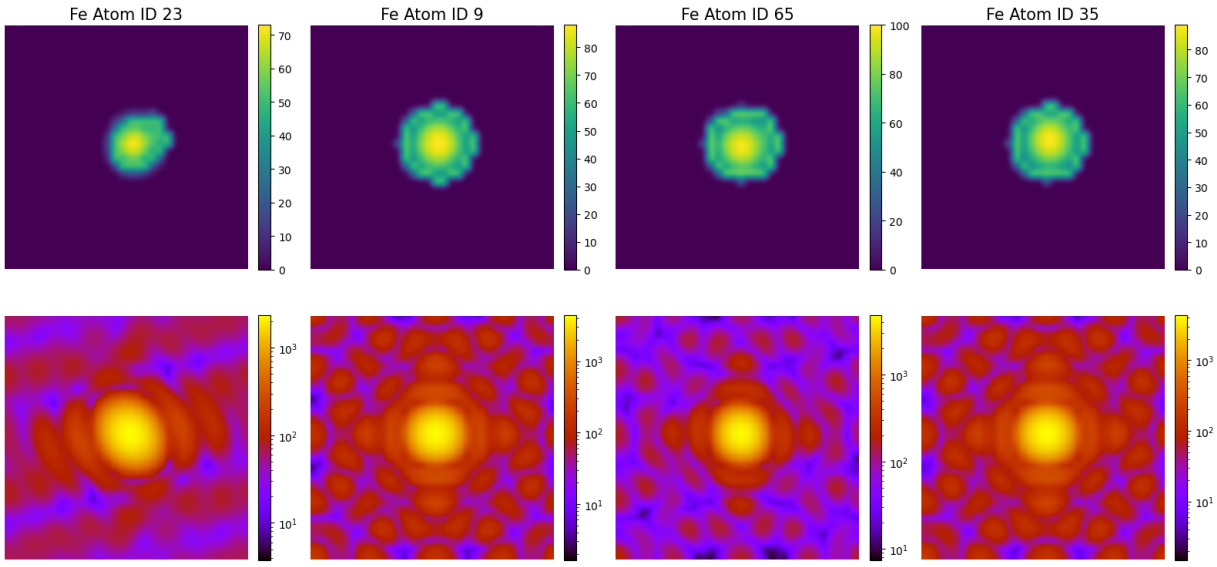
Supplementary Materials C: Class-conditioned Subclassification on Fe-classified Sites



(a)

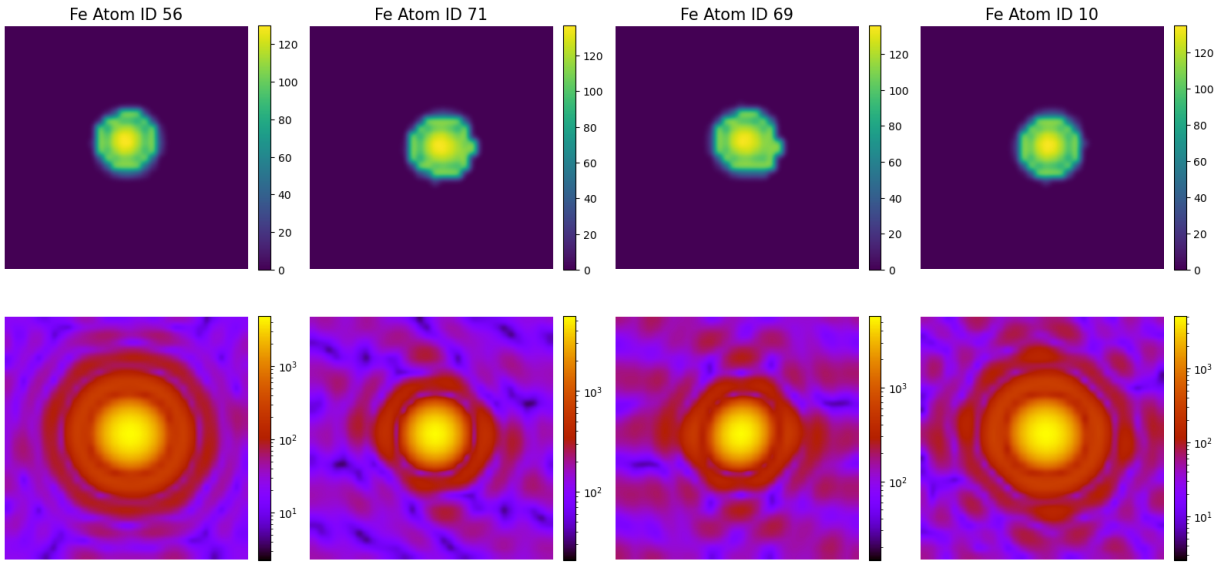


(b)

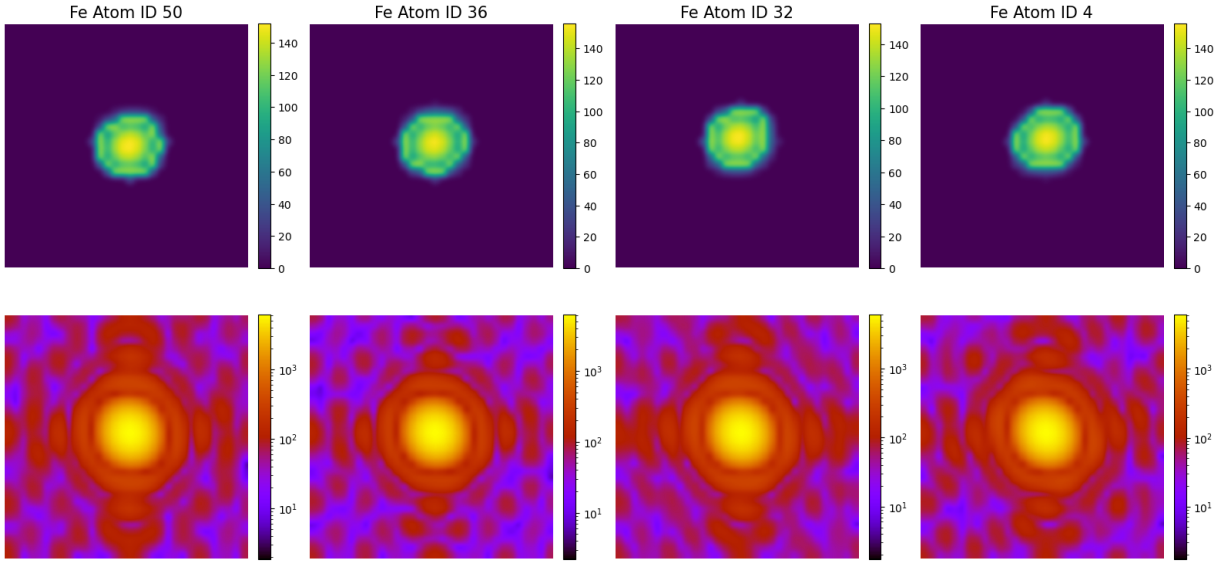


(c)

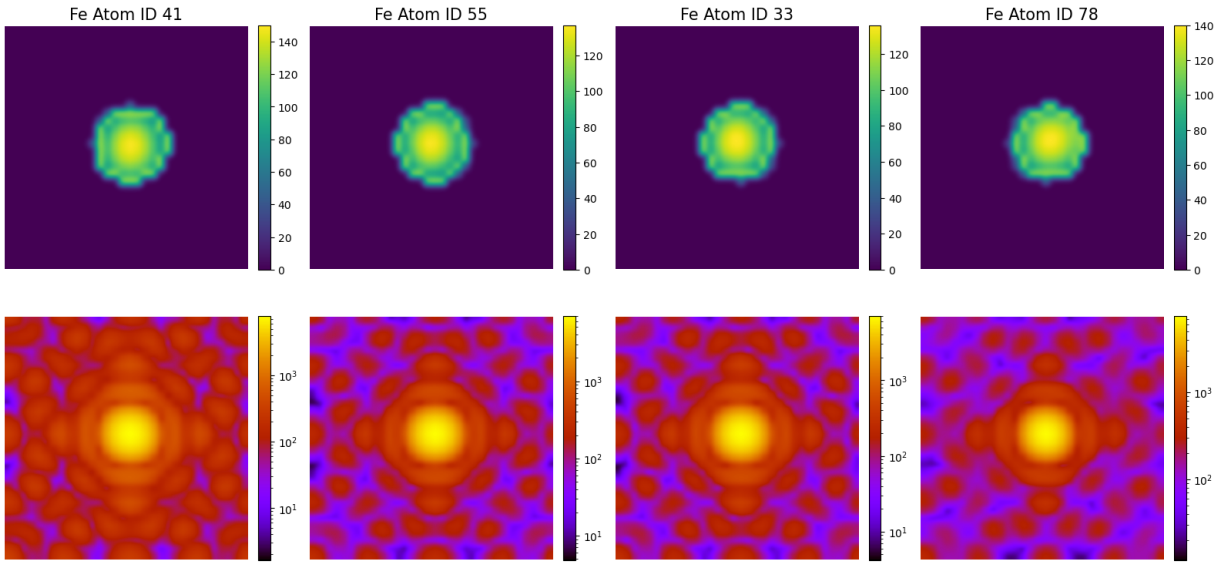
Figure C.1: Randomly sampled data points from the Fe class-conditioned GMM classification and their Fourier transforms, class 1 (outliers). We draw the contour (bright green) around each atom site to extract its shape and orientation. (a) Topo A001. (b) Topo B0376. (c) Topo B20627.



(a)



(b)



(c)

Figure C.2: Randomly sampled data points from the Fe class-conditioned GMM classification and their Fourier transforms, class 2 (Fe atoms). We draw the contour (bright green) around each atom site to extract its shape and orientation. (a) Topo A001. (b) Topo B0376. (c) Topo B20627.

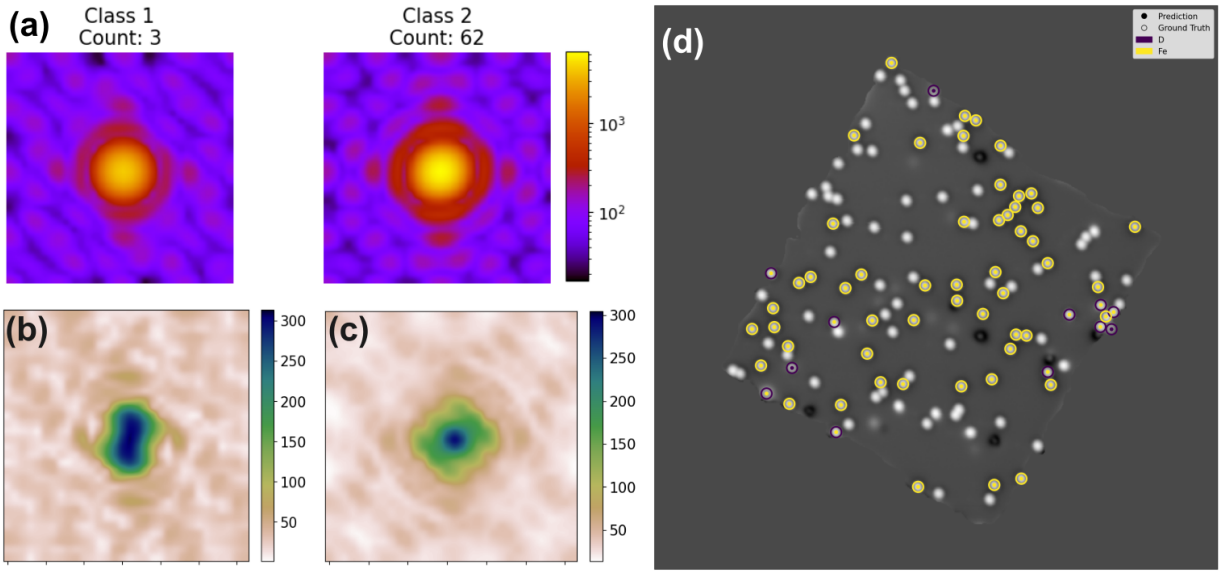


Figure C.3: Topo A228 class-conditioned GMM analysis on Fourier transformed Fe-classified sites. (a) Mean Fourier modulus of the atom-cropped image within each learned class. (b and c) Standard deviation in Fourier space within each learned class. (b) Class 1 (outliers). (c) Class 2 (Fe atoms). The outliers in this case skew the classification. (d) Plot of the GMM results with an overlay of the ground truth labels.

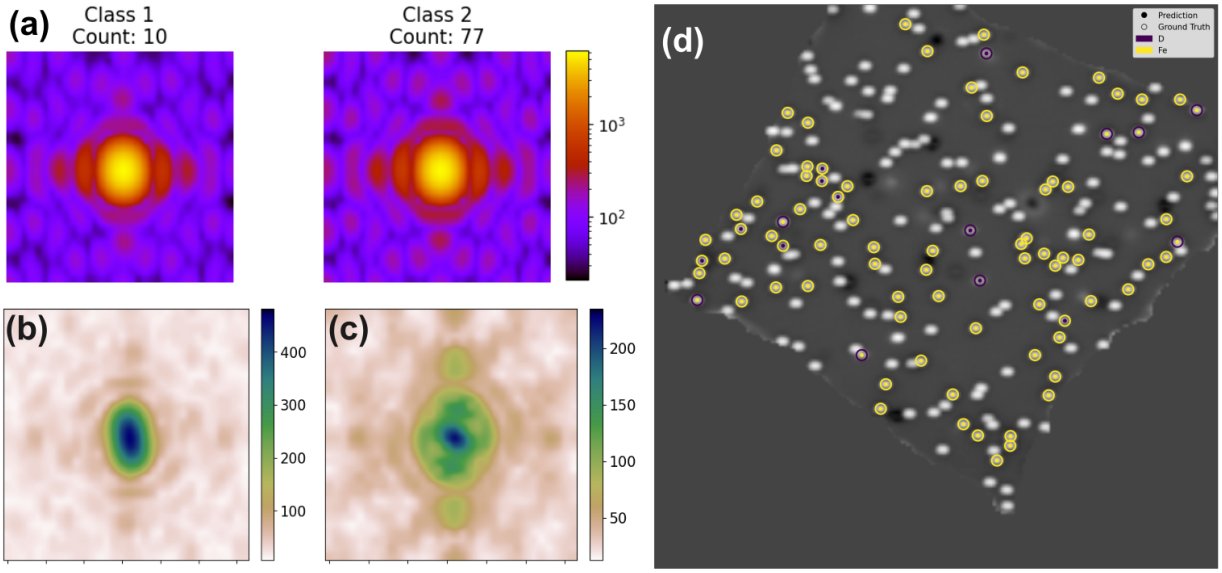


Figure C.4: Topo B2060 class-conditioned GMM analysis on Fourier transformed Fe-classified sites. (a) Mean Fourier modulus of the atom-cropped image within each learned class. (b and c) Standard deviation in Fourier space within each learned class. (b) Class 1 (outliers). (c) Class 2 (Fe atoms). Notice that the value ranges are widely different for each class. (d) Plot of the GMM results with an overlay of the ground truth labels.

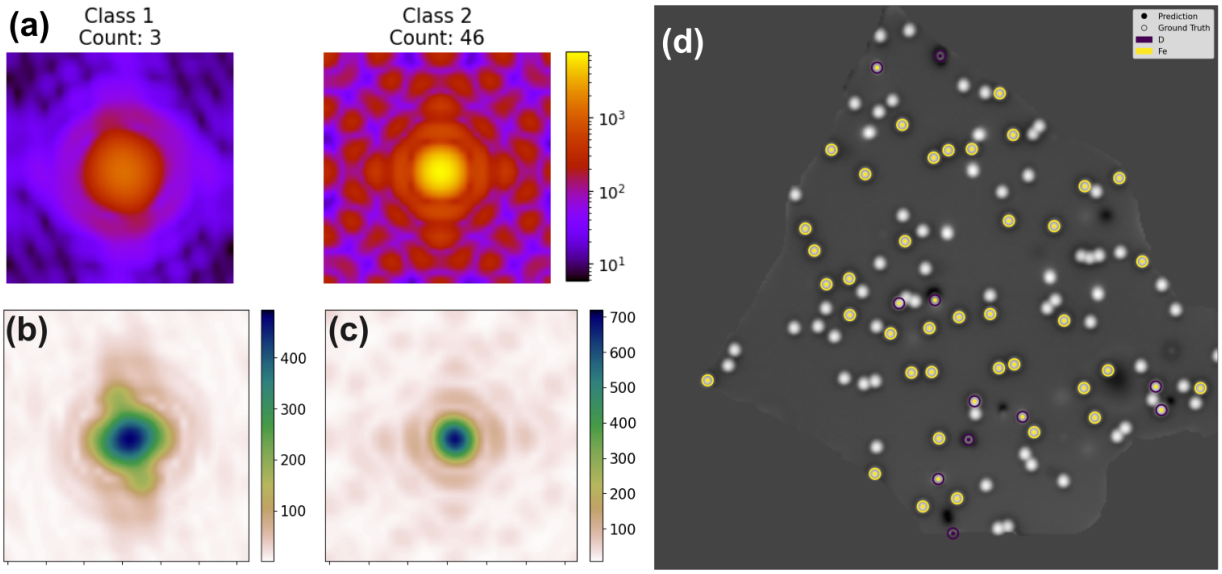


Figure C.5: Topo B0917 class-conditioned GMM analysis on Fourier transformed Fe-classified sites. (a) Mean Fourier modulus of the atom-cropped image within each learned class. (b and c) Standard deviation in Fourier space within each learned class. (b) Class 1 (outliers). (c) Class 2 (Fe atoms). The outliers in this case skew the classification. (d) Plot of the GMM results with an overlay of the ground truth labels.

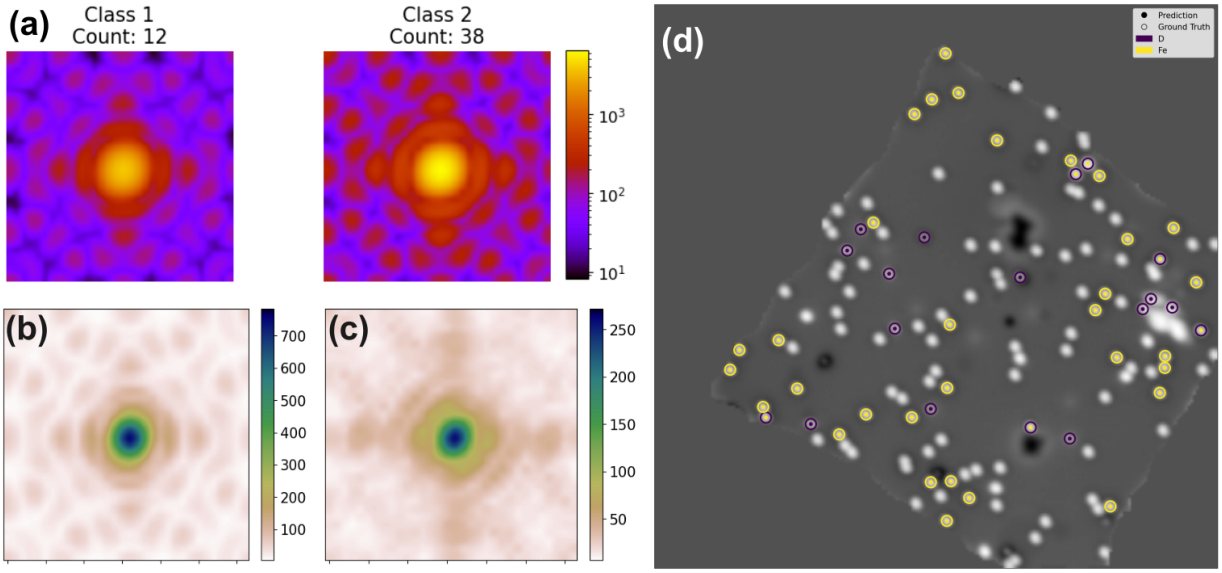


Figure C.6: Topo B1544 class-conditioned GMM analysis on Fourier transformed Fe-classified sites. (a) Mean Fourier modulus of the atom-cropped image within each learned class. (b and c) Standard deviation in Fourier space within each learned class. (b) Class 1 (outliers). (c) Class 2 (Fe atoms). Notice that the value ranges are widely different for each class. (d) Plot of the GMM results with an overlay of the ground truth labels.

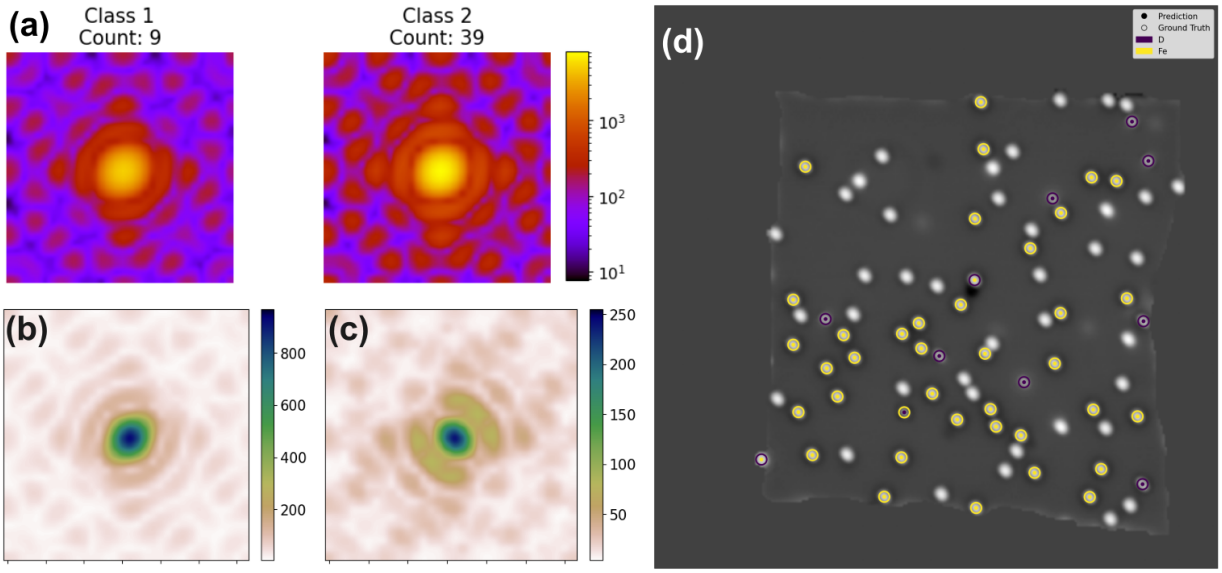
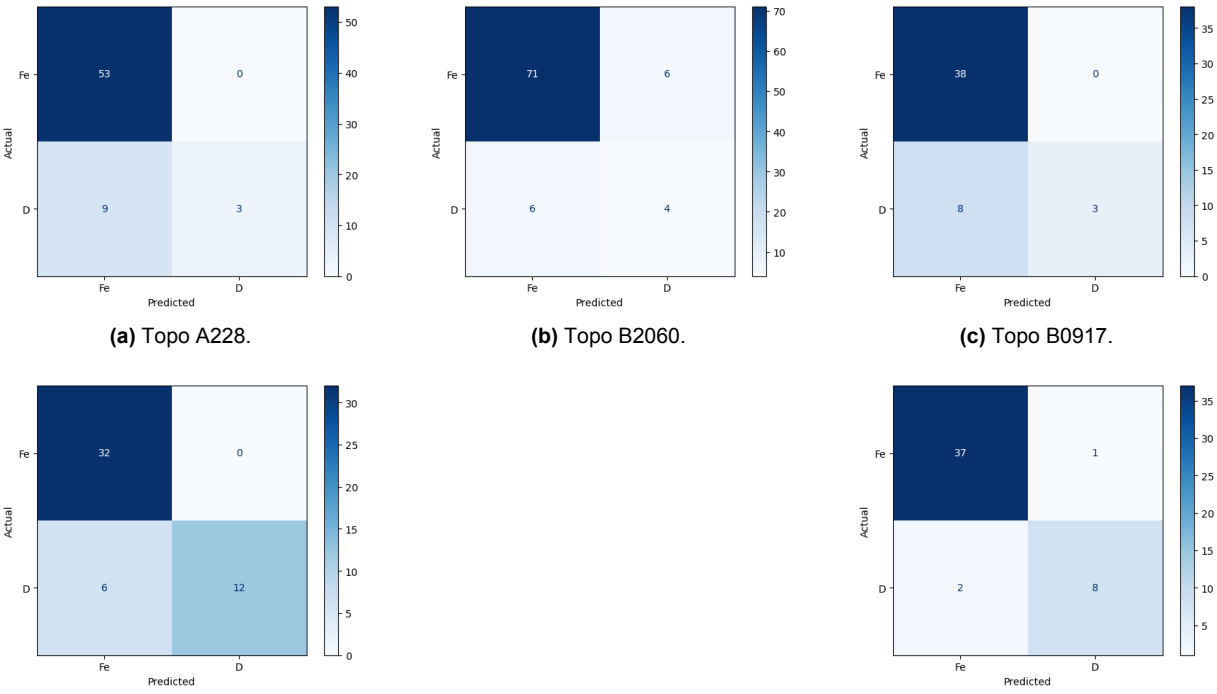


Figure C.7: Topo B2731 class-conditioned GMM analysis on Fourier transformed Fe-classified sites. (a) Mean Fourier modulus of the atom-cropped image within each learned class. (b and c) Standard deviation in Fourier space within each learned class. (b) Class 1 (outliers). (c) Class 2 (Fe atoms). Notice that the value ranges are widely different for each class. (d) Plot of the GMM results with an overlay of the ground truth labels.



(a) Topo A228.

(b) Topo B2060.

(c) Topo B0917.

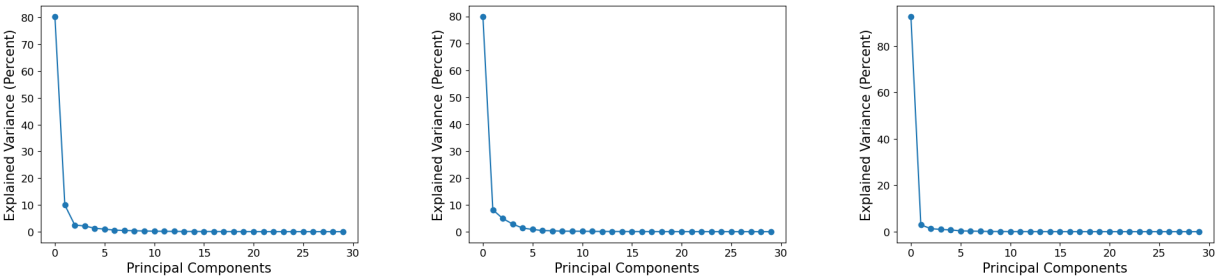
(d) Topo B1544.

(e) Topo B2731.

Figure C.8: Confusion matrices of results from the Fe class-conditioned GMM classification of Fourier transforms.

D

Supplementary Materials D: Quantifying the Variation in Fe-classified Sites



(a) Topo A001. **(b)** Topo B0376. **(c)** Topo B0627.

Figure D.1: Plotting the explained variance by principal component against the principal components learned from the Fourier transformed Fe-classified data. (a) Topo A001. (b) Topo B0376. (c) Topo B0627.

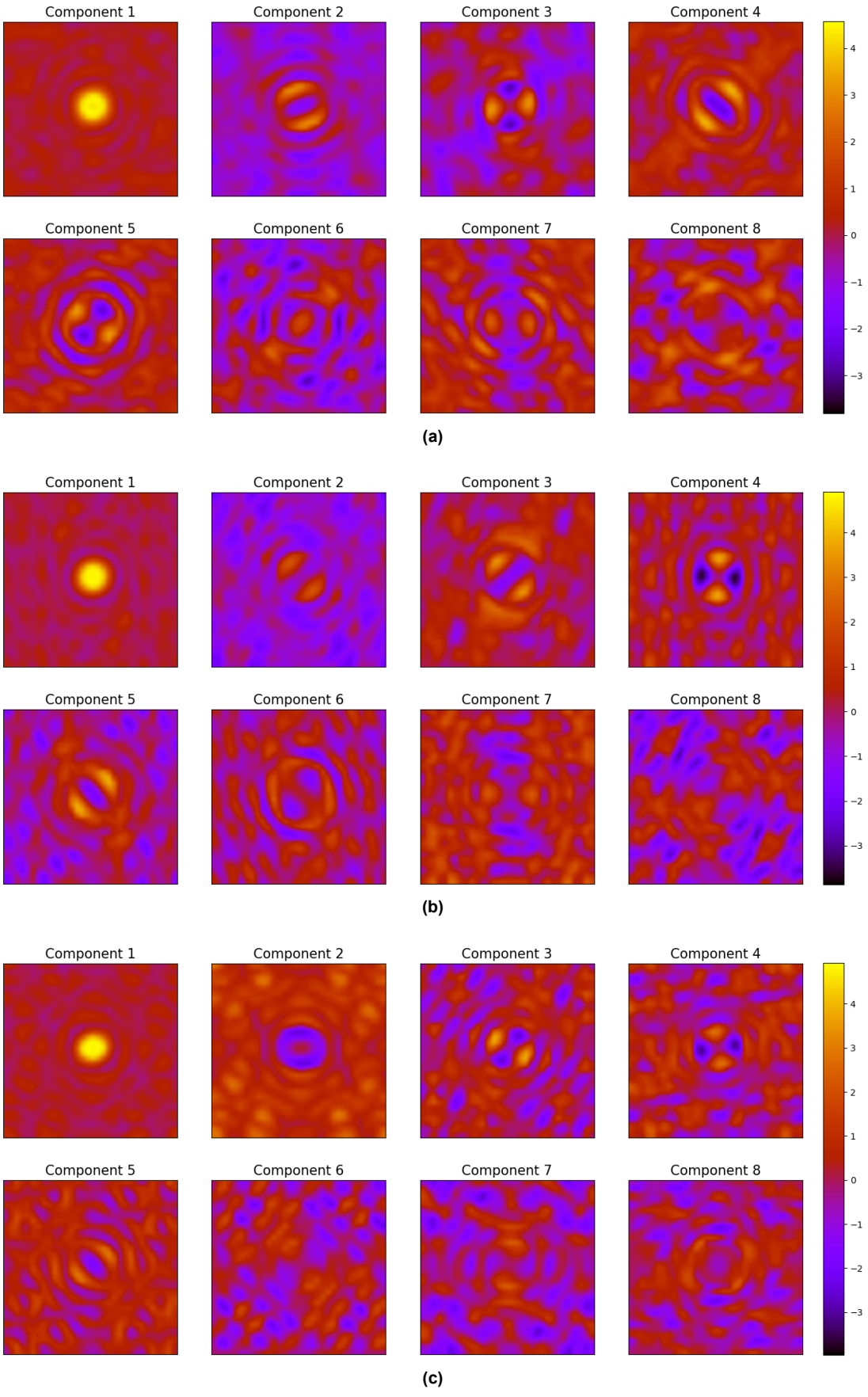


Figure D.2: Visualizing the components learned by PCA on the Fourier transformed Fe-classified data. (a) Topo A001. (b) Topo B0376. (c) Topo B0627.

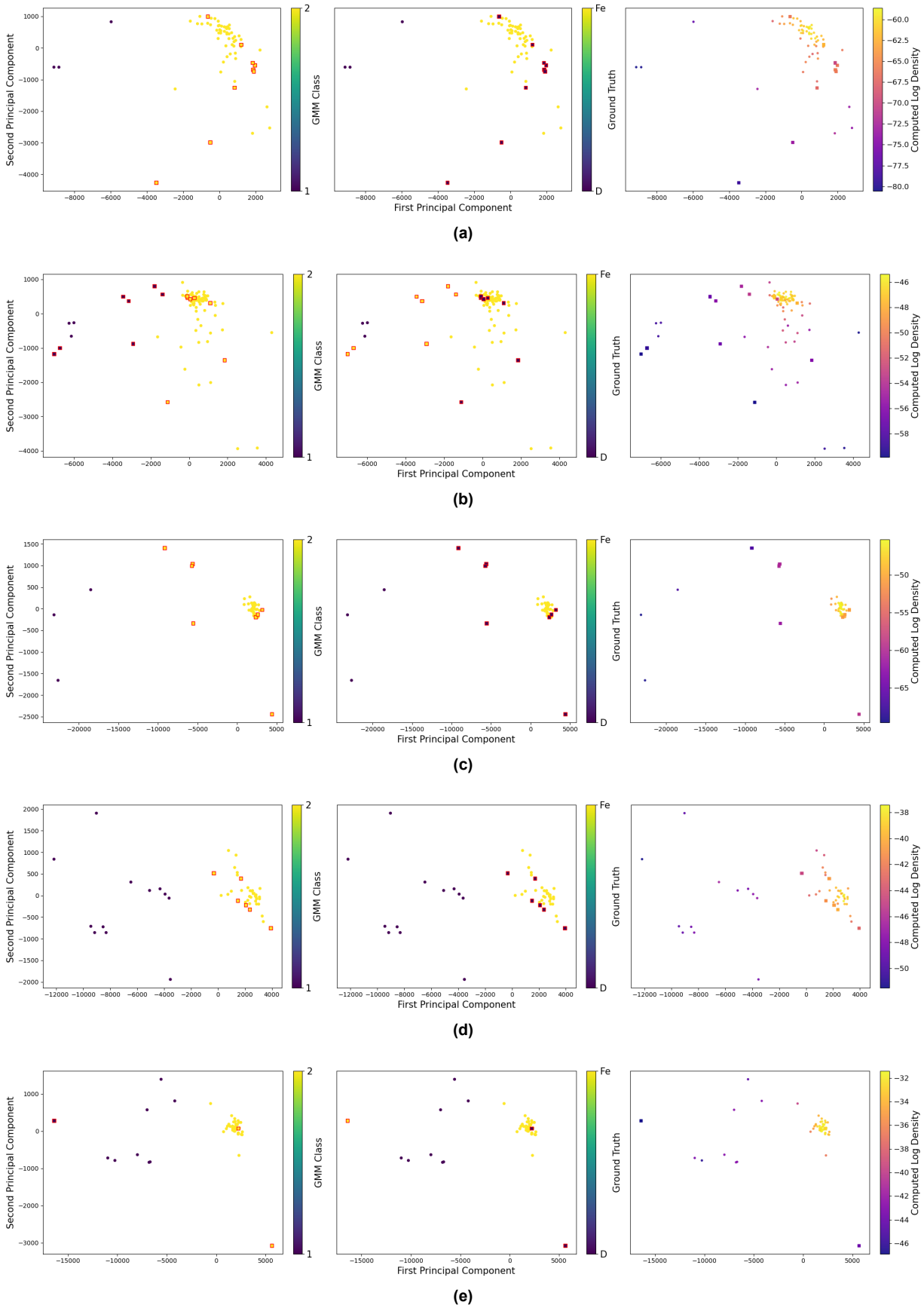


Figure D.3: PCA space plot of our Fe data points, colorized by GMM classification, ground truth, and density-based clustering results. Adequately classified sites are represented by circles. Misclassified sites are represented by red squares. (a) Topo A228. (b) Topo B2060. (c) Topo B0917. (d) Topo B1544. (e) Topo B2731

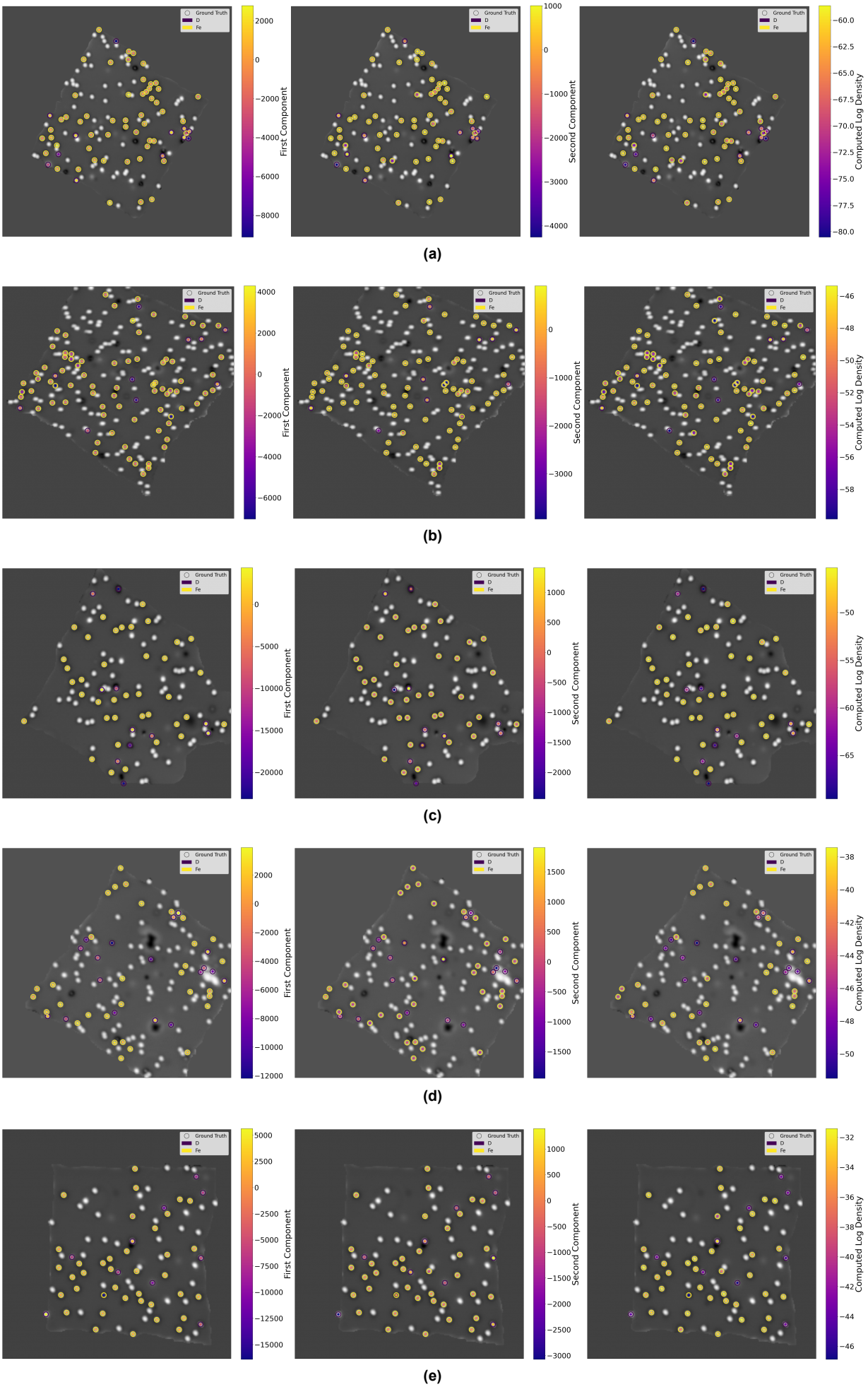
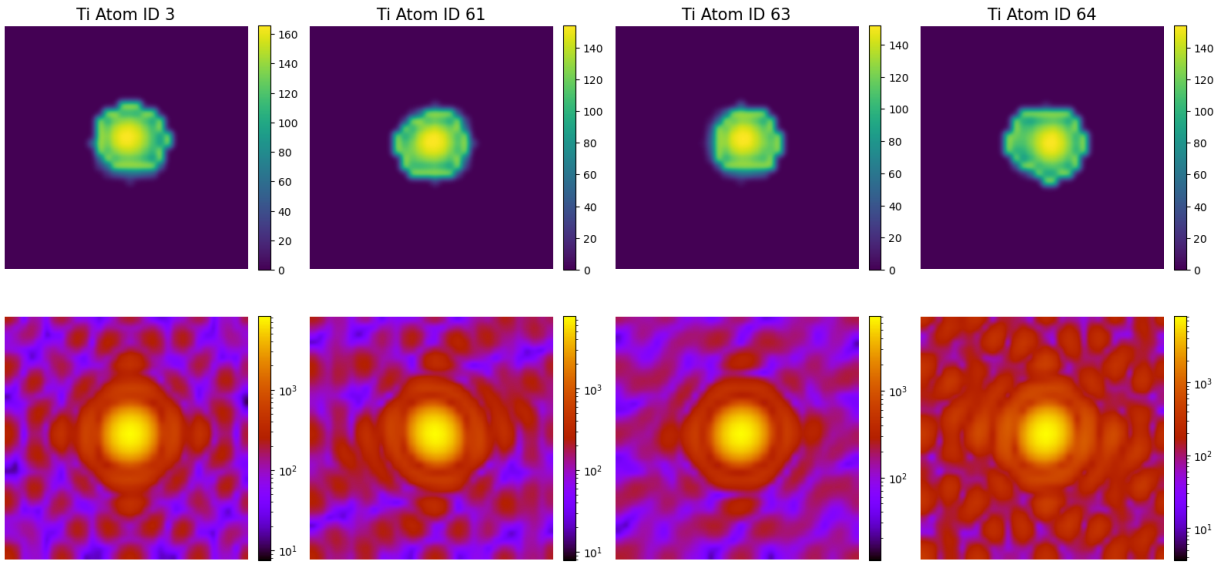


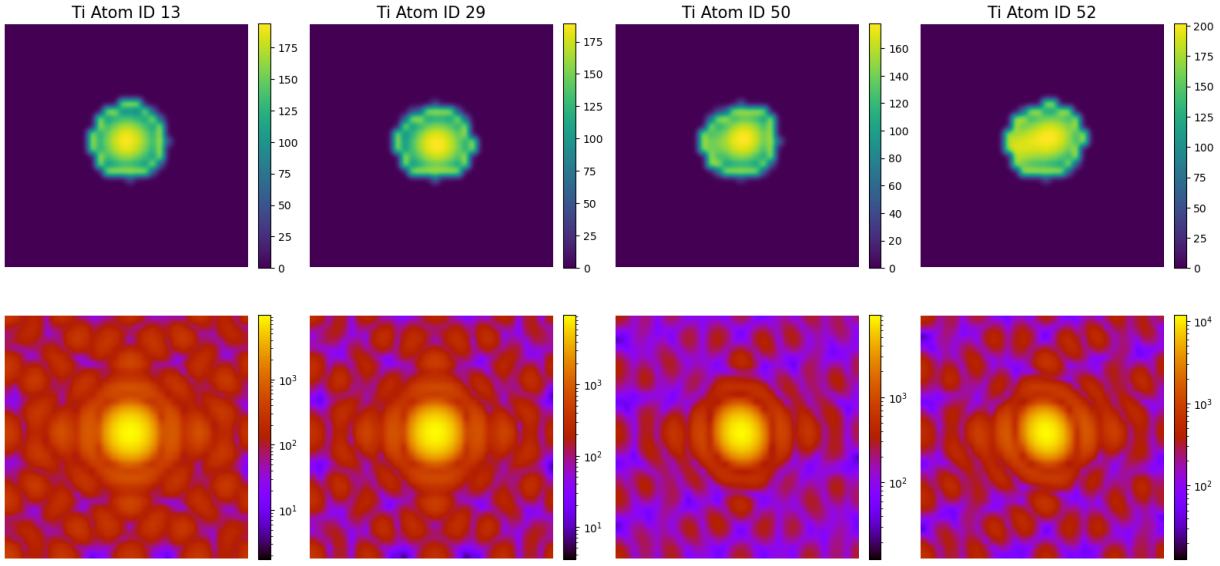
Figure D.4: Projecting the PCA and density-based clustering results onto our Fe-classified coordinate sites. (a) Topo A228. (b) Topo B2060. (c) Topo B0917. (d) Topo B1544. (e) Topo B2731

E

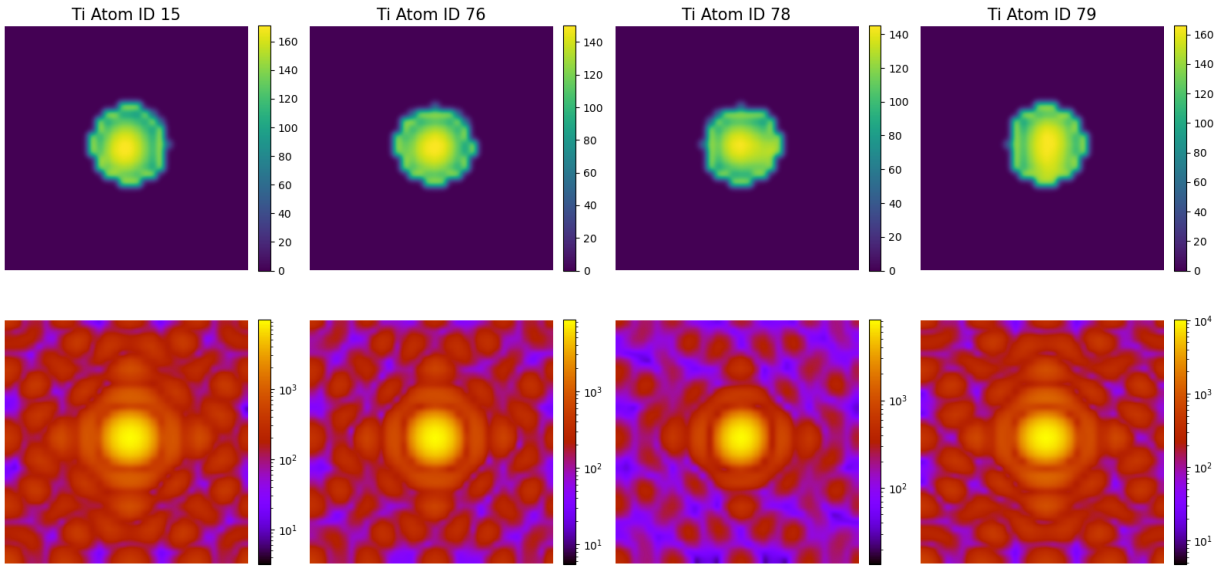
Supplementary Materials E: Class-conditioned Subclassification on Ti-classified Sites



(a)

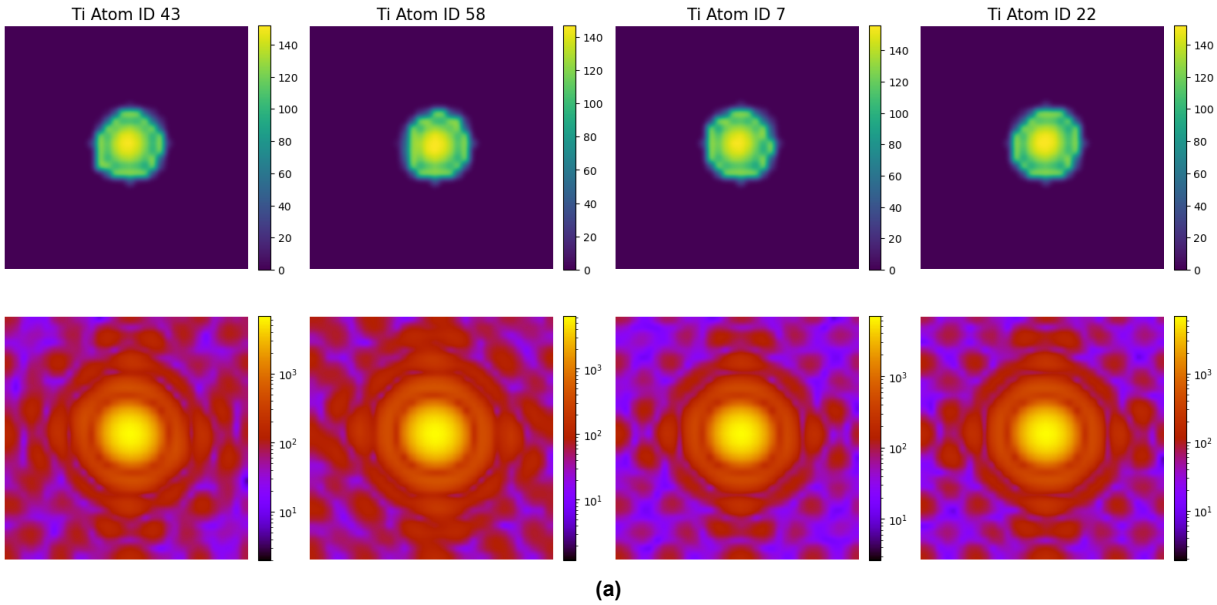


(b)

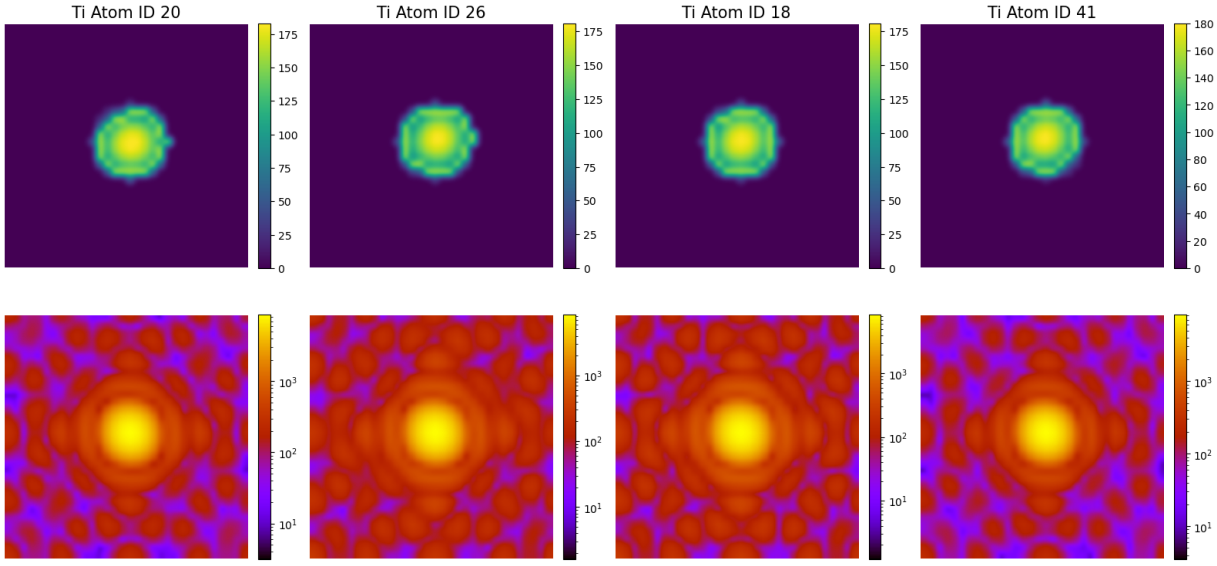


(c)

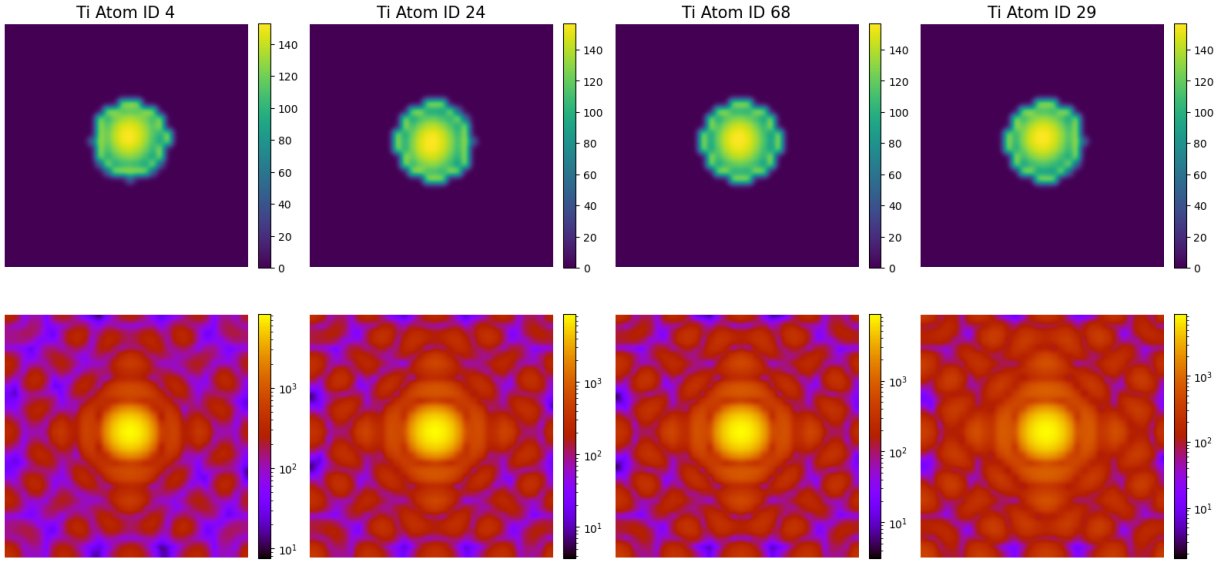
Figure E.1: Randomly sampled data points from the Ti class-conditioned GMM classification and their Fourier transforms, class 1 (outliers). We draw the contour (bright green) around each atom site to extract its shape and orientation. (a) Topo A001. (b) Topo B0376. (c) Topo B20627.



(a)



(b)



(c)

Figure E.2: Randomly sampled data points from the Ti class-conditioned GMM classification and their Fourier transforms, class 2 (Ti atoms). We draw the contour (bright green) around each atom site to extract its shape and orientation. (a) Topo A001. (b) Topo B0376. (c) Topo B20627.

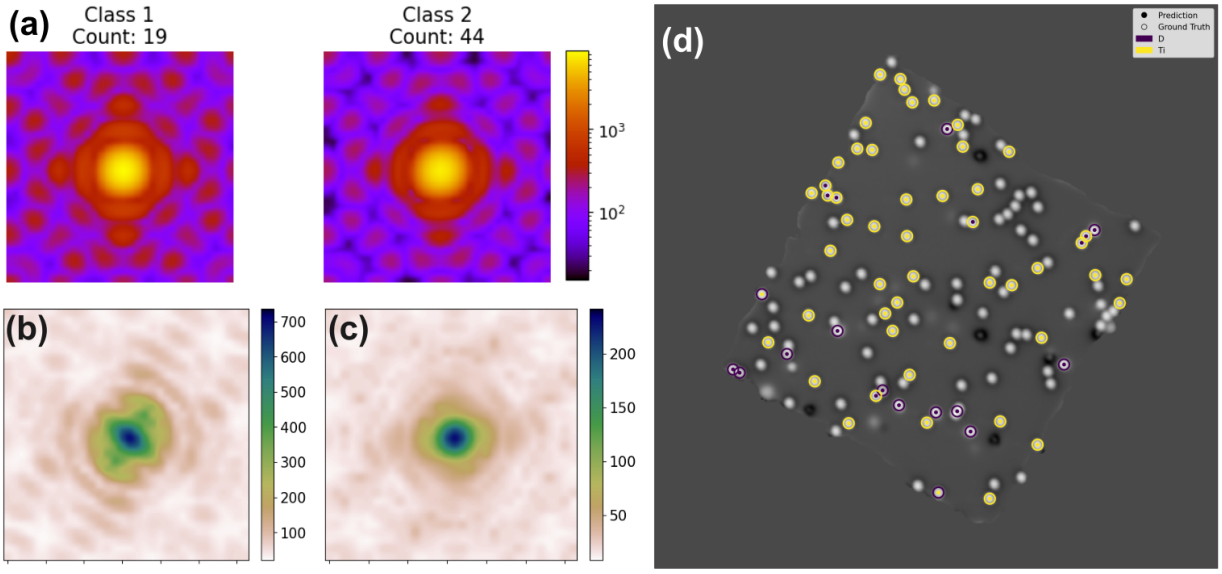


Figure E.3: Topo A228 class-conditioned GMM analysis on Fourier transformed Ti-classified sites. (a) Mean Fourier modulus of the atom-cropped image within each learned class. (b and c) Standard deviation in Fourier space within each learned class. (b) Class 1 (outliers). (c) Class 2 (Ti atoms). Notice that the value ranges are widely different for each class. (d) Plot of the GMM results with an overlay of the ground truth labels.

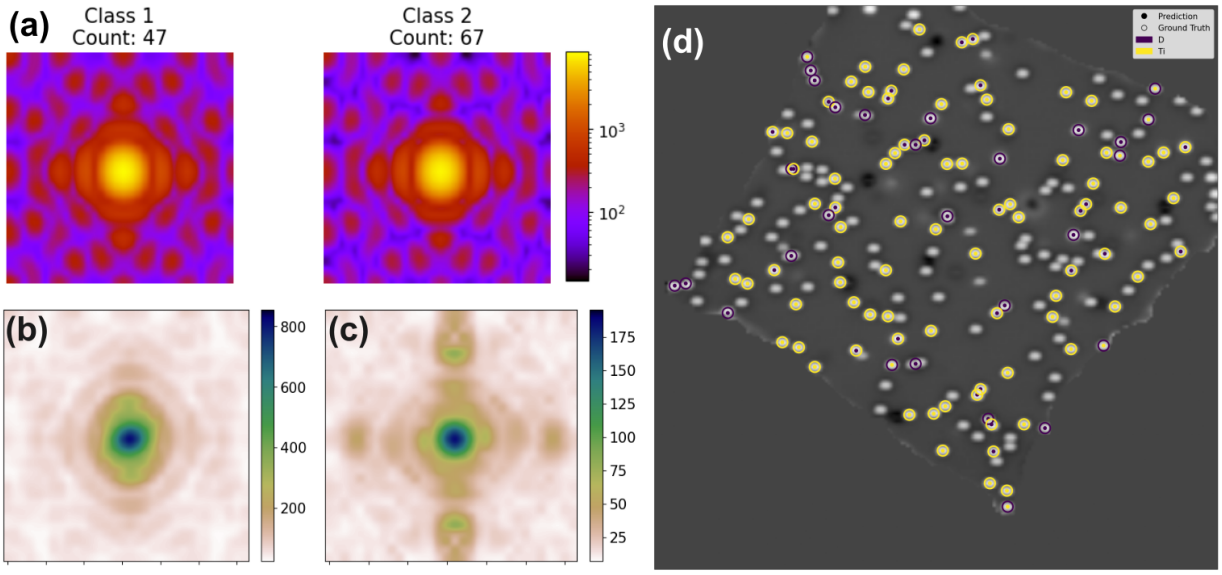


Figure E.4: Topo B2060 class-conditioned GMM analysis on Fourier transformed Ti-classified sites. (a) Mean Fourier modulus of the atom-cropped image within each learned class. (b and c) Standard deviation in Fourier space within each learned class. (b) Class 1 (outliers). (c) Class 2 (Ti atoms). Notice that the value ranges are widely different for each class. (d) Plot of the GMM results with an overlay of the ground truth labels.

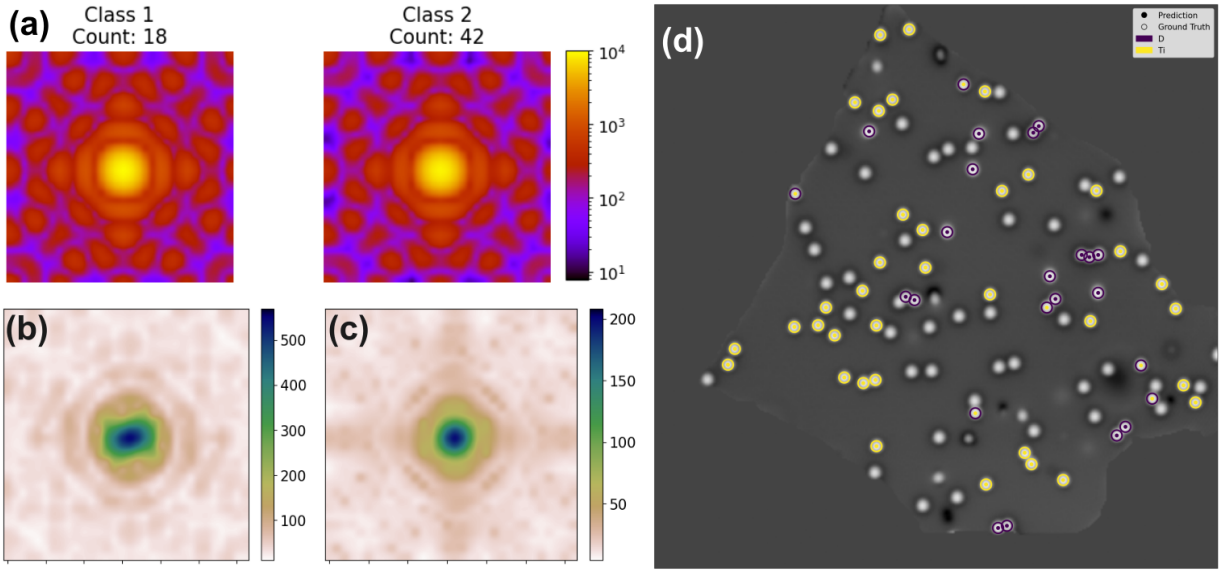


Figure E.5: Topo B0917 class-conditioned GMM analysis on Fourier transformed Ti-classified sites. (a) Mean Fourier modulus of the atom-cropped image within each learned class. (b and c) Standard deviation in Fourier space within each learned class. (b) Class 1 (outliers). (c) Class 2 (Ti atoms). Notice that the value ranges are widely different for each class. (d) Plot of the GMM results with an overlay of the ground truth labels.

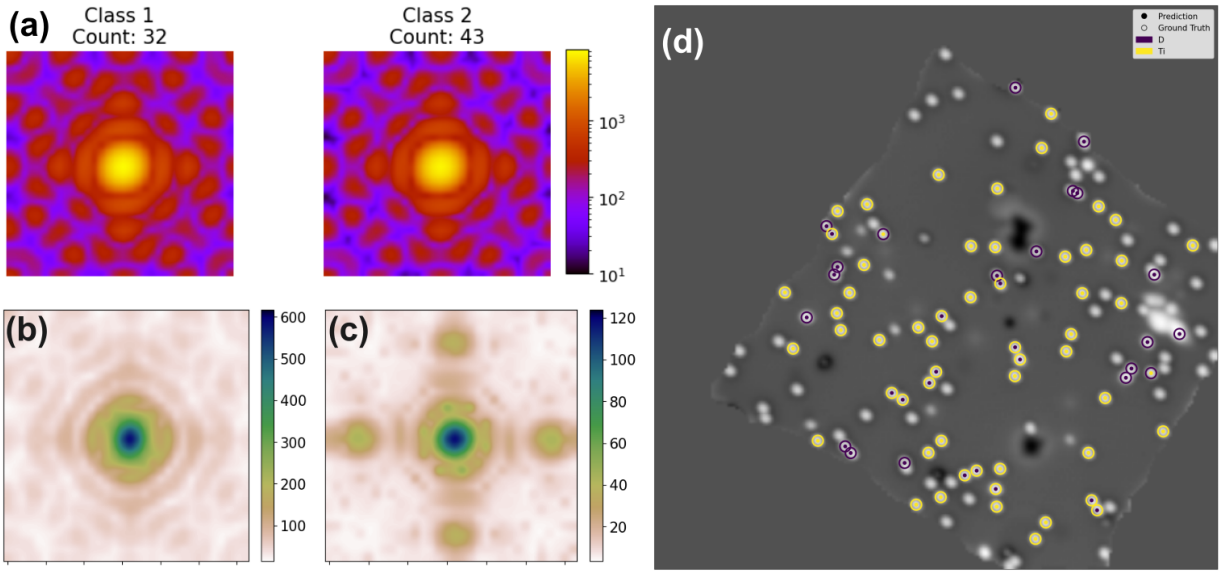


Figure E.6: Topo B1544 class-conditioned GMM analysis on Fourier transformed Ti-classified sites. (a) Mean Fourier modulus of the atom-cropped image within each learned class. (b and c) Standard deviation in Fourier space within each learned class. (b) Class 1 (outliers). (c) Class 2 (Ti atoms). Notice that the value ranges are widely different for each class. (d) Plot of the GMM results with an overlay of the ground truth labels.

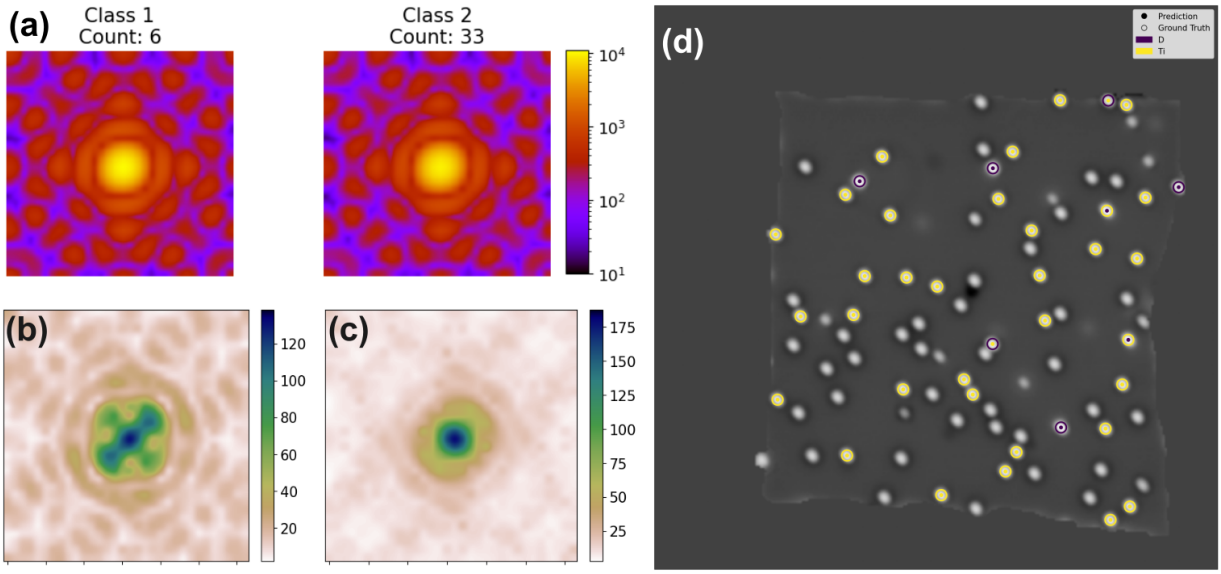
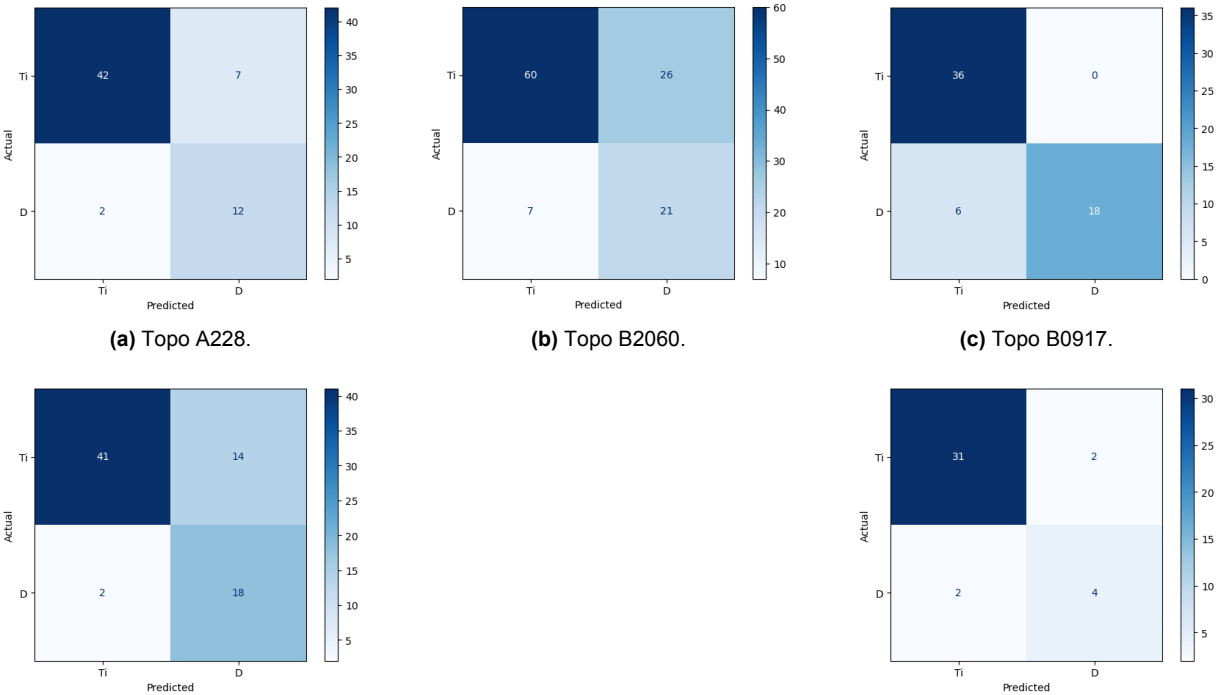


Figure E.7: Topo B2731 class-conditioned GMM analysis on Fourier transformed Ti-classified sites. (a) Mean Fourier modulus of the atom-cropped image within each learned class. (b and c) Standard deviation in Fourier space within each learned class. (b) Class 1 (outliers). (c) Class 2 (Ti atoms). The outliers in this case skew the classification. (d) Plot of the GMM results with an overlay of the ground truth labels.



(a) Topo A228.

(b) Topo B2060.

(c) Topo B0917.

(d) Topo B1544.

(e) Topo B2731.

Figure E.8: Confusion matrices of results from the Ti class-conditioned GMM classification of Fourier transforms.

Supplementary Materials F: Quantifying the Variation in Ti-classified Sites

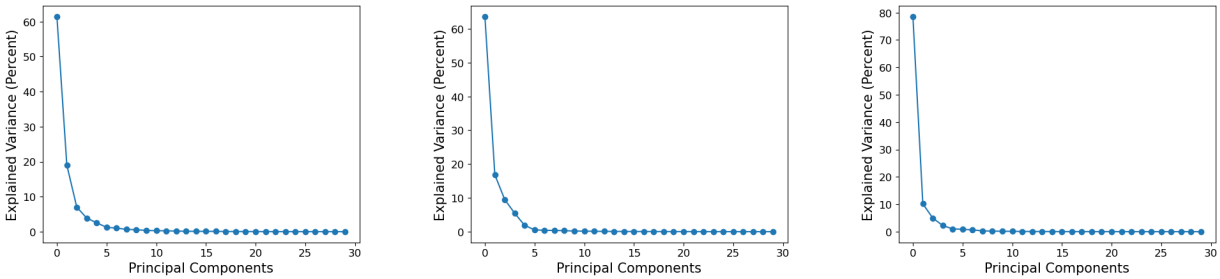
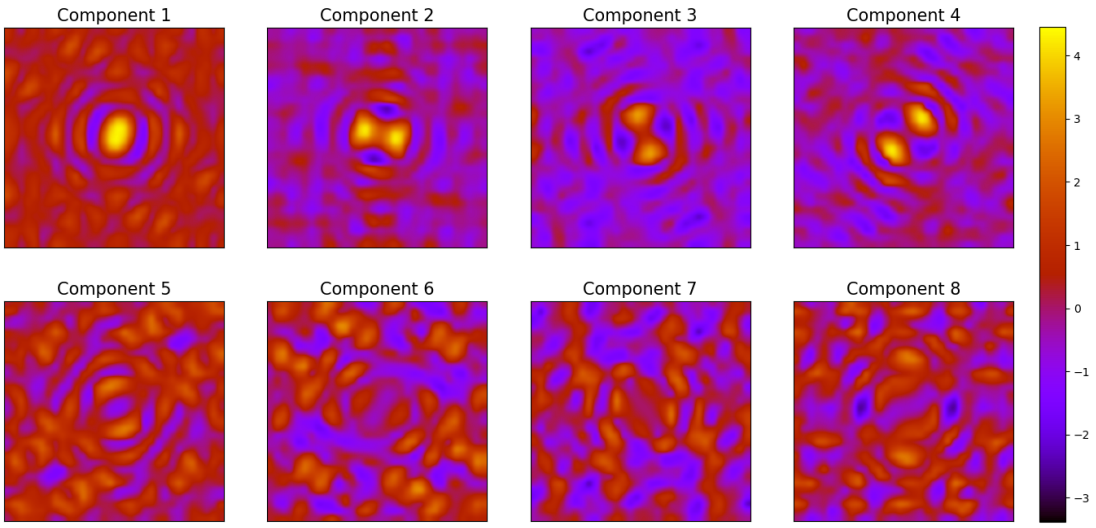
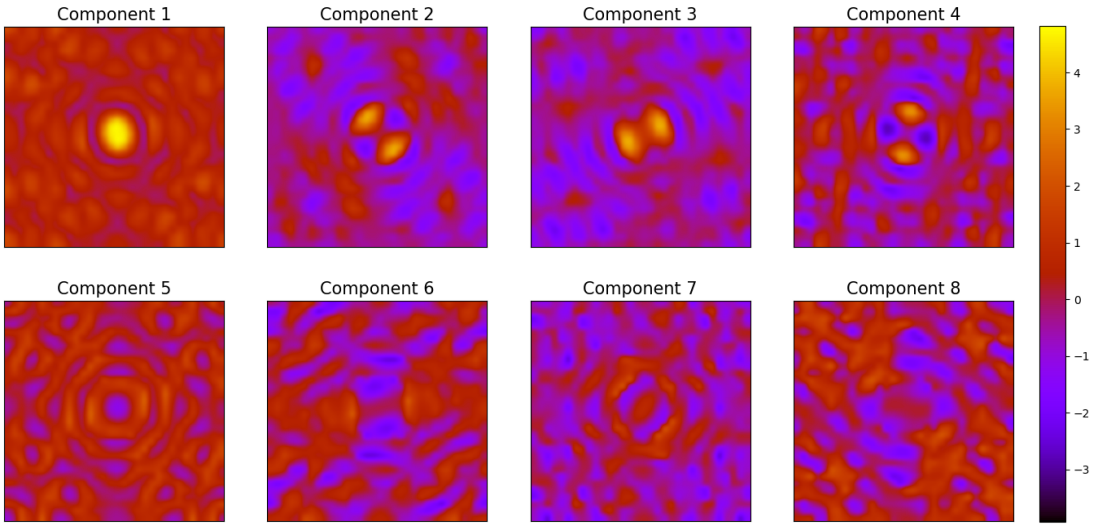


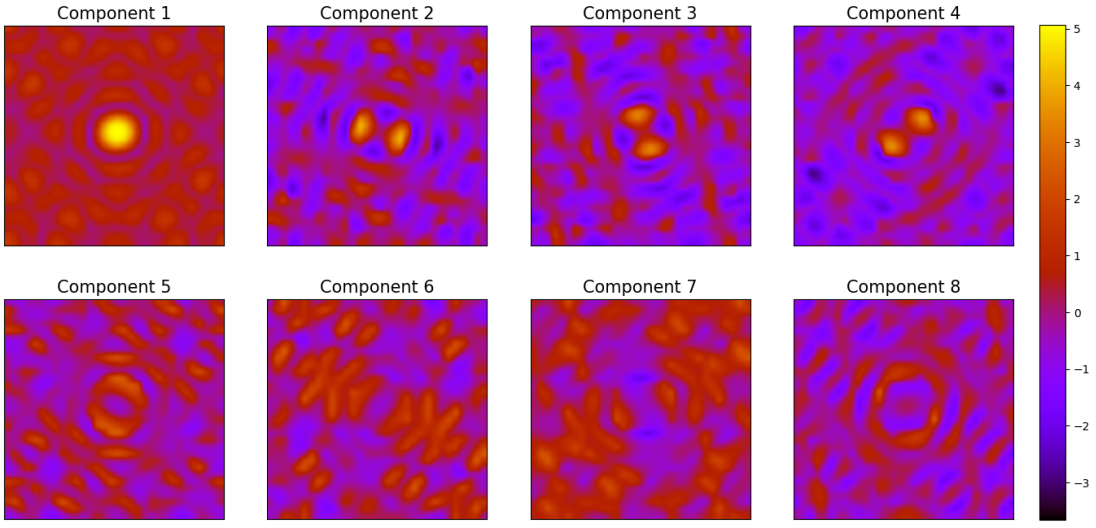
Figure F.1: Plotting the explained variance by principal component against the principal components learned from the Fourier transformed Ti-classified data. (a) Topo A001. (b) Topo B0376. (c) Topo B0627.



(a)



(b)



(c)

Figure F.2: Visualizing the components learned by PCA on the Fourier transformed Ti-classified data. (a) Topo A001. (b) Topo B0376. (c) Topo B0627.

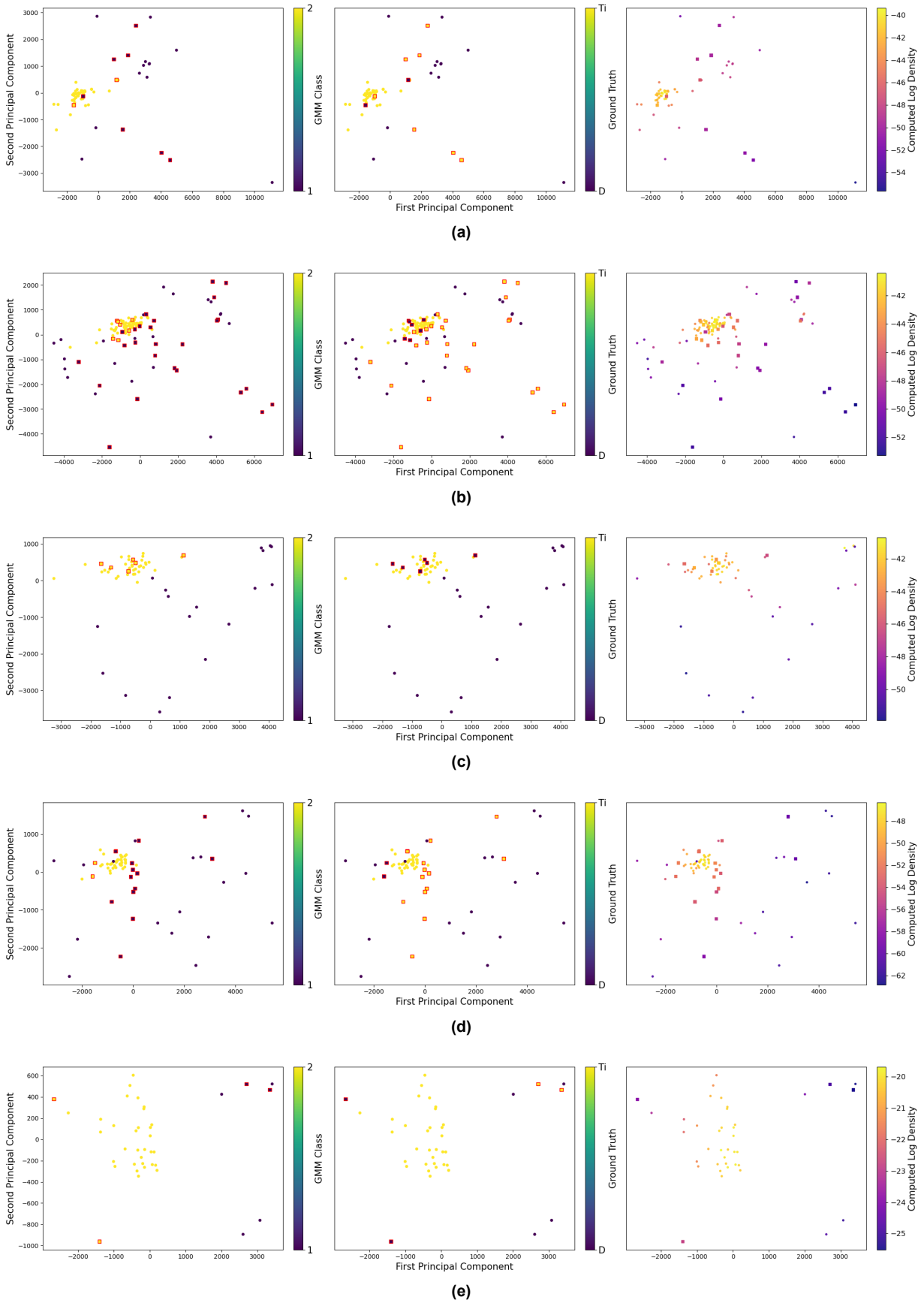
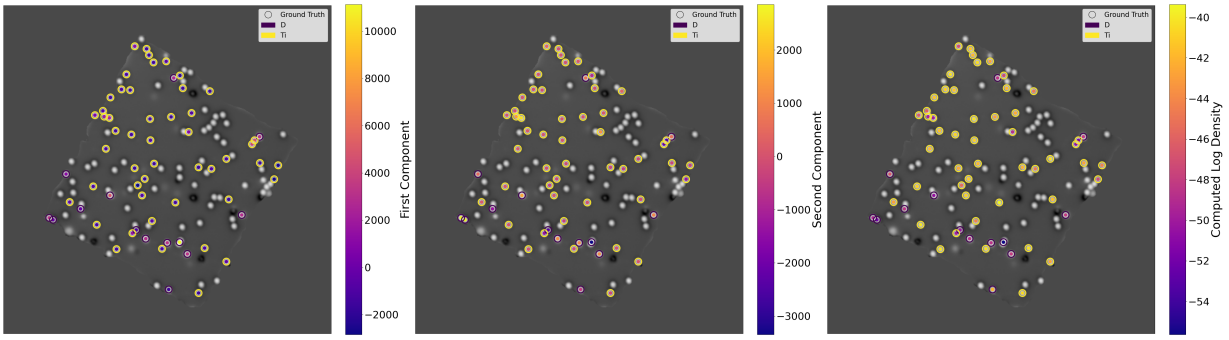
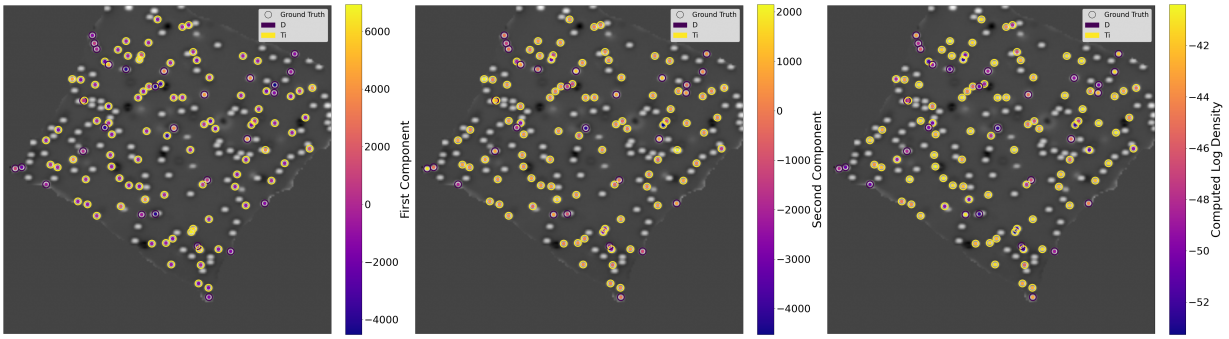


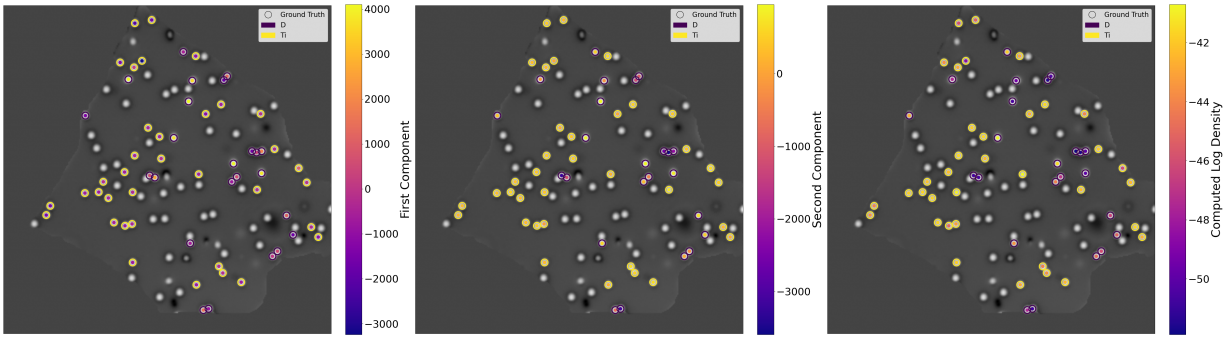
Figure F.3: PCA space plot of our Ti data points, colorized by GMM classification, ground truth, and density-based clustering results. Adequately classified sites are represented by circles. Misclassified sites are represented by red squares. (a) Topo A228. (b) Topo B2060. (c) Topo B0917. (d) Topo B1544. (e) Topo B2731



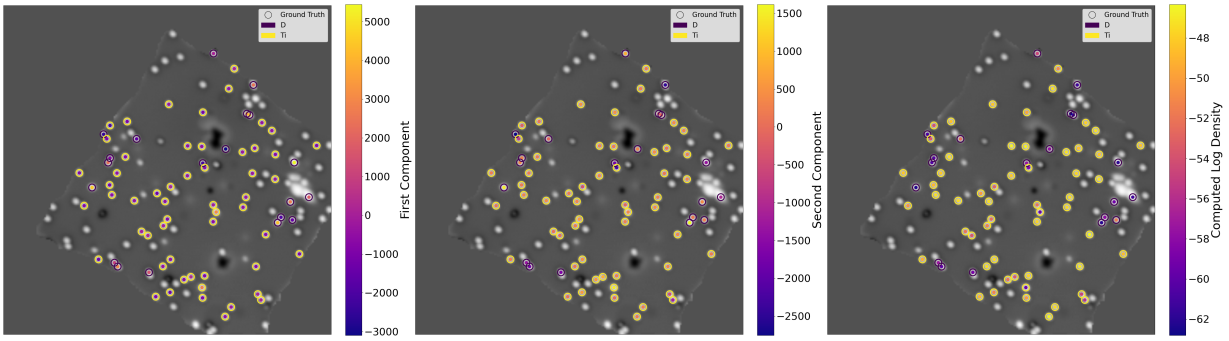
(a)



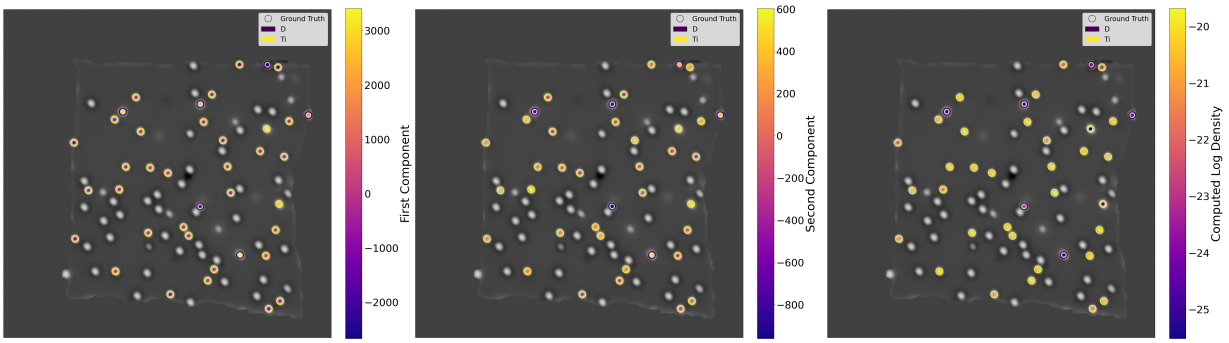
(b)



(c)



(d)



(e)

Figure F.4: Projecting the PCA and density-based clustering results onto our Ti-classified coordinate sites. (a) Topo A228. (b) Topo B2060. (c) Topo B0917. (d) Topo B1544. (e) Topo B2731