Development and Expansion of Tools for Data-Driven Materials Development

Jessica Kong

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Jim Pfaendtner, Chair

David A.C. Beck

Anne B. McCoy

Program Authorized to Offer Degree:

Chemistry

University of Washington

Abstract

Development and Expansion of Tools for Data-Driven Materials Development

Jessica Kong

Chair of the Supervisory Committee:
Jim Pfaendtner
Department of Chemical Engineering

Machine learning and natural language processing techniques are being integrated into chemistry and materials science, finding utility at field and domain levels of research. While these tools have existed, the relative recent emergence of these tools within high-level programming languages like Python means that they have only recently begun to be utilized at scale. In this dissertation, I explore the ways in which these tools can be applied in field-specific settings and a general, domain-level one. In one, I develop a new analysis methodology utilizing image registration, dimensionality reduction, and multivariate analysis to derive information from multimodal atomic force microscopy images. In a second, I utilize and develop reusable code for a Python package within the scanning probe community to obtain insights about and examine impacts of different physical contributions to a measured signal in a specialized atomic force microscopy technique. In another, I introduce a practitioner-centric framework for evaluating topic models that moves away from the dichotomic approach utilized in model development with a critical downstream benefit of advancing data-driven materials research via natural language processing. These works illustrate the ways in which existing machine learning and natural language processing are powerful tools and makes a case for the need of domain expertise in their development, much like the symbiotic work of computationalists, experimentalists, and theorists.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

The biggest thank you to my family for their unequivocal support without which none of this would have been possible. Thank you to my mother for being the first person to show me how to approach experiments: with perceptive iteration, a sense of playfulness, and openness to unexpected results, all through cooking. She imbued in me qualities of a scientist long before graduate school. She has also gifted me the privilege of focusing on myself by shielding me from financial and caretaking responsibilities; and space to come into my own being with her receptivity to my perspectives, especially when they are different. I feel so lucky to have her as my mother. Thank you to my father for being one of my first educators by teaching me his native language with a sense of wonder and nurturing my curiosity by answering all my questions about what random characters in a newspaper meant. His ability to endure and take big leaps of faith with potentially dire outcomes are the reason I am here. To my cousin and her spouse, for showing me that there's fun to be had in working and so much more to life than work. To my older brother who created memorable experiences to the library when I was younger and who continues to gift me all sorts of books; I attribute a lot of my love for intellectual exploration to you. To my younger brother, for showing me that we collectively owe each other so, so much. To my younger sister, for teaching me about sensitively relating to people. To my cousin, for showing me what it's like to be firmly true to yourself. Most of all, for their unconditional love, consistent presence, and simply being themselves. My individual relationships with each of you along with our shared time together shape who I am at my core. You all mean the world to me.

I am so grateful for my relationships with people that span a variety of different contexts, identities, and lived experiences. Thank you all for seeing me and being part of my comically absurd and statistically unlikely 24 year-long academic process. Thank you especially to Beth for enlightening conversations. I hold you all close to my heart.

To my colleagues and collaborators at the University of Washington, Pacific Northwest National Laboratory, and Oak Ridge National Laboratory, thank you. While I only mention a small portion of all that I have worked on throughout my time here, you are each a part of why I've been able to

think and tinker in many different fields. Pfaendtner Research Group, my time with you all has been as much restorative as it has been joyful; thank you all so much.

To all the educators I have had in my life, thank you for laying the foundation on top of which this work sits. I am indebted to the teachers and role models I had in the Oakland Unified School District who have believed in and encouraged me for as long as I can remember. Thank you to the exceptional and caring professors I had at Middlebury College, especially those in the Chemistry Department. My interactions and experiences with Dr. AnGayle Vasiliou and Dr. Sunhee Choi are the reason why I considered and decided to go to graduate school.

In the spirit of the liberal arts, I'm also thankful for works by scholars in various fields; they have allowed me to think expansively and engage thoughtfully. Thank you to James Baldwin, Dr. Tressie McMillan Cottom, Cathy Park Hong, Dr. bell hooks, Dr. Anthony Abraham Jack, Mikki Kendall, Kiese Laymon, Dr. Jennifer Morton, Dr. Andrew Solomon, Dr. Amia Srinivasan, Jia Tolentino, and many others. Your works have collectively given me the ability to articulate my experiences while enabling me to integrate seemingly dichotomic truths.

Finally, a heartfelt thank you to my advisors Dr. Jim Pfaendtner and Dr. David Beck for creating an environment in which I could live out my undergraduate dreams of graduate-level research. Jim, thank you for welcoming me into your lab; providing the space for me to grow as a person and scientist; and meeting me where I am, always. Thank you especially for being reflective and receptive. Dave, thank you for your humanistic approach to our meetings and providing guidance at pinpoint moments when I have felt completely lost. My experience with you two has given me clarity and confidence on things that have made a difference so other worldly I could not have imagined them from where I started. Both of you together created a transformative experience for me which I cognitively know is a result of finely balancing purpose, practicality, and trust. I can only imagine how difficult this is in practice and my gratitude for your Herculean efforts is immeasurable. I feel incredibly fortunate to have you both as advisors and will always draw positively from this experience.

DEDICATION

For my mother and father,

who are wonder and resilience embodied.

Chapter 1. INTRODUCTION

At the foundation of machine learning and artificial intelligence are mathematical theories that have existed for decades.[1–7] Despite this, adoption of these tools in the field of chemistry at mass throughout the entirety of the research pipeline has been slow, with recent publications utilizing brute force, human-powered approaches toward tasks that would benefit either from existing tools or their continued development.[8–11] Accessible high-level programming languages like Python along with the development of educational curriculum incorporating computer-science based skills for domain researchers has aided the incorporation of these tools that were previously inaccessible.[12] While these tools grounded in mathematical and statistical frameworks have found widespread utility, significant work remains in making them a central component of experimental chemistry. As a start, the works discussed herein look at how techniques that are broadly categorized as image registration, dimensionality reduction, multivariate analysis, and natural language processing can be utilized in chemistry.

In chapter 2, I introduce a new methodology utilizing multimodal atomic force microscopy involving the sequential application of statistical learning methods. This analysis pipeline extracts compositional information that cannot be derived from the analysis of images from any one technique alone; it is a demonstration of how data science techniques can be merged with microscopy techniques to fully leverage the gestalt of a multimodal approach. It is also one of the first experimental validations in an applied research setting of an emerging technique called photoinduced force microscopy, providing the scanning probe microscopy community with a sense of how this nanoscale infrared-based technique compares to its established, macroscale analog: Fourier Transform Infrared Spectroscopy. The analysis pipeline introduced is broadly

generalizable to any image data, encouraging researchers to move beyond non-rigorous methods of analyzing multimodal data.

In chapter 3, I study the role of electrostatics and its impact on piezoresponse-based force microscopy methods. A method called Band-Excitation Piezoresponse Force Microscopy (BE-PFM) is used to collect time-dependent piezoresponse 'relaxation' spectra on exemplary non ferroelectric and ferroelectric materials. The resulting spectra are analyzed by curve fitting to extract information telling of the physical phenomena occurring. I decouple the electrostatic contribution from the measured signal in this technique look at the impact this electrostatic contribution has on a related technique, contact Kelvin Probe Force Microscopy. Based on the results of this work, make practical experimental and analysis recommendations for removing its contribution in piezoresponse-based force measurements. A component of this work entailed writing a module for analyzing the results from BE-PFM and was incorporated into Pycroscopy, a Python package in part for scanning probe microscopists.

In chapter 4, I explore how different topic models can form the basis of a machine-focused information retrieval system for scientific text that is content based. This work has downstream implications on curating field-specific databases and ultimately, the quality of high-throughput efforts for materials design. I evaluate their performance with a prevailing metric and introduce a practitioner-centric one based on mutual information that is a proxy for their ability to return high quality results based on an abstract query. These metrics are correlated with human judgments of the semantic structure of the models. Pending results of two human evaluation tasks, we find that in most cases, the topic models examined can be seamlessly integrated into a domain setting according to the prevailing metric. We find that the mutual information based metric is more telling of how well a model performs as the basis of an information retrieval system. This work establishes

a practitioner-centric, domain-agnostic framework for evaluating topic models, regardless of model assumptions. The broader implications of this work include guiding annotation efforts for downstream tasks and increasing the rate at which novel ideas are adopted all of which accelerate materials research. Lastly, a high-level overview of the value of this work is presented along with suggestions for future work.

Chapter 2. NANOSCALE STRUCTURE-FUNCTION PROPERTIES WITH IMAGE REGISTRATION, DIMENSIONALITY REDUCTION, AND REGRESSION [1]

2.1 ABSTRACT

Correlating nanoscale chemical specificity with operational physics is a long-standing goal of functional scanning probe microscopy (SPM). We employ a data analytic approach combining multiple microscopy modes using compositional information in infrared vibrational excitation maps acquired via photoinduced force microscopy (PiFM) with electrical information from conductive atomic force microscopy. We study a model polymer blend comprising insulating poly(methyl methacrylate) (PMMA) and semiconducting poly(3-hexylthiophene) (P3HT). We show that PiFM spectra are different from FTIR spectra but can still be used to identify local composition. We use principal component analysis to extract statistically significant principal components and principal component regression to predict local current and identify local polymer composition. In doing so, we observe evidence of semiconducting P3HT within PMMA aggregates. These methods are generalizable to correlated SPM data and provide meaningful technique for extracting complex compositional information that is impossible to measure from any one technique.

2.2 INTRODUCTION

Chemical structure and composition are inextricably linked to the functional properties that give rise to materials with promising applications in systems from solar cells to batteries.[13,14]

[1] Adapted with permission from Kong, J.; Gridharagopal, R.; Harrison, J.S.; Ginger, D.S. Identifying Nanoscale Structure–Function Relationships Using Multimodal Atomic Force Microscopy, Dimensionality Reduction, and Regression Techniques. *J. Phys. Chem. Lett.* **2018.** https://doi.org/10.1021/acs.jpclett.8b01003

Quantitative structure-function models are key to efficiently developing paradigms that enable materials discovery at a rate that outpaces the classical synthesis-characterization-theory approach.[15] Images from multimodal atomic force microscopy (AFM) techniques which provide spatially resolved, quantitative structural and functional information make for particularly well suited data sets for this purpose.[16] Statistical learning tools can be leveraged to apply detailed quantitative analysis on structural and functional maps to develop insight for optimized materials and devices.

AFM-based techniques have become ubiquitous in functional nanoscale characterization.[17–21] Imaging chemical composition on this scale however, has been challenging; historically, such data were acquired using techniques including near-field scanning optical microscopy (NSOM) and tip-enhanced Raman spectroscopy (TERS), both of which can image chemical structure far beyond diffraction-limited optics.[22–24] NSOM and TERS, however, impose a limitation of requiring careful optical alignment to "directly" collect optical information.[23,25] To circumvent this issue, scanning probe methods like photoinduced force microscopy (PiFM)[26,27] and photothermal infrared microscopy (PTIR, also called AFM-IR)[28,29] have been developed to mechanically probe vibrational modes. These techniques use an infrared (IR) laser that spans the vibrational excitation range for many common organic materials. Instead of measuring the excitation via a change in absorption as in NSOM, they measure the response via a mechanical response of the tip. Of these, PiFM has the distinction of being a nominally "non-contact" technique as opposed to a direct thermal expansion measurement in PTIR,[30] though the ultimate result in either paradigm are IR spectra with nanoscale resolution measured via cantilever dynamics.

Briefly, PiFM generates spatial maps with chemical contrast by measuring the time-integrated photoinduced gradient force between a tip and sample due to the interaction between their

polarizabilities at wavenumbers corresponding to vibrational modes of different chemical species.[26] PiFM can also generate hyperspectral images containing a photoinduced force infrared spectrum at each pixel, enabling visualization of absorption maps at all wavelengths. The current understanding of time-integrated photoinduced force[31–35] and theory of PiFM can be found elsewhere.[26,36]

One of the principal advantages of chemical imaging via PiFM is that correlated functional and chemical maps can be utilized to advance our understanding of the relationships between chemical structure and composition and electrical properties on the nanoscale. While hybrid systems that can obtain registered images of materials like AFM/mass spectrometry[37] are useful, they are more costly to implement and therefore motivate data-driven approaches to extract such information. Gleaning this level of insight requires combining domain knowledge and data science tools which in the field of scanning probe microscopy (SPM) have largely been applied to multidimensional SPM data sets.[16] For example, principal component analysis (PCA) has been used as a tool to analyze band excitation SPM data[38] and extract statistically significant contributions to signals in general mode and general dynamic AFM.[39–41] To gain physical insights from multidimensional first-order reversal curve current-voltage (FORC-IV) SPM data, techniques like *k*-means clustering and Bayesian linear unmixing have been applied.[42,43] With the continued development of multidimensional SPM techniques, statistical learning tools will have an increasingly important role in processing, analyzing, and extracting information from progressively larger data sets.

A natural extension of the permeation of these data science techniques into the SPM community is to apply these tools to multimodal images with the goal of understanding the relationship between chemical structure and functional properties. As a simple demonstration of the utility of this approach, we apply statistical learning techniques to registered hyperspectral photoinduced

force and conductive AFM (cAFM) images of a model polymer blend system: a phase-segregated blend of semiconducting poly(3-hexylthiophene) (P3HT) and insulating poly(methyl methacrylate) (PMMA). These polymers' distinct chemical fingerprints and the strong contrast between their electronic properties make this blend an ideal system for benchmarking analysis techniques to be used in increasingly complex systems with optoelectronic and chemical heterogeneity such as emerging mixed cation hybrid perovskites.[8,44]

In this work, we employ affine transformations as an image registration technique, PCA as a dimensionality reduction tool, and principal component regression (PCR) as a statistical analysis method. This sequence of techniques is a representative analysis flow in deriving correlations between chemical and functional image data. (See section 2.7.1 for graphical overview.) We apply this approach to our model polymer blend to demonstrate the robustness of the registration and interpretability of the data resulting from PCA and PCR. Despite the linear nature and relative simplicity of PCR, it provides a sensible prediction of current with the larger deviations providing useful insight into the tomographic structure of the film. Importantly, this method is generalizable to correlated datasets from other chemical mapping tools as well.

2.3 Experimental Methods

2.3.1 *Materials*

All materials were purchased from Sigma-Aldrich and used as received unless otherwise stated.

A 1:1 ratio, 1 wt.% solution of poly(3-hexylthiophene) (P3HT) and poly(methyl methacrylate) (PMMA) blend was prepared following methods in Kergoat et al.[45] PMMA was dissolved in dichlorobenzene by stirring overnight. Subsequently, P3HT was added to the PMMA solution and

stirred until completely dissolved. The solution was then filtered with a 0.45 µm polytetrafluorethylene (PTFE) filter.

2.3.2 *Polymer Film Fabrication*

Films were fabricated on ITO substrates (TFD, Inc.) which were sequentially cleaned by sonicating in dilute detergent, acetone, and propan-2-ol, and subsequently treated with oxygen plasma for 10 minutes. A thin layer of NiO was deposited by spin-coating in air a precursor consisting of 129 mg nickel(II) acetylacetonate dissolved in 5 mL of anhydrous ethanol with addition of 50 µL of 38% HCl at 5000 rpm at 11000 rpm/s for 30 s. The NiO layer was annealed at 320 °C for 45 min in air. Subsequently, the polymer precursor solution was spin-coated onto the NiO at 1500 rpm and 1500 rpm/s for 30 s in air. Films were then heated at 120 °C for 20 min.

2.3.3 *Hyperspectral Infrared Imaging*

A Molecular Vista Inc., Vista Scope coupled to a LaserTune QCL with wavenumber resolution of 0.5 cm^{-1} was used for hyperspectral photoinduced force hyperspectral imaging. The PiFM was operated in sideband mode where the first and second mechanical resonance frequencies at 320 kHz and 1.9 MHz of a Si cantilever with Pt coating were used to detect the photoinduced force and surface topography, respectively. During imaging, the laser was swept from 760 cm^{-1} to 1875 cm^{-1}. The sample was housed in a vacuum shroud back filled with nitrogen.

2.3.4 *Conductive Atomic Force Microscopy*

Conductive atomic force microscopy measurements were taken with an Asylum Research MFP-3D BIO atomic force microscope (Oxford Instruments) installed on an inverted Nikon Eclipse Ti inverted microscope and Table Stable vibration isolation stage. A surface bias of +5 V was applied

as a Budget Sensors ContE-GB tip with spring constant ~0.2 N/m was raster scanned across the surface of the sample to simultaneously measure the topography and current between the tip and sample.

2.3.5 *FTIR*

FTIR spectra were collected on a Thermo Fischer Nicolette 8700 with a nitrogen cooled MCT-A detector. 256 scans, 1.928 cm^{-1} resolution, with an optical velocity of 0.9494, an aperture of 3.

2.4 Results and Discussion

The distinct chemical nature of P3HT and PMMA result in unique vibrational modes in their respective IR spectra. Figure 2.1a shows attenuated total reflectance Fourier-transform IR (ATR-FTIR) spectra of PMMA, P3HT, and their blend in the range 1700-1760 cm^{-1}. (For full range spectra see Figure 2.7.) PMMA exhibits a vibrational mode corresponding to a carbonyl stretch at 1734 cm^{-1} that is also present in the blend but not in P3HT.[46] Fig. 2.1b shows that, within the range 780-840 cm^{-1}, P3HT contains an aromatic C-H out of plane bending mode at 824 cm^{-1}.[47] As expected, this vibrational mode is also present in the blend but not in PMMA. Importantly, vibrational modes present in the bulk material are also present at the nanoscale as shown in PiFM point spectra of films made of neat PMMA and P3HT in Figure 2.1c and 2.1d, indicating that they can be used as spectroscopic handles for nanoscale characterization.

According to the theoretical description of PiFM, the PiFM signal response originates from the real part of the polarizability,[26] though photothermal contributions can arise due to the imaginary part of the polarizability under certain imaging conditions.[36] In contrast, the absorption in FTIR spectra is proportional to the imaginary part of the linear susceptibility.[48] Thus, the signal in PiFM and FTIR spectra, at least in current understanding, probe different parts of a material's response

to an electromagnetic field. Therefore, it is expected that there will be differences between local PiFM and macroscopic FTIR spectra.

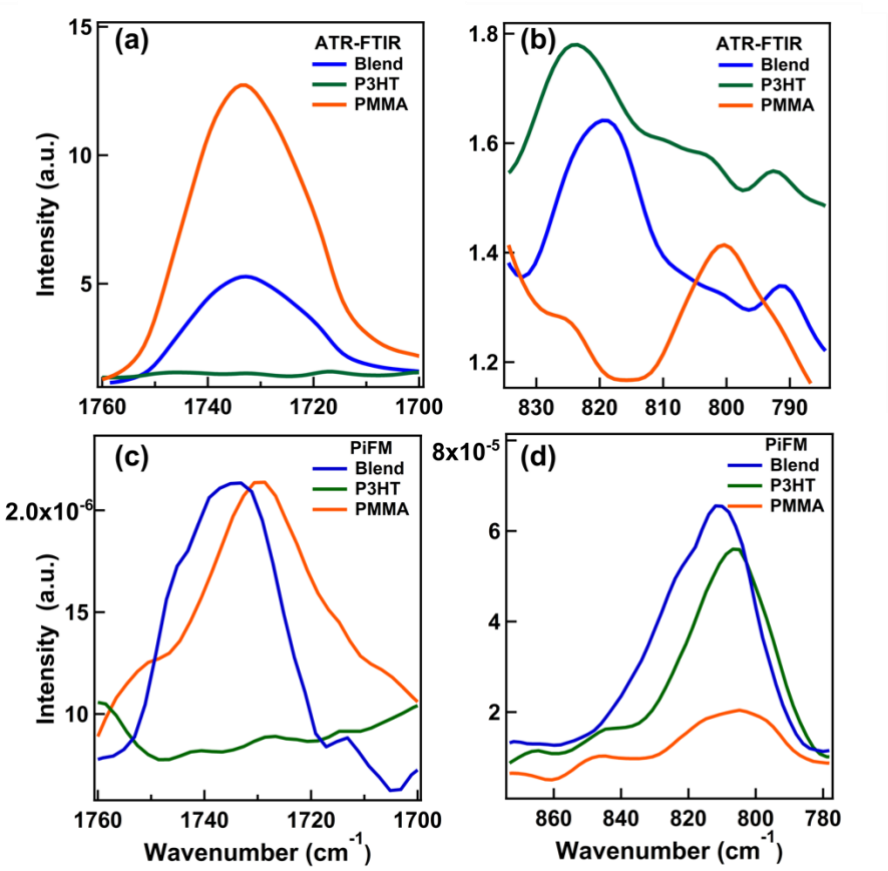

Figure 2.1. A comparison of attenuated total reflectance Fourier transform infrared (ATR-FTIR) **(a, b)** and photoinduced force microscopy (PiFM) point spectra **(c, d)** of neat polymer and blend films of poly(3-hexylthiophene) (P3HT) and poly(methyl methacrylate) (PMMA) showing that vibrational modes present in bulk ATR-FTIR spectra are also present in local PiFM spectra. **(a, c)** The spectral range of ~1700 to 1780 cm^{-1} corresponds to a C=O stretch in PMMA. **(b, d)** The spectral range of ~780 to 870 cm^{-1} corresponds to a C-H out-of-plane bending mode in P3HT. For the purposes of this work, however, it is only important that the PiFM spectra of PMMA and P3HT both exhibit distinct spectral fingerprints, a condition we expect to hold generally for materials that have distinguishable FTIR spectra. For instance, Figure 2.1c shows the presence of a carbonyl (C=O) stretch in neat PMMA and the polymer blend around 1728 cm^{-1} and 1735 cm^{-1}, respectively; its absence in P3HT is in excellent agreement with the bulk ATR-FTIR spectra.

However, the aromatic C-H bend of P3HT appears to shift from 824 cm^{-1} in the ATR-FTIR spectrum to 806 cm^{-1} in the PiFM point spectrum as shown in Figure 2.1d. Since P3HT exhibits no other vibrational modes in this region, it is reasonable to deduce that this peak corresponds to the aromatic C-H bend of this molecule.[47] While this shift places this vibrational mode in the vicinity of an unidentified vibrational mode of PMMA around 810 cm^{-1}, this peak is higher intensity in the P3HT blends and characteristic of P3HT.

We interpret the lineshape differences between our PiFM spectra compared to some of those previously reported in the literature (dispersive vs. quasi-Lorentizian) based on the physical origin of the PiFM signal. Notably, Jahng, et al. and Yang, et al. have reported that there are both thermal and photoinduced force contributions in PiFM, with the former giving rise to dissipative, Lorentzian line shapes and the latter to dispersive lineshapes.[36,49] However, if the sample's thermal relaxation rate is shorter than the interpulse time separation of the laser, the sample does not have enough time to achieve significant thermal expansion and the photoinduced force gradient can be separated from the thermal contribution.[36] As the thermal contribution is proportional to a wavelength-dependent thermal expansion coefficient, we expect there to be a wavenumber dependence to the thermal and photoinduced force contributions in most spectra. Therefore, a superposition of Lorentzian and dispersive line shapes can be anticipated under many common imaging conditions. Specifically, some studies have shown PiFM spectra that match well with those of FTIR in the 1300 - 1800 cm^{-1} window[27,50] and therefore, at least in current understanding, those spectra would be interpreted as having a greater thermal contribution. In Fig. 2.5a-d, the PiFM spectral peak at 823 cm^{-1} is not in agreement with those of FTIR, but appear more dispersive. Therefore, we interpret these peaks as having a greater photoinduced force gradient contribution.[36] The unidentified vibrational mode of PMMA around 800 cm^{-1} and carbonyl stretch around 1734

cm^{-1}, are in good agreement with those of FTIR so we interpret these peaks as having a greater thermal contribution."

To identify structure-function relationships, we take a hyperspectral PiFM image over the range 760 – 1875 cm^{-1} and current image at +1 V sample bias with a gold tip. Prior to any data analysis, we register the hyperspectral PiFM and current images using the affine transformation capabilities in Dipy.[51] Briefly, translation, rigid, and affine transformations are sequentially optimized for the topography obtained in parallel with the cAFM scan. These transformations are obtained by automated maximization of the statistical dependence (as measured by mutual information) between the intensity distributions of the topography images obtained during hyperspectral PiFM and current imaging. Additional details regarding the image registration can be found in the SI and Jupyter notebook online in the associated repository.[52]

We used separate instruments for PiFM and cAFM. In this case, image registration is important because differences due to variations in scan angle, piezo drift, and the different imaging modalities (intermittent contact for PiFM and contact mode in cAFM) can lead to distortions which can mask the sought-after structure-function relationships. We achieved successful co-localization across both microscopy methods with the following protocol: (1) we registered the images optically within several μm using visible alignment marks; (2) we imaged roughly the same region by locating the region of interest via large scans; and (3) we precisely aligned the images post-capture using affine transformations. (See methods and section 2.7.3 for details.)

Figure 2.2 shows the registered integrated hyperspectral PiFM and current images along with the topography. The integrated hyperspectral PiFM image (Fig. 2.2a) is obtained by integrating the area underneath the IR spectrum, spanning 760-1875 cm^{-1} at every pixel (see Figure 2.7 for associated full range spectra at isolated points). Contrast in this image is largely due to the presence

of an intense C=O stretch in PMMA; areas rich in PMMA give a greater integrated signal due to this large peak. Since P3HT is a hole conductor, the domains in the cAFM image (Fig. 2.2b) correspond to hole extraction from the P3HT domains through the grounded tip.[53] As expected, the current reflects the strong contrast in conductivity of this semiconductor-insulator blend, with the current over PMMA being significantly lower than that over P3HT. The topography of the blend (Fig. 2.2c) shows formation of PMMA aggregates (bright regions) within a P3HT matrix (dark background), a natural result of phase segregation within the blend.[54]

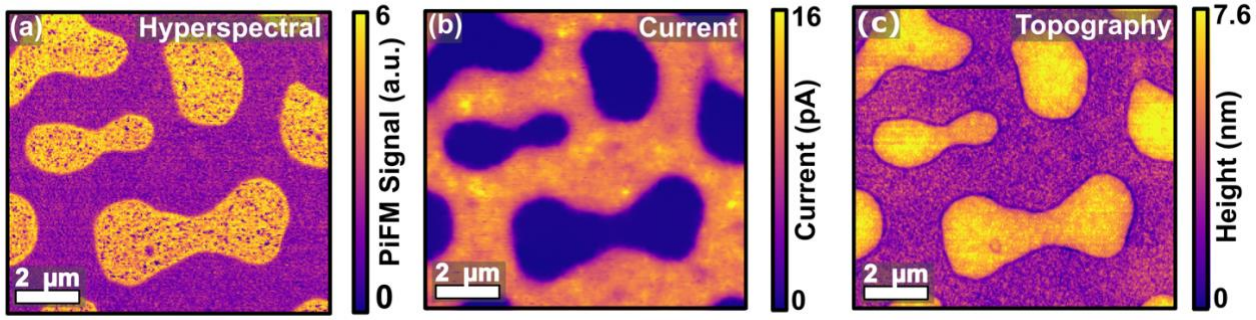

Figure 2.2. Registered multimodal atomic force microscopy data: **(a)** integrated hyperspectral PiFM signal over the range 760-1875 cm^{-1}, **(b)** current under +1 V sample bias, and **(c)** topography images of the same 10 x 10 μm^2 area of a P3HT/PMMA blend. In **(c)** the bright aggregates are PMMA and the dark matrix is P3HT.

We then utilize PCA to extract statistically important components from the hyperspectral image data cube of dimensionality 231 x 247 x 548 (see section 2.7.7 for further information). With PCA, the hyperspectral PiFM image is re-expressed as a linear combination of the principal components (PCs). Often, the variance of the original data captured by the first few principal components is large and quickly drops off for subsequent components, making PCA a useful data reduction technique allowing one to work under the premise that most interesting information is contained

in first few, *K*, principal components and reduce the number of variables we examine from 548 to *K*.[55]

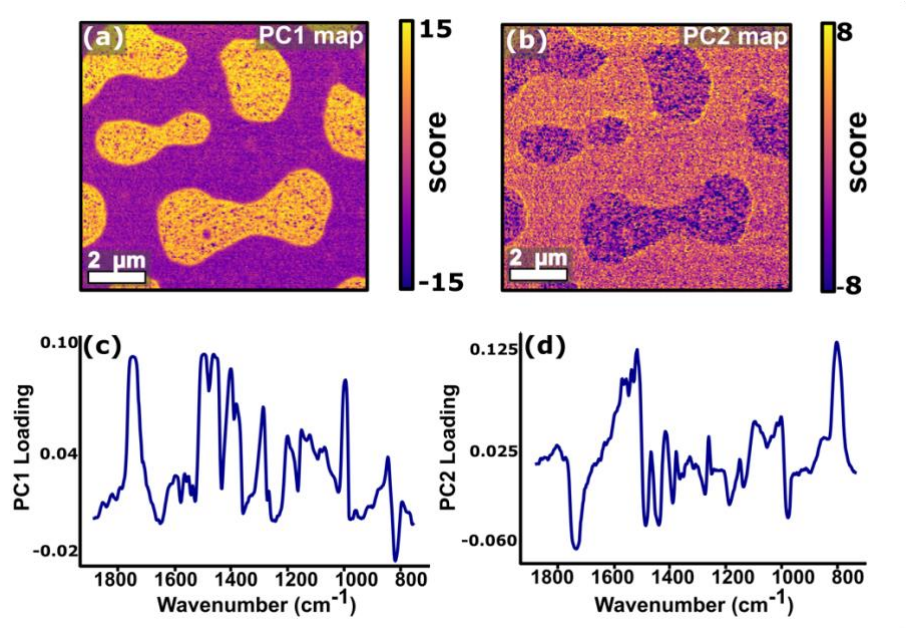

Figure 2.3. Principal component score images and principal components of the hyperspectral PiFM image as infrared (IR) spectra. Loading maps for the **(a)** first and **(b)** second eigenvectors showing that principal component analysis can be used to differentiate the PMMA from P3HT. IR spectra of the **(c)** first and **(d)** second eigenvectors which shows that the principal components are not necessarily physically interpretable in that they maximize variance, but may not correspond to physically observable spectra, a result that is well-known for PCA.

We apply PCA on the hyperspectral PiFM data set.[56] Figure 2.3 shows images of the first two scores along with their respective principal components. The score images (loading maps) (Fig. 2.3a-b) are obtained by plotting the value of each principal component (Fig. 2.3c-d) at each spatial pixel. The resulting score images (Fig. 2.3a-b) can be used to differentiate between polymers as seen in the loading maps, where the contrast in the image is the integrated intensity of the spectral loading at each pixel (Fig. 2.3c-d).

Attempting to decipher a structure-function relationship from the principal component images of hyperspectral PiFM data alone, however, proves difficult. Within the framework of PCA, the loadings define the contributions of the original wavenumbers to the principal components. Loadings large in magnitude at particular wavenumbers signify greater contributions to the principal component. Thus, contrast seen in a score image can be largely attributed to wavenumbers with large loadings in the corresponding principal component. If one desires to interpret the principal components as spectra, neither can be done so in the physical sense, as each has negative absorbance values. This outcome, namely non-intuitive principal components without obvious physical analogs, is a well-known result of PCA.[42,43,57] We highlight it here merely as a reminder that while sometimes PCA of scanning-probe data may appear to decompose into desired physical components,[38,41,58] one should not expect such results to always be the norm.

We note that numerous other techniques for spectral extraction such as, independent component analysis (ICA), nonnegative matrix factorization (NMF), automatic target generation process (ATGP), pixel purity index (PPI), vertex component analysis (VCA), N-FINDR, and numerous others exist.[59–65] Many methods available in software libraries rely on the presence of pure endmember spectra[59–61], while others require more overhead to implement. (See 2.7.6 for additional discussion regarding hyperspectral unmixing.) We plan to compare these approaches in more detail in the future, but here, we focus on what can be achieved with relatively simple and widely-accessible PCA methods.

Turning back to the PCA data, in principal component 1 (Fig. 2.3c), peaks at 1739 cm^{-1}, 1453 cm^{-1},and 1489 cm^{-1} which are known vibrational modes of PMMA[46], have the greatest loadings. In addition, there is a strong negative peak at 823 cm^{-1}, which is spectrally correlated with the P3HT aromatic C-H bending mode.[47] These peaks are therefore primarily responsible for the contrast

observed in the first loading map (Fig. 2.3a). Intuitively, pixels with high principal component 1 scores, will have original spectra with strong absorption peaks at 1793 cm^{-1}, 1453 cm^{-1} and 1489 cm^{-1} corresponding to a C=O stretch, an unassigned vibrational mode, and O(CH_3) bend, respectively, in PMMA. In contrast, pixels with high principal component 1 scores will have original spectra with a weak absorption peak at 823 cm^{-1}. The peaks at 1739 cm^{-1} and 823 cm^{-1} are largely responsible for the contrast generated in the second loading map (Fig. 2.3b) as indicated by principal component 1 (Fig. 2.3d). We interpret principal component 2 therefore, broadly as a lack of PMMA and presence of P3HT.[46,47]

We now turn to extract physical insight by combining PCA data with cAFM information. We use principal component regression (PCR), which uses the principal components as the independent variables in a multiple linear regression model, to understand the relationship between local chemical composition (even without assigning composition fractions or endmember spectra) and electronic function.

To perform the analysis, we randomly selected 10% of the image pixels and fit a multiple linear regression model by regressing the electrical current on the first 10 principal components. With this model, we can reasonably predict the current of the remaining pixels from the hyperspectral information alone as shown in Figure 2.4. (For quantitative measures of error and spatially resolved training and test data sets, see section 2.7.7.) Fig. 2.4a shows that the predicted current is in good overall qualitative agreement with the measured current (Fig. 2.2b). Overall, the PMMA aggregates are predicted to be more insulating than the P3HT components. Importantly, this model also reasonably predicts the current images for different image sets of the same composition polymer blend. (See Section 2.7.7 for more information.)

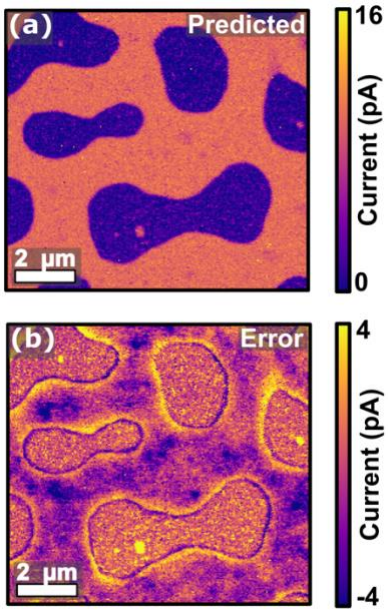

Figure 2.4. Current as predicted with a principal component regression model obtained by regressing the electrical current onto the first ten principal components. **(b)** Error image obtained by subtracting the real current, shown in Figure 1b, from the predicted current.

Figure 2.4b shows the error image (measured current – predicted current), from the principal component regression. We speculate that regions where the PCR model under predicts the current could be regions of better ordered P3HT with higher mobility.[66–68] Interestingly, this image contains additional insight: regions that diverge most from the PCR model inform the identification of the polymer and appear to reveal tomographic structure indiscernible from PiFM or cAFM alone. Figure 2.5a shows the integrated hyperspectral PiFM image with the spatial distribution of the pixels (in yellow) *where the PCR model over-predicts the real current by more 3 standard deviations* (See 2.7.8 for statistical rationale). Upon closer inspection, we see these pixels reside within regions that have negligible electrical current in cAFM, yet where the PCR model predicts they are P3HT. The average spectrum, reconstructed from the first 10 principal components, for these pixels, denoted P3HT*, along with that of the PMMA aggregates, and P3HT matrix are shown in the spectral windows corresponding to the distinct vibrational modes of P3HT and PMMA in Figure 2.5b and c, respectively (see Figure 2.13 for masks used to obtain average

spectra). The average spectrum for these pixels is similar to that of the P3HT matrix as it exhibits a negative peak at 1733 cm^{-1} (Fig. 2.5c), distinct from the PMMA aggregates. In addition, it also has a peak at 823 cm^{-1}; the weaker intensity is consistent with these regions being within PMMA.

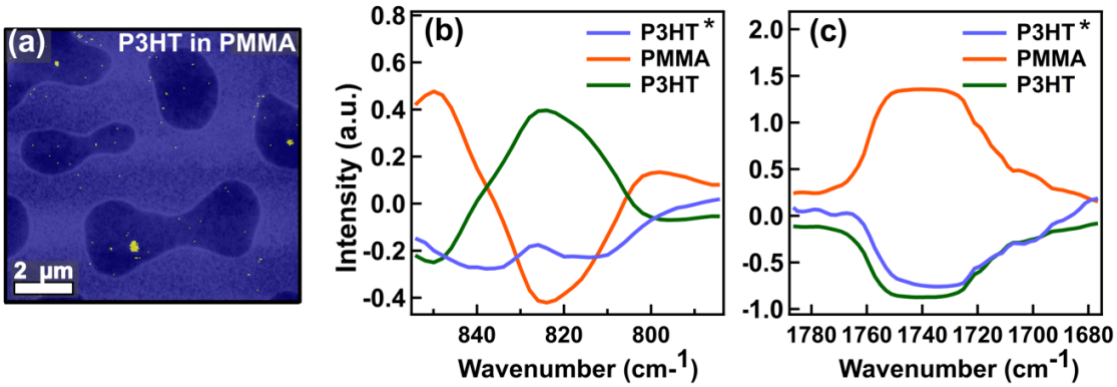

Figure 2.5. (a) Regions with predicted current that deviate more than 3 standard deviations from the real current (yellow pixels) overlaid on the hyperspectral PiFM image. **(b)** The average spectrum, reconstructed from the first 10 principal components only, of pixels in **(a)** are in blue, PMMA aggregates in orange, and P3HT matrix in green for the range 850-780 cm^{-1} and **(c)** 1670-1790 cm^{-1}. The regions highlighted in **(a)** are likely to be P3HT regions within the PMMA aggregates that are electrically insulated.

This interpretation, combined with the presence of a vibrational mode at 823 cm^{-1} suggests that these regions are likely P3HT regions trapped within the PMMA aggregates. The advantage of the PCA/PCR approach as shown in Fig. 2.5 is clear when compared to the seemingly straightforward approach of using the raw hyperspectral PiFM signal at 823 cm^{-1} to correlate with the cAFM current image. The result (see Fig. 2.14) is less clear than the PCA/PCR approach because of the noise at 823 cm^{-1}, while PCA allows more use of the acquired information, while also denoising the data. The pixels shown in Fig. 2.5a exhibit the spectral signatures of P3HT regions, but lack an electrically-connected pathway for hole transport through the entire film. Also notable is that these regions (Fig. 2.5a) are *not* directly correlated with regions of lower integrated hyperspectral

PiFM intensity (see Fig. 2.15) in the same region as might be expected if this information were contained within the top few nanometers accessible to PiFM.[69] The error from PCR, combined with our *a priori* knowledge about the semiconducting and insulating nature of the two polymers suggests an explanation for why P3HT aggregates within PMMA do not appear in the cAFM image: it is due to the lack of an electrically connected pathway to the extracting contact. Such inferences are impossible with either the cAFM or PiFM alone and reveal a level of tomography (sub-surface structure) that is unique to a data-driven approach to scanning probe regression of electrical information on chemical nanostructure.

2.5 CONCLUSION

Herein, we have acquired hyperspectral PiFM and cAFM images of the same area within a PMMA/P3HT blend film. We show that while the local PiFM spectra do not necessarily match the macroscopic FTIR, they are still useful in differentiating between the polymers. We described a representative workflow to register and analyze hyperspectral PiFM and cAFM images with PCA and PCR. Using these statistical tools, we demonstrate that the two techniques provide complementary information that when combined, provide additional insight into the tomography of P3HT/PMMA that neither provides in isolation. While the PCR model yields a reasonable prediction of the current image, we note that these results show there could be additional value in moving beyond basic PCA approaches in the future. Furthermore, to improve predictions of the PCR model, it is likely that nonlinear techniques such as multiple adaptive regression splines (MARS)[70], which can flexibly model more sophisticated relationships by including nonlinearity and interaction terms, will be necessary. Incorporating prior information about systems can also improve the results of dimensionality reduction and regression; therefore, these techniques may be more appropriate when applied within a Bayesian framework.[71]

2.6 ACKNOWLEDGEMENTS

This paper is primarily based on work supported by the Department of Energy BES under award number DESC0013957. J.K. is supported in part by the Clean Energy Institute and National Science Foundation Research Traineeship under award NSF DGE-1633216. We thank Ariel Rokem (The University of Washington eScience Institute) for help with image registration. We thank Daniel J. Graham (University of Washington) for helpful discussions regarding PCA. We also thank Derek B. Nowak (Molecular Vista) and William A. Morrison (Molecular Vista) for help with PiFM. Part of this work was conducted with instrumentation supported by the University of Washington Student Technology Fee at the Molecular Analysis Facility, a National Nanotechnology Coordinated Infrastructure site at the University of Washington, which is a user facility supported in part by the National Science Foundation (Grant ECC-1542101), the University of Washington, the Molecular Engineering & Sciences Institute, the Clean Energy Institute, and the National Institutes of Health.

2.7 SUPPLEMENTARY INFORMATION

2.7.1 *Graphical Overview of Workflow*

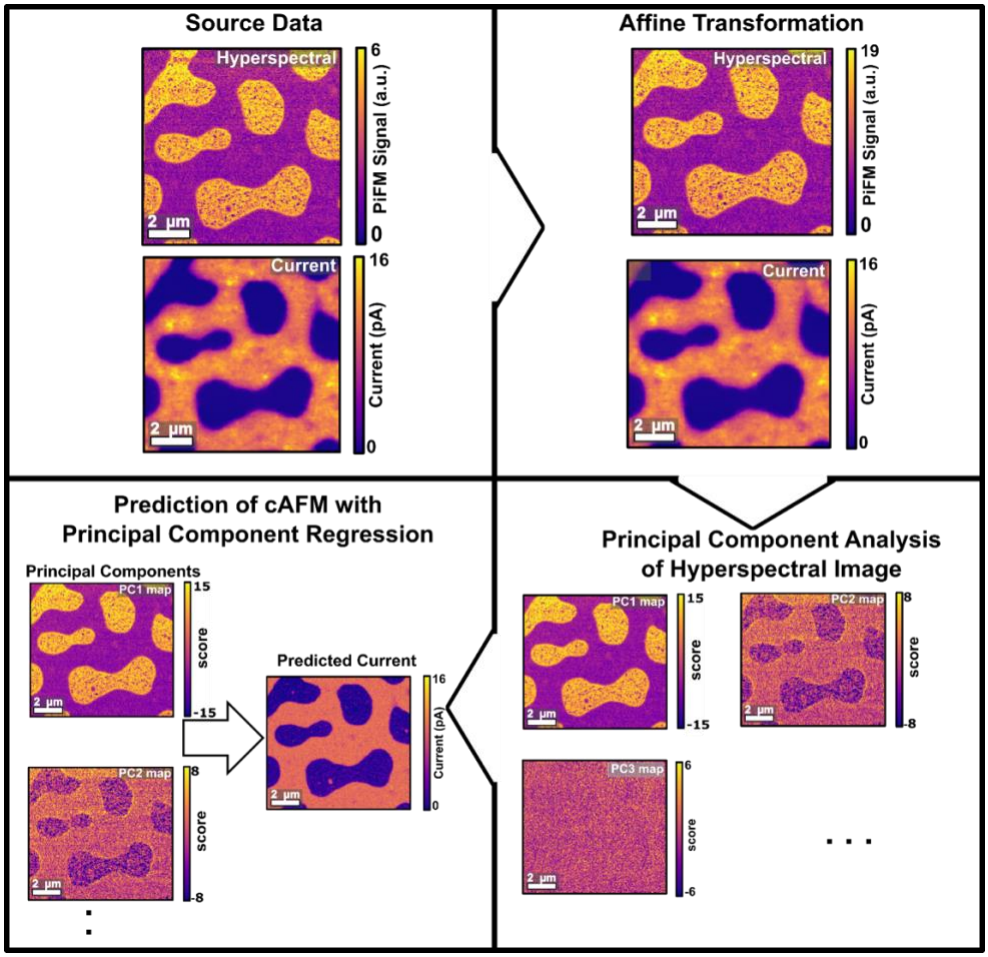

Figure 2.6. Overview of work flow. The hyperspectral photodinduced force microscopy (PiFM) and conductive atomic force microscopy (cAFM) images are aligned with affine transformations prior to data analysis. Principal component analysis (PCA) is then applied to the hyperspectral PiFM image to extract statistically significant principal components (PCs). Principal component regression (PCR) where the electrical current is regressed onto the first ten principal components is used to develop a model that quantifies their relationship with the current.

2.7.2 *Full Range FTIR and PiFM Spectra*

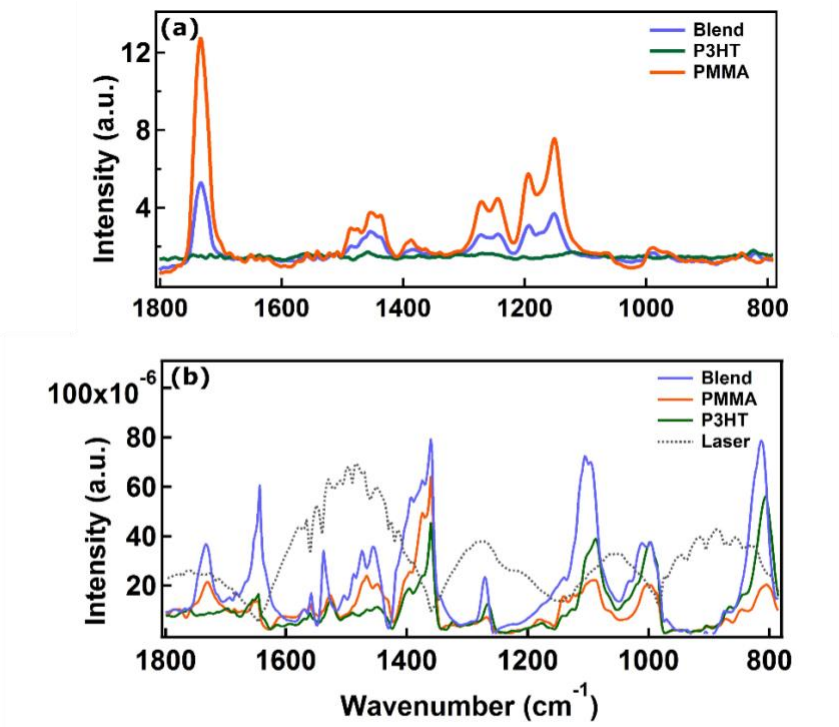

Figure 2.7. (a) Attenuated total reflectance Fourier-transform infrared (FTIR) and **(b)** photoinduced force microscopy (PiFM) point spectra of neat PMMA (orange traces), neat P3HT (green traces), and blend (blue traces) films along with the laser profile of the QCL (grey, dashed trace) in the region 800 to 1800 cm^{-1}.

2.7.3 *Image Registration*

The hyperspectral PiFM image was obtained in an area located near a visible “X” as described in the methods section. After imaging, the tip was used to scratch a 17 µm box around the region imaged as shown in Fig. 2.8a. The region of interest was located following the procedure described in the methods section. The raw, unregistered topographies obtained in hyperspectral and conductive AFM imaging are shown in Fig. 2.8b and c, respectively.

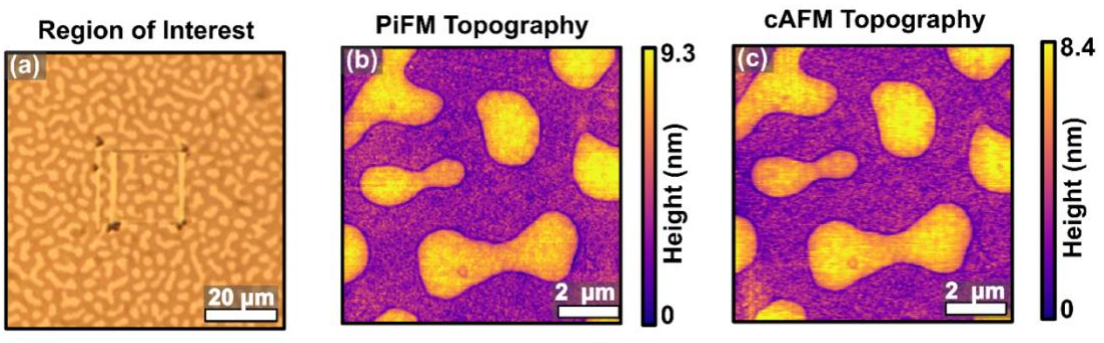

Figure 2.8. (a) Optical microscopy image of the 17 μm box used for locating the region of interest. Unregistered topography scans obtained during **(b)** hyperspectral PiFM imaging and **(c)** conductive AFM imaging.

A multiresolution approach using affine transformations with mutual information (MI) as a similarity metric was employed to register the current and hyperspectral PiFM images. Image registration was performed in Dipy[51], following a package-specific tutorial.[72] An optimized affine transformation was found with the corresponding topography images and applied to the current image to register it with the hyperspectral PiFM image. To calculate the MI, the joint and marginal probability distribution functions were estimated using the Parzen-window method.[73] Distributions were computed using all pixels and discretized to 32 bins. A multi-resolution strategy featuring a three-level Gaussian Pyramid containing sub-sampled topographies at quarter, half, and full resolutions was used. The coarsest, medium, and finest resolutions were Gaussian-smoothed with $\sigma = 3$, 1, and 0, respectively. At each of the aforementioned resolutions, 10,000, 1,000, and 100 optimization iterations were performed. The registration was refined in three stages where the translation, rigid, and affine transformations were sequentially optimized using results from the prior stage. A limited memory Broyden-Fletcher-Goldfarb-Shanno box (L-BFGS-B) optimization algorithm was used. After registration, the images were cropped to a common area of spatial dimensions 231 x 247 pixels.

Mutual information (MI) measures the statistical dependence between image intensities of corresponding pixels in two images, A and B.[2] When aligned, the statistical dependence is measured by the Kullback-Leibler distance[74] is assumed to be a maximal.

2.7.4 *PiFM Point Spectra*

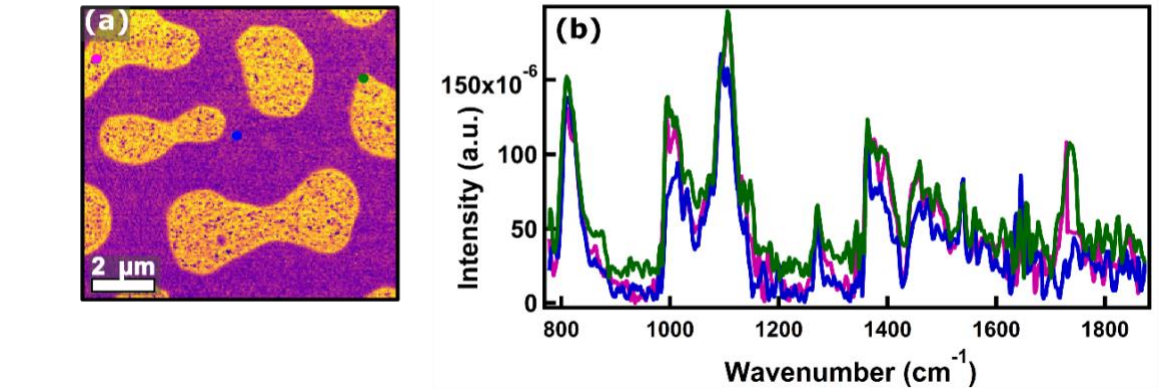

Figure 2.9. (a) Various points in the region imaged and **(b)** their associated photoinduced force microscopy point spectra in the range 775-1875 cm^{-1}.

2.7.5 *Principal Component Analysis*

Principal component analysis was performed with sci-kit learn using singular value decomposition after correcting for laser power. As the laser profile was unreliable in the range 775 to 760 cm^{-1} and this section of the spectra does not contain relevant spectroscopic information, this range was removed from the hyperspectral data cube. The intensities at each wave number were mean centered prior to PCA to ensure that the first principal component was not simply the average spectrum across all pixels. The data were also scaled to unit variance.

2.7.6 *Hyperspectral Unmixing*

Hyperspectral unmixing aims to decompose the hyperspectral PiFM image into constituent endmember spectra and determines fractional abundances of each at every pixel. However, the

techniques listed in the main text, when used out of the box, with standard inputs, do not produce desired endmember spectra. Figure 2.10 shows the two resulting endmember spectra from nonnegative matrix factorization (NMF), independent component analysis (ICA), vertex component analysis (VCA), automatic target generation process (ATGP), NFINDR, and pixel purity index (PPI) when default arguments are used in the extraction methods. The resulting spectra either are not physically meaningful or unable to separate the characteristic vibrational modes of P3HT and PMMA at 823 cm^{-1} and 1735 cm^{-1}, respectively.

The results from NMF and ICA were obtained with scipy version 1.0.0 in Python 3.5 using default arguments with two endmembers specified. Figure 2.10a shows the resulting two endmember spectra. Like PCA, NMF is a matrix decomposition technique but is constrained by non-negativity which often enables physically meaningful interpretation of components.[63,75] In this case, however, NMF gives spectral components that contain peaks that are concave up which is not observable physically. Figure 2.10b shows the component spectra obtained from ICA which assumes statistical independence between each component, a condition we do not expect to be true as the presence of PMMA indicates the absence of P3HT in this phase segregated polymer blend. The extracted components from ICA are able to separate the defining vibrational modes of P3HT and PMMA but contain negative values as can be seen in Figure S5b and therefore, do not align with physical reality.

More advanced hyperspectral unmixing techniques such as vertex component analysis (VCA), automatic target generation process (ATGP), N-FINDR, and pixel purity index (PPI) assume the presence of a pure endmember in the image and extract basis spectra without any *a priori* information. VCA was implemented following the method described in Nascimento and Dias[76]

using code by Laadr on Github with two endmembers specified.[77] The results from N-FINDR, ATGP, and PPI were obtained with pysptools version 0.14.2 in Python 3.5.

When default arguments with two endmembers specified, these techniques do not appear to separate the defining features of P3HT and PMMA. VCA finds endmembers by iteratively projecting data onto a spectrum orthogonal to the subspace already spanned by determined endmembers and stops when the specified number of basis spectra are found.[76] Fig. 2.10c shows the extracted endmembers from VCA; we see that this method gives spectra with negative values and cannot better differentiate between the vibrational modes corresponding to P3HT and PMMA compared to ICA. ATGP extracts endmembers one-by-one by projecting pixels onto orthogonal basis spectra to identify spectrally distinct pixels.[62] The resulting spectra (Fig. 2.10d), while containing some peaks with similar shapes as those acquired in PiFM (Fig 2.7b), still do not match the spectra of neat P3HT or PMMA. N-FINDR assumes that each pixel is a linear combination of endmember spectra and simplistically, searches for the purest pixel in a given hyperspectral image. The resulting spectra (Fig. 2.10e) from this technique are similar to ATGP in that they capture some, but not all, peak positions and shapes. PPI, like N-FINDR, also searches for the purest spectral signatures in a hyperspectral image.[78] The extracted basis spectra (Fig. 2.10f), while not perfect, are distinct in the 1730 cm^{-1} region. These spectra are appear similar to those from ATGP and N-FINDR.

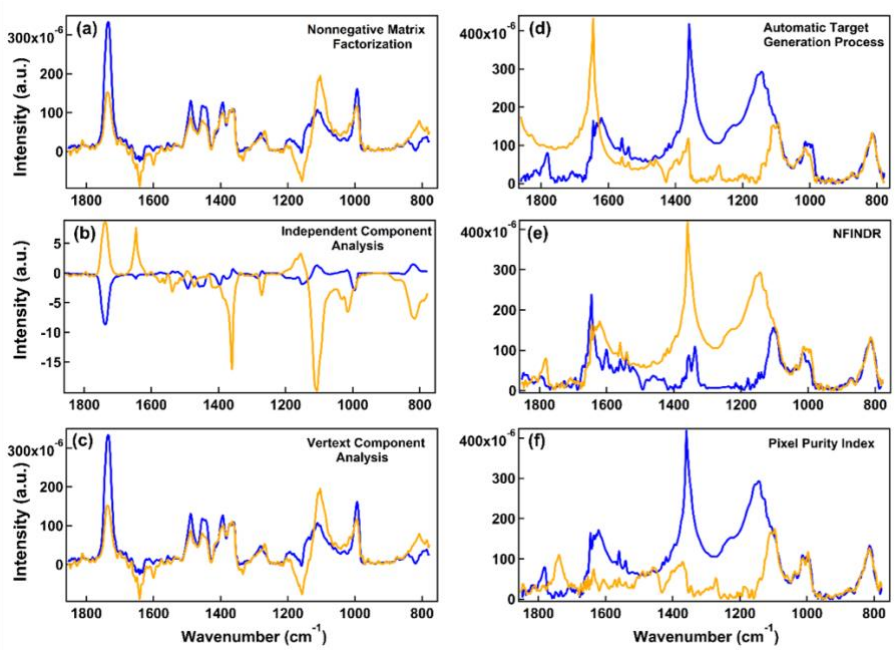

Figure 2.10. Extracted endmember components from **(a)** nonnegative matrix factorization (NMF), **(b)** independent component analysis (ICA), **(c)** vertex component analysis (VCA), **(d)** automatic target generation process (ATGP), **(e)** NFINDR, and **(f)** pixel purity index (PPI). These methods used out of the box either do no return physically meaningful spectra or are unable to separate the characteristic of P3HT from PMMA in the extracted endmembers.

2.7.7 *Principal Component Regression*

Regressing the first ten principal components of 10% of randomly selected pixels (see Fig. 2.11) to the electrical current gave a regression equation of:

$$Current\ (pA) = 6.6 - 4.3 \cdot 10^{-1}X_1 + 3.5 \cdot 10^{-1}X_2 - 2.9 \cdot 10^{-2}X_3 - 1.6 \cdot 10^{-1}X_4 - 4.1 \cdot 10^{-2}X_5 - 5.5 \cdot 10^{-2}X_6 + 8.3 \cdot 10^{-2}X_7 + 3.3 \cdot 10^{-2}X_8 + 1.5 \cdot 10^{-2}X_9 + 8.4 \cdot 10^{-2}X_{10} + \epsilon$$

where X_i denotes the *i*th principal component. By several quantitative measures, the extent to which the model fits the data, while not perfect, is reasonable. For linear regression models, the fit is usually evaluated with two related metrics: the residual standard error (RSE) and R-squared statistic (R^2). The (RSE) is an estimate of the standard deviation of ϵ in the model.[79] Here, an RSE of 1.7 pA indicates on average, the PCR model will give a predicted current that deviates 1.7 pA from the true current. It was computed with the test data using the formula:

$$RSE = \sqrt{\frac{1}{n-2}RSS}$$

where RSS is the residual sum of squares and is given by:

$$RSS = \sum_{i=1}^{n}(current_{measured} - current_{predicted})^2$$

The R^2 statistic is a value between 0 and 1 that indicates the fraction of variance explained by the model.[79] Here, the R^2 statistic of 0.85 indicates the PCR model explains 85% of the variance in the current and suggests that model is an approximation of the data. It was calculated using the following equation:

$$R^2 = 1 - \frac{RSS}{TSS}$$

where TSS is the total sum of squares given by:

$$\sum_{i=1}^{n}(current_{measured} - current_{average})^2$$

In addition, the F-statistic can be used to test whether there is a relationship between the current and principal components.[79] If there is a relationship between the current and one of the principal components, the F-statistic is expected to be larger than 1. It can be calculated using:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

Here, an F-statistic of 154955 with associated p-value of 1.1×10^{-16} (obtained by summing the tail of the probability density function of all F-statistic values greater than 154955) signifies that there is a relationship between at least one of the principal components and the current.

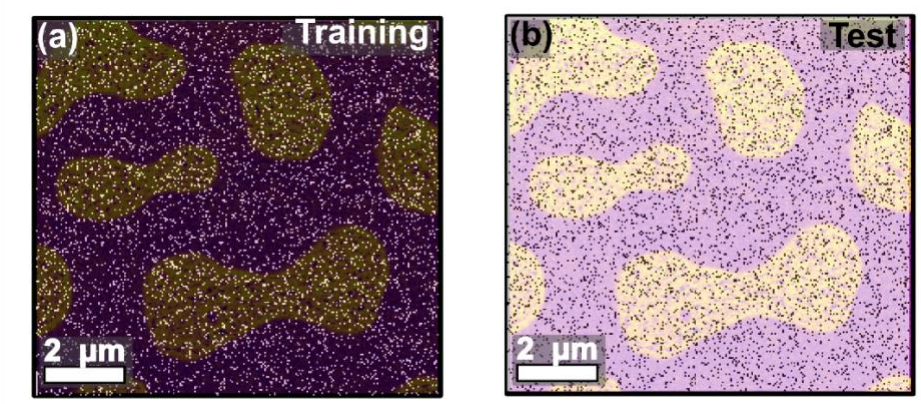

Figure 2.11. Randomly selected pixels (white) overlaid on the integrated hyperspectral PiFM image used to generate the **(a)** training and **(b)** test data sets.

Importantly, the model can use the principal components of a different sample and predict its current in a qualitatively consistent manner. We take a hyperspectral PiFM image of a second sample as shown in Figure 2.12a, apply the same dimensionality reduction on the data, and predict the current using the model described in the main text. Similar to the hyperspectral PiFM shown in the main text, contrast in this image is largely due to the presence of a carbonyl stretch in PMMA and its absence in P3HT. The pattern observed over the PMMA aggregates are a result of patches of P3HT over the aggregates. Figure S6b shows the measured current image, obtained by applying a +1V bias between the tip and sample of the same area shown in Figure 2.12a. These images were aligned using the same affine transformation procedure as described for the data set in the main text. The predicted current is shown in Figure 2.12c and the error image obtained by subtracting the measured current from the predicted current is shown in Figure 2.12d. Overall, bright regions in the hyperspectral PIFM image (PMMA aggregates), map to a low predicted currents and darker regions (P3HT) map to high predicted currents, consistent with the data set in the main text and with the materials' expected electrical properties. The model gives, on average, a prediction that deviates 83 pA from the measured current. This metric is unsurprising as the results of PCA, and therefore the predicted current, are affected significantly by the magnitude of the spectra in the

hyperspectral PiFM image which can vary across images due to differences across tips and in laser power. The result of applying this model on a different sample shows that the qualitative results and physical model obtained are generalizable.

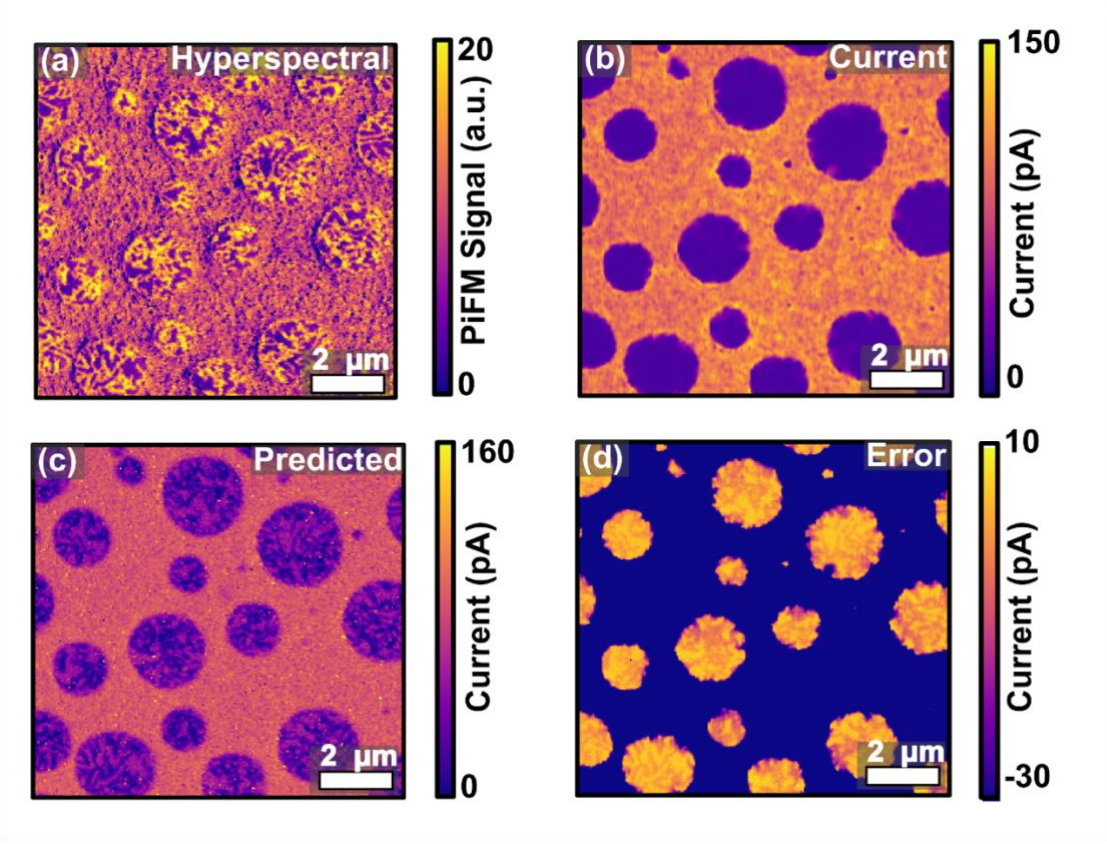

Figure 2.12. (a) Hyperspectral photoinduced force image of a polymer blend comprising poly(methyl methacrylate) (PMMA) and poly(3-hexylthiophene) (P3HT). **(b)** Current measured by applying a surface bias of +1V of the same area. **(c)** Predicted current from applying principal component analysis and regression obtained with the image set in the main text, showing that the model holds qualitatively. **(e)** Error image obtained from subtracting the measured from predicted current.

2.7.8 *Selection of Pixels in Error Analysis*

The distribution of error values in current, I, (error = I(measured) – I(predicted)) was roughly normal, with 68.3%, 96.5%, and 99.7% of the values lying within one, two, and three standard

deviations of the mean. We chose pixels with error greater than three standard deviations (5.3 pA) more than the mean to analyze as they were poorly fit by the regression model.

2.7.9 *P3HT and PMMA Masks*

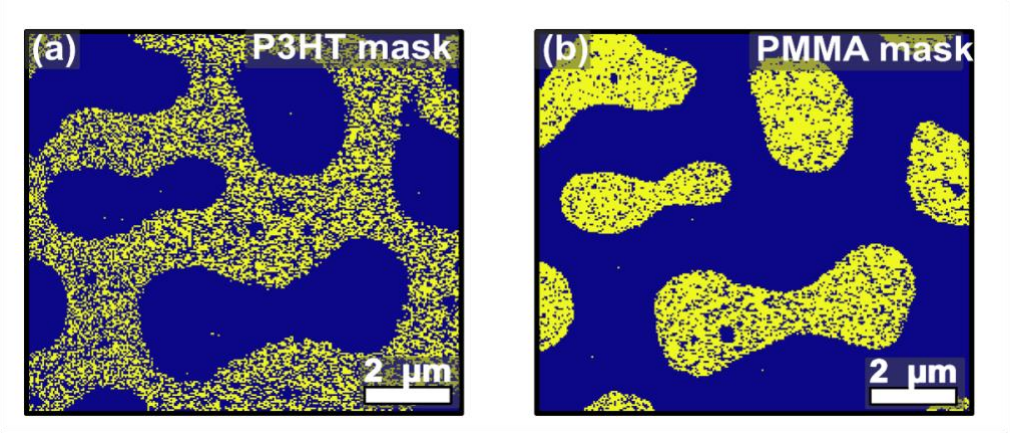

Figure 2.13. Pixels within the **(a)** P3HT matrix and **(b)** PMMA aggregates, both in orange, for which the average reconstructed spectra in Fig. 2.5 was obtained.

2.7.10 *Integrated Hyperspectral Intensity at 823 cm^{-1}*

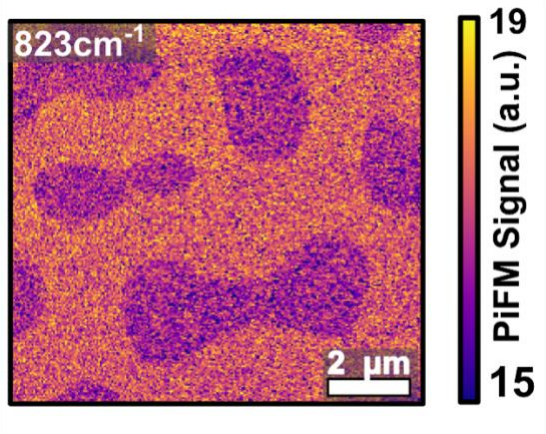

Figure 2.14. Integrated hyperspectral intensity about the vibrational mode corresponding to a C-H out-of-plane bend in P3HT.

2.7.11 *Pixels with Low Hyperspectral Intensity and Low Current*

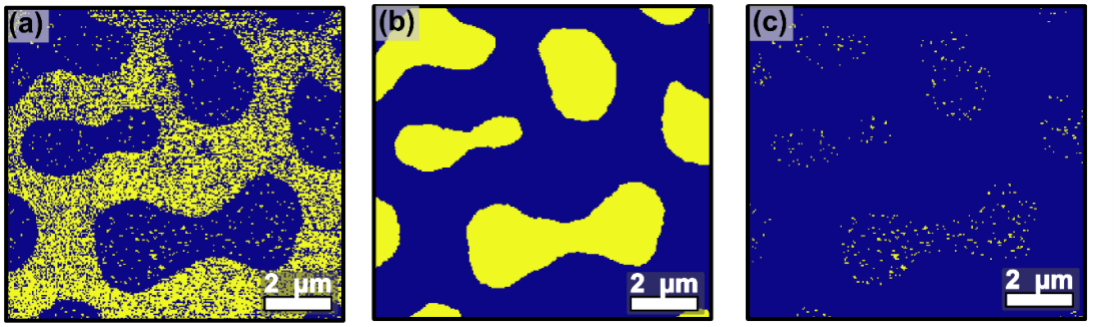

Figure 2.15. In yellow are: **(a)** pixels with integrated hyperspectral intensity (<0.2 a.u.) which selects for P3HT; **(b)** pixels with low measured current (<2 pA) which selects for PMMA; and **(c)** pixels with low integrated hyperspectral intensity and low predicted current (<2 pA). Simply selecting for pixels with low integrated hyperspectral intensity and low predicted current does not provide the same result shown in Fig. 5b in the main text.

Chapter 3. CHARGE RELAXATION IN PIEZORESPONSE FORCE MICROSCOPY AND ITS IMPACT ON CONTACT KELVIN PROBE FORCE MICROSCOPY[2]

3.1 ABSTRACT

Piezoresponse force microscopy (PFM) and associated spectroscopies are a key tool in the study of nanoscale ferroelectrics, ionic systems, and complex electromechanical behaviors in soft matter and biological systems. Yet, the origin of the electromechanical response even in simple inorganic materials can be nonsingular, confounding attempts to interpret results in terms of physical material parameters. This often leads to purported observations of ferroelectric behavior in nanoscale systems of bulk non-ferroelectric materials. Here, we investigate the origins of relaxation in PFM, considering the dynamics of surface charges and explore how this influences the measured electromechanical response. We perform a controlled study by measuring the relaxation of the electromechanical response in non-ferroelectric amorphous HfO_2 and a ferroelectric $BaTiO_3$ thin film as a function of reading voltage. These results are then used to explain a commonly observed concavity in contact Kelvin probe force microscopy measurements. This study has implications for the interpretation of all piezoresponse force spectroscopies, and points towards the need for time-dependent measurement of the electrostatic contribution to the measured electromechanical response.

[2]Adapted with permission from Kong, J.; Kelley, K.P.; Collins, L.F.; Kalinin, S.V.; Balke, N.; Vasudevan, R.K.

3.2 INTRODUCTION

Piezoresponse force microscopy (PFM) has been an in invaluable tool for imaging and measurement of piezo and ferroelectric samples for more than 20 years.[80] Its utility stems from the wide variety of samples that can be probed from biological samples[81] to ceramics[82] and nanostructures.[83] In addition, the ease of sample preparation for PFM along with its ability to image ferroelectric domains with sub 10-nm resolution make it a powerful, accessible tool.[84] Of equal importance are the various spectroscopic modes of PFM[85] that enable the correlation of functional with microstructural features such as domains, dislocations, and grain boundaries.[86] However, despite its widespread use, the signal interpretation in PFM remains a proverbial thorn, leading to many anomalous reports of ferroelectric behaviors in nanoscale systems that otherwise show no macroscopic evidence of ferroelectricity.[87–93]

In PFM, the electromechanical response is defined as the measured deflection signal from the standard atomic force microscope (AFM) setup and can stem from multiple physical origins including piezoelectricity and electrostatics. The electrostatic contribution can arise from a variety of phenomena including charge injection and ion migration as illustrated in Fig. 3.1a. This contribution to the overall signal comes from the difference in the cantilever and surface potentials and has largely been studied in the context of Kelvin probe force microscopy (KPFM).[87,88,94] The magnitude of the electrostatic contribution (D_{ES}) is traditionally approximated by $k^{*\,-1}C'(V_{DC} - V_{SP})V_{AC}$ where k^* is the effective contact stiffness, V_{DC} the applied DC potential, V_{AC} the AC voltage applied to measure the electromechanical response, C' the effective capacitance gradient, and V_{SP} the surface potential.

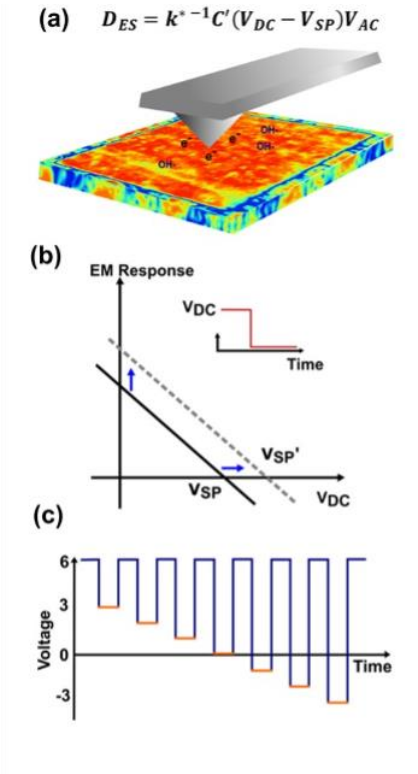

Figure 3.1. (a) An AFM tip, when biased, can cause charge injection and/or migration of ions, that can than affect the subsequent electrostatic response. **(b)** Schematic of the KPFM response for a non-ferroelectric sample, where the bias perturbs the local surface potential and therefore affects the measured response in time. When the magnitude of the applied DC is equal to the new local surface potential, the relaxation, assuming that local effects dominate, is annulled. **(c)** Outline of the DC waveform applied to the tip; measurements are taken at each time step, but the reading voltage is varied while the writing voltage is kept constant.

The surface potential generally varies across the sample and will therefore be a function of (x,y). Biases applied to the tip are only expected to change the V_{SP} in a small region near the location of the tip, (x_i, y_i). If we assume a constant V_{AC} = 1 V, then measuring the response as a function of V_{DC} should result in an electrostatic contribution to the electromechanical response that is linear (in the absence of all other mechanisms). An example is shown in Fig. 3.1b. – when V_{DC} is equal to the surface potential, the magnitude of the electrostatic response goes to zero.[95] This is indeed observed, and the measurement of this response with respect to applied 'writing' voltages in

contact (usually used for probing ferroelectric hysteresis loops) constitutes the technique called contact Kelvin Probe Force microscopy (cKPFM).[88,96,97]

As a foundation, the above discussion simplistically assumes that the electrostatic response D_{ES} is time independent. However, changes in the electromechanical response with respect to time known colloquially as relaxation is often observed in both PFM[98–100] and KPFM; this phenomenon is the basis for techniques like time-dependent KPFM.[101] While relaxations in PFM can have a ferroelectric origin like the motion of domain walls within a probed volume, this phenomenon can also be a result of charge relaxation.[102] For example, it has been shown that even on soda-lime glass, hysteretic electrostatic response can be obtained when the tip is not in contact with the sample at distances tens to hundreds of microns above the sample, further obfuscating the origins of the measured electromechanical response.[103] We cite these examples to show that an electromechanical response can be obtained in both ferroelectric and non-ferroelectric materials and originate from a superposition of electrostatic and piezoelectric properties but that it is difficult to isolate the contributions of each.

Understanding the origins of relaxation in the measured signal is important for assigning mechanisms from relaxation spectroscopy data and subsequent materials characterization. For example, when relaxation times are slow (~>100 ms) traditional hysteresis loop acquisition will average over a time-dependent signal and thus include the dynamic component as part of 'stationary' loop data. In PFM spectroscopy, the average measurement time for a hysteresis loop occurs within 1 s, which means that charging phenomena can significantly contribute to the measured dynamic responses. Therefore, understanding relaxation is crucial even for quasi-static PFM spectroscopies which assume time-independence (i.e. fast relaxations with respect to measurement time). For PFM imaging, the role of charging phenomena can be more complex, as

imaging times far exceed those which are required to obtain spectra. We pose that relaxation times for oxide surfaces can differ by many orders of magnitude, from well below seconds to tens of hours.[104–106] As such, these phenomena are of fundamental interest for both PFM imaging and spectroscopy.

3.3 Experimental Design and Methods

To experimentally investigate the origin of relaxation in the electromechanical response measured by PFM**,** we apply a DC write voltage to the tip and measure the electromechanical response as a function of time *and* DC read voltage in two exemplary materials: non-ferroelectric, amorphous HfO_2 and ferroelectric $BaTiO_3$ films using the band excitation (BE) technique applied to scanning probe microscopy.[107] In BE, the frequency-dependent tip deflection signal around the contact resonance is measured and subsequently fit to a single harmonic oscillator model to extract the amplitude. The electromechanical response is calculated via $Acos\theta$ where θ is the measured phase and A is the fitted amplitude.

We operate under reasonable assumptions and specific experimental conditions that mitigate the complexity of the PFM signal. It is known that the DC write pulse, $V_{WRITE,}$ will change the charge state of the surface or tip-surface junction (see Figure 3.1a), regardless of whether the material is ferroelectric or not. With regards to this, we assume that a new electrochemical potential equilibrium is established and the effect of charging, due to the DC write pulse, can be represented as a change in local surface potential V_{SP} ($V_{SP} \rightarrow V_{SP'}$), effectively shifting the response line in Fig. 1b vertically. We note that this field can also cause local migration of pre-existing surface charges including hydroxyl ions, as shown in Fig. 3.1a.

In addition, relaxation in ferroelectric materials is expected to be complex. Motion of ferroelectric domain walls changes the volume fraction of switched domains, thereby affecting the polarization

in the probed volume and ultimately, the measured piezoresponse. To address this, we measure the relaxation with V_{READ} and V_{WRITE} aligned with the polarization direction of the domain being investigated such that the ferroelectric contributions can be effectively negated. We note that our approach is dichotomic with dynamic switching spectroscopy piezoresponse force microscopy (D-SSPFM)[100] which induces switching of ferroelectric domains and includes the contribution of nanodomains also contributes to the measured signal. Our alignment of both V_{READ} and V_{WRITE} in the direction of the polarization circumvents this concern. While separation of response from relaxation curves on ferroelectric materials with complex domain structures is beyond the scope of this report, a promising avenue is using the interferometric sensing to effectively annul the electrostatic response.[103,108] However, since the piezocoefficient d_{ij} is not necessarily independent of V_{DC} within a domain (e.g., if there is a field-induced phase transition[109,110]), we consider the dynamics of screening charges.

We operate under reasonable assumptions and design our experiments to mitigate the complexity of the PFM signal. Relaxation in ferroelectric materials is expected to be complex; motion of ferroelectric domain walls change the volume fraction of switched domains, thereby affecting the polarization in the probed volume and subsequently, the measured piezoresponse. To address this, we measure the relaxation with V_{READ} and V_{WRITE} aligned with the polarization direction of the domain being investigated such that the ferroelectric contributions can be effectively negated. We note that our approach is dichotomic with dynamic switching spectroscopy piezoresponse force microscopy (D-SSPFM)[100] which induces switching of ferroelectric domains and includes the contribution of nanodomains also contributes to the measured signal. Our alignment of both V_{READ} and V_{WRITE} in the direction of the polarization circumvents this concern. While separation of response from relaxation curves on ferroelectric materials with complex domain structures is

beyond the scope of this report, a promising avenue is using the interferometric sensing to effectively annul the electrostatic response.[103,108] However, since the piezocoefficient d_{ij} is not necessarily independent of V_{DC} within a domain (e.g., if there is a field-induced phase transition[109,110]), we consider the dynamics of screening charges.

To modify the relaxation, we apply a constant write voltage but measure the electromechanical response as a function of DC read voltage as represented in orange in the waveform in Fig. 3.1c. Measurements were performed over a 5 x 5 grid with the electric field in the same direction during V_{WRITE} and V_{READ} steps. (See SI for additional experimental details.)

3.4 Results and Discussion

The averaged relaxation traces as a function of read voltages for HfO_2 after -6 V_{WRITE} and +6 V_{WRITE} are shown in Fig. 3.2a and b, respectively. For both, as V_{READ} moves away from V_{WRITE}, the absolute value of the response decreases. To analyze the relaxations quantitatively, we fit the curves to a dual exponential fit of the form $A_1 * \exp(-t/\tau_1) + A_2 * \exp(-t/\tau_2) + c$; the model was selected over the stretched single exponential models given a superior Akaike information criterion[111] score for most measured pixels (see Fig. 3.6). Figure 3.2c shows A_0, the first relaxing amplitude in a dual exponential fit. While a linear trend is expected with respect to V_{READ}, we suspect that the relaxing amplitude shows a weakly linear trend due to the dual exponential fit. For HfO_2, the signal is expected to be purely electrostatic in origin which is corroborated by the equal spacing of relaxation curves. The time constants, plotted in Fig. 3.2d show a weak dependence on the read voltages. These data validate that relaxation, due to electrostatics, occurs and can contribute to measurements done in short time windows that capture the relaxation phenomenon.

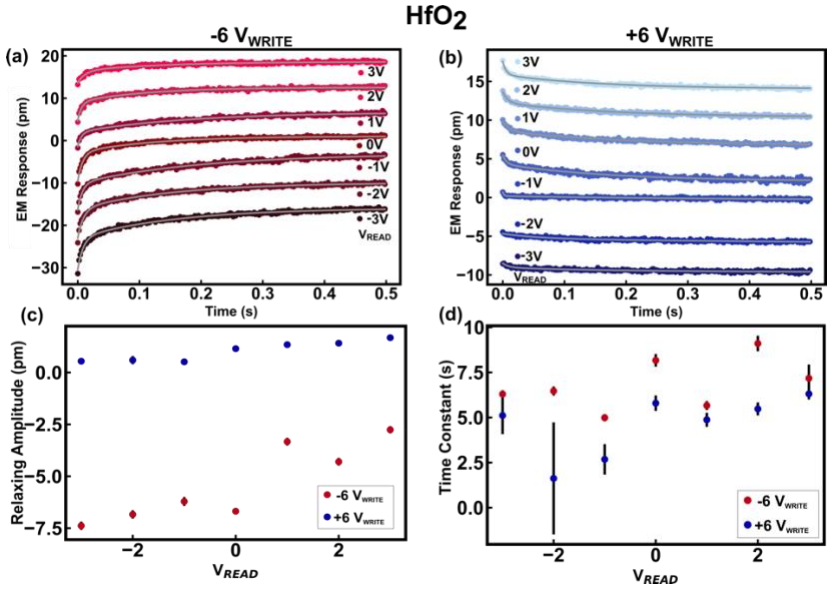

Figure 3.2. (a) Relaxation response as a function of read voltages for -6 V and **(b)** +6 V. **(c)** Fitted relaxing amplitude for both write voltages, where the relaxing amplitude is first (out of two) for the dual exponential fit. Error in fit parameter is smaller than marker size. **(d)** Time constant for both write voltages, where the τ plotted is τ_0 from the dual exponential fit.

Next, we perform the same measurement for ferroelectric $BaTiO_3$ and compare the relaxation curves to those for HfO_2. For $BaTiO_3$, we align the polarization direction with the DC write step by poling an area with the write voltage prior to performing the spectroscopic experiment. (See SI for additional information.) Relaxation traces for $BaTiO_3$ as a function of read voltages after -6 V_{WRITE} (Fig. 3.3a) and +6 V_{WRITE} (Fig. 3.3b) applications were recorded and remarkably, show features very similar to HfO_2. Both materials have electromechanical signals that appear to qualitatively relax in similar manner. A strong linear dependence on the relaxing amplitude (A_0) is found for both polarities (see Fig. 3.3c), but the time constant (τ_0) dependence appears more complex (Fig. 3.3d). We rule out the instability of switched nanodomains as the origin of relaxation in $BaTiO_3$ since the polarization direction of the sample is aligned with both V_{WRITE} and V_{READ}. Instead, we propose that relaxation occurs due to electrostatics, injected charges, and possible surface diffusion of pre-existing ions. We note that long time constants for $BaTiO_3$ have been identified on oxide surfaces including nanostructured Ceria[112] and $LiNbO_3$[113] using

time-resolved KPFM. In those cases, it was posited that long time constants are likely due to proton transport across the surface. Further work involving chemical imaging in combination with AFM will be required to better isolate the specific physical mechanisms at play.[112]

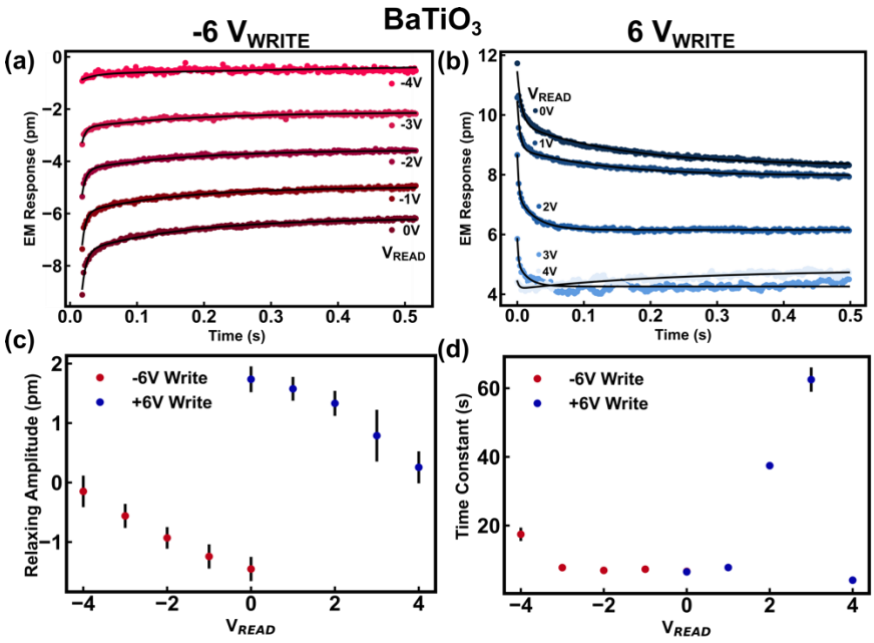

Figure 3.3. Relaxation curves measured for -6 V**(a)** and +6 V**(b)** write voltages. Note that these are measured after pre-poling to align the polarization orientation with the subsequent applied field from the writing pulse, to eliminate domain wall relaxations. Fitted relaxing amplitudes **(c)** and time constants **(d)** for both write polarities. Only the first relaxing amplitude A_0 and time constant τ_0 are shown.

We now investigate how relaxation can impact the form of the measured piezoresponse hysteresis loops and in particular, those derived from contact-KPFM (cKPFM).[88] This technique aims to utilize the linear relationship between the electrostatic response and DC read bias to disentangle it from the ferroelectric contribution, which to the first approximation is invariant to the read voltage. We start by considering the contribution of electrostatics to concavities present in cKPFM loops. Figure 4a shows the average of 25 cKPFM response curves of the $BaTiO_3$ thin film after write steps in the range -9 V and +9 V. (For cKPFM response curves of amorphous HfO_2, see Fig. 3.8.) We first look at the portions of traces acquired with negative write and read voltages (e.g., -9 V_{WRITE} write with negative $V_{READ,}$ shown in red to the left of the dashed line) and note that they

exhibit a linear dependence with respect to the read voltage. This linear dependence with respect to the read voltage is also present in the relaxation curves in Figure 3b before significant relaxation occurs, as corroborated by the linear relationship between the relaxing amplitude and read voltage shown by the red dots in Figure 3.3c. The cKPFM traces in Figure 3.4a are also linear with respect to each other, exhibiting a stacking that is parallels that of the relaxation curves in Figure 3.3b. However, the cKPFM traces acquired with a positive write and read voltage (Fig. 3.4a, shown in blue to the right of the dotted line), have a slight downward concavity. This downward concavity is mimicked in the relaxation curves captured with positive write and read voltages in their initial relaxing amplitude shown by the blue dots in Figure 3.3c. In addition, the cKPFM traces in this portion of the graph exhibit a 'bunching' and are not linearly spaced with respect to each other. This is also reflected in the relaxation curves shown in Figure 3.3b where the relaxation curves are not linearly spaced, and at a read voltage of 4V, shows markedly different behavior.

We now look more specifically at how relaxation impacts the concavity of cKPFM curves by constructing pseudo-cKPFM traces using relaxation spectra acquired by BE-PFM. Figure 3.4b shows the pseudo-cKPFM traces constructed by taking the electromechanical response of the relaxation curves immediately (t = 0 s) after V_{WRITE} is turned off for $BaTiO_3$. The portion constructed with -6 V_{WRITE} and negative read voltages (in red) is linear whereas that constructed with 6 V_{WRITE} and positive read voltages (in blue) exhibits a slight concavity. Figure 3.4c shows the pseudo-cKPFM loops constructed by taking the electromechanical response of the relaxation curves 0.5 s after V_{WRITE} is turned off. The portion constructed with -6 V_{WRITE} and negative read voltages (in red) remains linear whereas the concavity present in that constructed with 6 V_{WRITE} and positive read voltages (in blue) becomes even more dramatic. As a control, we collect cKPFM curves and construct pseudo-cKPFM traces with the relaxation spectra of HfO_2 (see Figure 3.8)

and note that there appears to be a voltage dependence as the curve constructed with 6 V_{WRITE} and positive read voltages (in blue) relaxes differently compared to the other constructed trace. Related to this, we also show that adding a voltage dependence to the relaxing amplitudes in a basic simulation can also introduce concavities in the cKPFM response (see Fig. 3.9). Collectively, these results show that nonlinearities in cKPFM have both a voltage-dependent and an electrostatic, time-dependent component.

We focus on the contribution of electrostatics to the measured PFM signal and nonlinearities in present in cKPFM in this paper and offer ideas for mitigating their contributions. The results presented in this paper suggest that the electrostatic contribution can be removed by acquiring the PFM and cKPFM signals in a relaxed state with the time-independent electrostatic contribution deduced from the total signal. The electrostatic contribution can be obtained by acquiring band excitation piezoresponse force spectroscopy with on- and off-field hysteresis loops and subsequently extracting the electrostatic contribution as described in Ref. 35.[114] Alternatively, an AFM equipped with an interferometric displacement sensor can be used to quantitatively measure the tip displacement.[108]

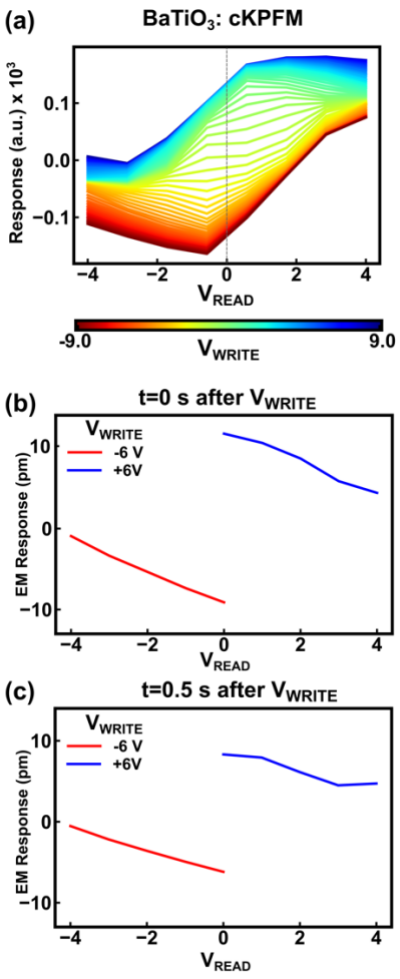

Figure 3.4. (a) Average of 25 cKPFM response curves acquired with V_{WRITE} and V_{READ} in the range -9 V to 9 V on the same $BaTiO_3$ thin film used to obtain relaxation spectra. **(b)** cKPFM traces constructed from the relaxation data immediately (0 s) after V_{WRITE} and **(c)** 0.5 s after V_{WRITE} is turned off.

3.5 Conclusion

In summary, we measured the electrostatic contribution to the relaxation of the electromechanical PFM response on an amorphous HfO_2 and a ferroelectric $BaTiO_3$ thin film. These relaxations have important implications for the interpretation of piezoresponse and contact Kelvin probe force microscopies. Beyond cKPFM and relaxation PFM spectroscopy, relaxation affects related techniques including hysteresis loop acquisition, where a measurement may average over the relaxations and induce strange effects, and standard PFM measurements. We show that concavities

present in cKPFM loops can be explained by relaxation via construction of pseudo-cKPFM loops with BE-PFM relaxation spectra. An understanding of these effects assists in interpretation of loops captured during cKPFM experiments and the electrostatic origin of features such as concavities can be used to remove their contributions.

3.6 Acknowledgements

This work was supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division and performed at the ORNL Center for Nanophase Materials Sciences, which also provided support (LC, SVK), and which is a US DOE Office of the Science User Facility. J.K was supported in part by an appointment to the Oak Ridge National Laboratory HERE program, sponsored by the U.S. Department of Energy and administered by the Oak Ridge Institute for Science and Education. J.K also acknowledges support from the National Science Foundation Research Traineeship under award NSF DGE-1633216 for support in data science training and NSF DMR-1842708 for the data analytics portion.

3.7 Supplementary Information

3.7.1 *Additional Experimental Details*

For BE-PFM, we use an external power source to apply a bias to the sample continuously during both the write and read steps. The sample bias was chosen to acquire a particular read voltage from the perspective of the tip; the tip bias during V_{READ} is 0 V. The voltage applied to the tip during the write steps were changed based on the sample bias to acquire the appropriate write voltage ($\pm$6 V). As an example, the DC biases used to acquire the write and read steps $BaTiO_3$ are shown in Table 3.1. This keeps the electric field in the same direction during both the write and read steps.

The relaxation curves are acquired starting immediately after the write bias is turned off (0 s) and for 0.5 seconds at 2 ms intervals. The rise time associated with the DC bias changes, dictated by the data acquisition card, is around 20 ns. As there is dependence of the measured PFM signal on the AC probing bias,[88,115] we use a 1 V probing AC bias for all measurements to acquire the relaxation curves and cKPFM loop. For HfO_2, measurements were performed on a grid over an 8 μm area. For $BaTiO_3$, measurements were performed on a grid over a 2 μm area within a 2.5 μm area that was pre-poled using the appropriate write voltage. In both cases, we used a 5 x 5 grid, resulting in 25 relaxation spectra per combination of V_{READ} and V_{WRITE}.

Table 3.1. Sample and tip biases used to obtain the desired read (V_{READ}) and write (V_{WRITE}) voltages for $BaTiO_3$.

Tip Bias (V)	-3	-4	-5	-6	6	5	4	3
Sample Bias (V)	3	2	1	0	0	-1	-2	-3
Actual V_{WRITE} (V)	-6	-6	-6	-6	6	6	6	6
Actual V_{READ} (V)	-3	-2	-1	0	0	1	2	3

3.7.2 *Supplemental Figures*

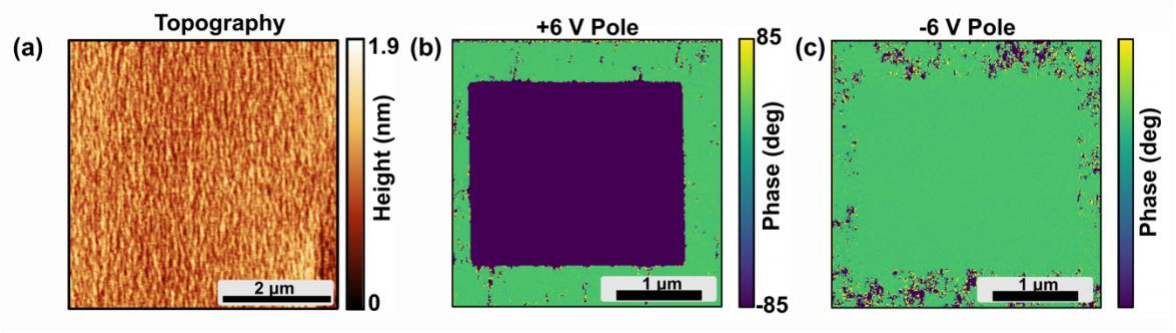

Figure 3.5. (a) Representative topography of the ~80 nm thick epitaxial $BaTiO_3$ thin film with a smooth surface and roughness RMS <0.5 nm. **(b)** Phase image after poling a 2.5 μm area with +6 V and **(c)** -6 V. For $BaTiO_3$, relaxation measurements were performed on a 2 μm area within the poled region.

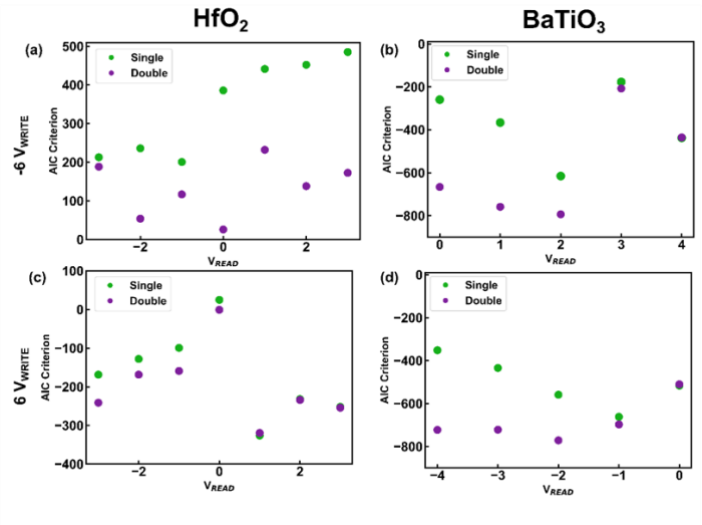

Figure 3.6. The Akaike information criterion (AIC) was used for model selection; a lower AIC indicates the preferred model. Here, the AIC for single and double exponential fits are shown for band excitation piezoresponse force microscopy relaxation curves acquired at -6 **(a, b)** and +6 V_{WRITE} **(c,d)** steps for HfOx **(a, c)** and BTO **(b, d)**. Lower values indicate that model is statistically preferred.

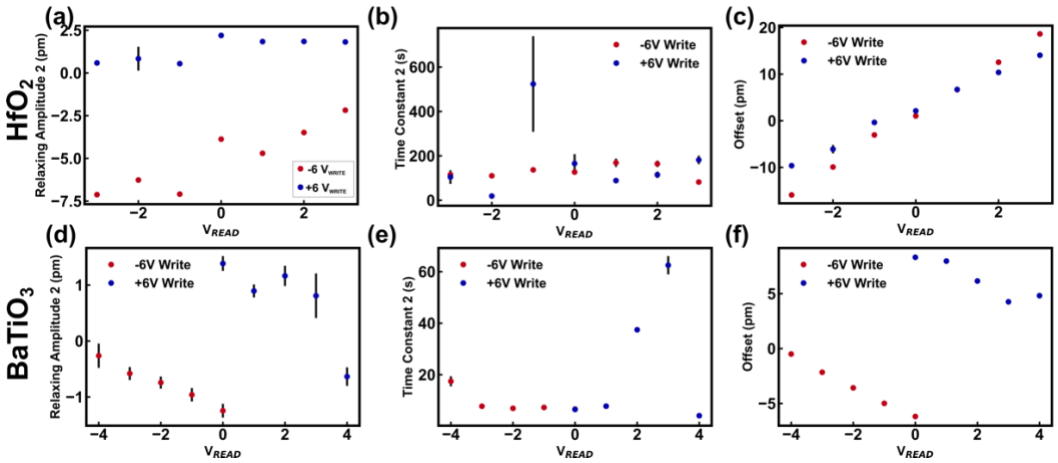

Figure 3.7. Remaining parameters from performing a double exponential fit to relaxation curves for HfO_2 **(a-c)** and $BaTiO_3$ **(d-f)**.

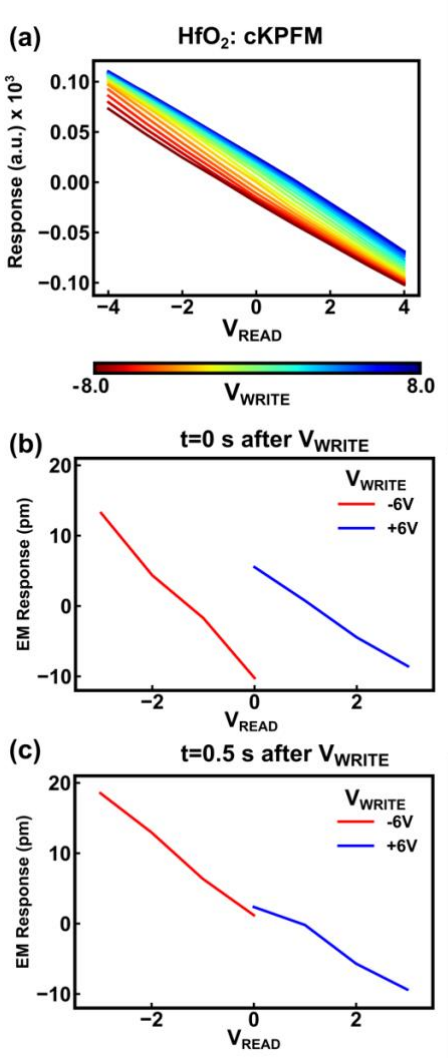

Figure 3.8. (a) Average of 25 Contact Kelvin Probe Force Microscopy (cKPFM) response curves acquired with V_{WRITE} and V_{READ} in the range -8 V to 8 V on amorphous HfO_2. **(b)** cKPFM traces constructed from the relaxation data immediately (0 s) after V_{WRITE} and **(c)** 0.5 s after V_{WRITE} is turned off.

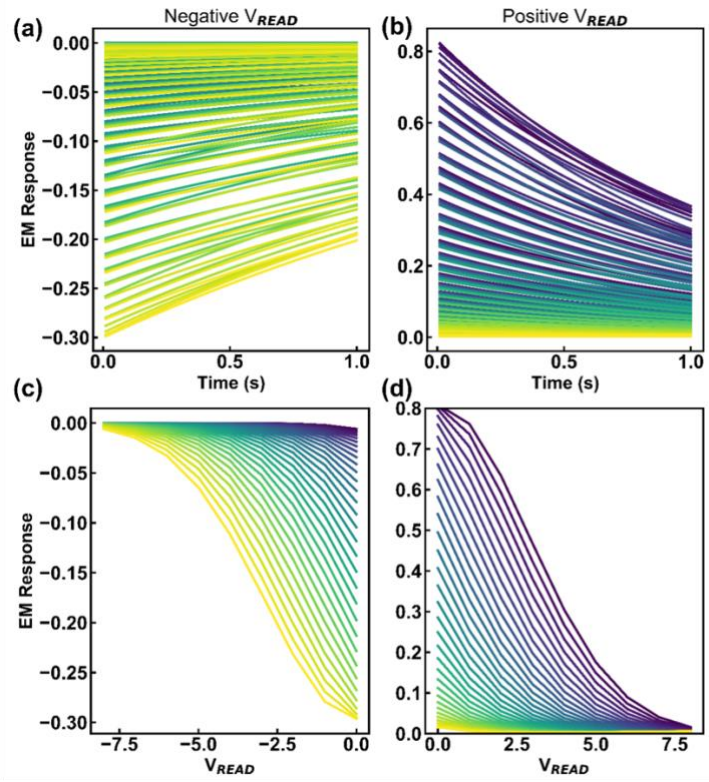

Figure 3.9. Simulated cKPFM traces for **(a,c)** positive and **(b, d)** negative V_{READ}. The traces for **(a,c)** are generated from the dual exponential function $A_1 * \exp(-k_1 * t) + A_2 * \exp(-k_2 * t)$ where A_1 and A_2 are the relaxation amplitudes k_1 and k_2 are the inverse time constants. Relaxation amplitudes are made dependent on the reading voltage multiplied by the write voltage, squared. Time constants are sampled from a normal distribution. The traces for **(c,d)** are obtained by taking the average of the signal in **(a,b)**, respectively. Collectively, these results show that nonlinearities (i.e., not completely linear or step-function responses) in cKPFM are a result of the voltage and time-dependent relaxation amplitude which is electrostatic in origin.

Chapter 4. MACHINE FRAMEWORK FOR EVALUATING TOPIC MODELS IN CONTENT-BASED INFORMATION RETRIEVAL SYSTEMS FOR SCIENTIFIC TEXT[3]

4.1 ABSTRACT

Topic models have the potential to form the basis of a content-based information retrieval system for developing corpora used in data-driven materials development. Here, we evaluate the performance of three topic models in an applied setting and compare their performance in a general language setting. We quantitively evaluate their performance with a ubiquitous metric used to benchmark topic models and introduce a mutual-information based metric that is a proxy for their performance in corpus development. These metrics are validated with human judgments on model coherence. This evaluation pipeline is domain-agnostic and practitioner centric, and can be broadly applied across a variety of settings.

4.2 INTRODUCTION

While the entirety of known chemical and property space is available for data-driven materials research in the form of scientific papers, only a small fraction contains information relevant to a particular field.[116] Extracting scientific data requires the curation of corpora that accounts for the context of pertinent data.[117,118] Constructing field-specific datasets has traditionally required manual curation of relevant papers by domain experts, metadata-based searches, or brute force methods that parse through all articles published in a wide time frame or specific journals.[119–126,11]

[3] Adapted with permission from Kong, J.; Dollar, O.; Pfaendtner, J.; Beck, D.A.C.

Of these approaches, the accessible ones are inherently limited to the knowledge contained within either a group of researchers or the scope of selected journals. With government-regulated pushes to make data and scientific papers open-access, an approach that combines the specificity and generality of demonstrated approaches is needed to parse the resulting papers.

A promising option is to utilize topic models which aim to derive a semantically meaningful latent topic representation of a corpus and its documents along with a predictive model for new texts. Within chemistry, topic modeling has largely been adapted for uses at the field-level. Examples include using them to classify materials synthesis paragraphs,[127] organize 1.6 million molecules of the ChEMBL22 dataset,[128] retrieve microscopy images of materials using queries,[129] and extract relevant features given mass-spectrometry fragmentation spectra.[130] At the domain level, it is primarily used in a manner that parallels aspects of the science of science to varying degrees of specificity.[126,131,132] For example, in one, word embeddings were trained on 3.3 million scientific abstracts and interpreted as latent topics to accurately predict pseudo-future thermoelectric materials.[126] In the context of corpora curation, topic models can be trained on a subset of a large corpus and subsequently be used as an information retrieval system to curate articles based on content.[116,133] This idea utilizes the topic model-generated representation of a document as a query to retrieve related results, combining the benefits of the two existing approaches.

4.3 Experimental Design and Methods

Like most natural language processing (NLP) tools, the development of topic models utilizes general language text, without being "chemistry aware."[5,134–137] Given the development of chemistry-specific models for entity extraction[116,138–145] and domain-specific pretrained models[146–148] along with use of these tools in topic modeling pipelines,[135,149] uncertainty surrounding topic

models performance on scientific text rife with specialized language precludes their usage. To add an additional layer of complication, the topic modeling field develops and benchmarks on metrics that measure the coherence of model-generated topics as opposed to their ability to cluster similar documents together.[149–151] While topic models have the potential to curate high quality, field-specific corpora based on content, a framework to evaluate their performance is needed to enable their usage in content-based information retrieval systems for targeted scientific data extraction.

In this work, we quantify the performance differences of topic models on general and scientific language texts and introduce a domain agnostic, application-centric metric to measure their performance as information retrieval models for corpus curation. The general language corpus is comprised of English articles from Wikipedia and the scientific one built from a subset of abstracts sampled from the S2ORC corpus published in the fields of chemistry and materials science.[118,152] We aim to keep the two corpora as similar as possible such that topic model performance differences can be attributed to the content within the two. Therefore, to parallel the breadth of content described in scientific abstract and obtain entries of similar length, we take the introduction section of Wikipedia entries to construct the general language corpus. Corpora statistics are shown in Table 4.2. (See Supplementary Information for preprocessing details.)

Table 4.2. Corpus information for Wikipedia and S2ORC. For scientific domain corpus used, we sampled a subset of the S2ORC corpus, targeting abstracts published within the fields of chemistry and materials science.

	General	**Chemistry**
Source	Wikipedia	S2ORC
Number of Docs.		
Train	1M	1M
Test	250	250
Mean Tok/Doc	94	95
Vocab Size	25,990	18,256

4.4 Results and Discussion

We focus on three topic models: Latent Dirichlet Allocation (LDA), product of experts LDA (ProdLDA), and contextualized topic model (CTM).[5,134,135,153,154] These models build on top of each other and thereby form a conceptually clean way to assess how incremental changes affect their performance. We start with LDA, which is widely used in applied settings and as a baseline model in topic model development.[127–130,135–137] The basic idea behind LDA is that documents are represented as a mixture of latent topics which are in turn represented by a distribution over a vocabulary of words. Its name is derived from the fact that one of its parameters θ, which affects the topic distribution for each document has a Dirichlet probability distribution.[5] LDA is a multinomial mixture model which means that topics are characterized by a multinomial distribution and each word in a document is sampled from a multinomial distribution preconditioned on a topic.[5] This multinomial mixture is replaced with a product of experts in ProdLDA, which from a statistics perspective can result in distributions sharper than their components and has a practical downstream effect of producing more coherent topics.[134,155] With respect to an information retrieval system, it is possible that a topic model utilizing a product of experts can focus on the variety of aspects present in an abstract that are relevant to researchers from different fields. CTM is an extension of ProdLDA and incorporates pre-trained language models via pre-trained document embeddings from SBERT.[135,156] Across the two corpora, we use correlation explanation to select the hyperparameter N, the number of topics; with the exception of the number of topics, we keep all other hyperparameters the same.[157] (See SI for additional information regarding these models and training hyperparameters.) These models are generative, and their outputs include probability distribution functions of latent topics for each document and words for each topic.

To evaluate these models, we first look at normalized pointwise mutual information (NPMI), which quantifies topic coherence via the pairwise co-occurrence of topic words.[158] For each model trained on the two corpora, we calculate NPMI for the top 10 words of each topic using the training corpus.[159] NPMI is a statistical measure of the pairwise occurrence of the top *N* word and scores topics highly if all pairs have high joint probability compared to their marginal probability. A positive correlation between NPMI and human-perceived topic coherence has been validated in LDA and has made it the prevailing metric used to evaluate newly developed topic models.[149,151] (See SI for additional details about NPMI.)

Table 4.3 shows the three topics with highest NPMI and the top five words associated with that topic for each model and corpus. In general, it is easier to form an intuitive sense of what a topic captures across all model-generated topics for the general language corpus compared to the domain one which is unsurprising due to the highly specialized nature of research articles. It is notable that the mean NPMI across topics in the general language corpus is lower for CTM compared to ProdLDA, given the incorporation of a pretrained model; that it is similar, for ProdLDA and CTM in the chemistry corpus; and that this metric is comparable between the general and domain corpus for CTM. While it is intuitive to qualitatively examine model-generated topics and their differences, we conduct two-sample, pairwise one-tailed Kolmogorov-Smirnov (KS) at α=0.05 to test the observed qualitative differences. (See SI for more information regarding the two-sample KS test, its assumptions, null and alternative hypotheses tested, interpretation of results.) We find that these qualitative observations are corroborated by the KS test. In addition, the results of this test show that despite the apparent larger values in topic NPMI for ProdLDA in the general language corpus compared to the domain one, this difference is not corroborated by the KS test; the corresponding difference is corroborated by the KS test for LDA. Collectively these results

suggest that given consistent selection criteria of the number of topics but otherwise same hyperparameters, ProdLDA and CTM can be used in a domain setting without significantly affecting topic NPMIs whereas LDA cannot.

Table 4.3. The highest-NPMI topics generated from LDA, ProdLDA, and CTM across the two corpora are shown along with the top-10 NPMI of each topic. Mean top-10 NPMI over all topics is shown on the last line for each corpus and model. Models are trained with fifty and sixty topics for the general and chemistry corpus, respectively.

General									
	LDA			ProdLDA			CTM		
	album	party	specie	hot	silver	sun	costa	art	golden
	song	election	plant	chart	olympic	earth	grey	exhibition	nominate
	release	member	find	hop	gold	moon	brown	exhibit	awards
	single	elect	family	billboard	bronze	naked	dot	museum	documentary
	chart	vote	tree	certify	olympics	bright	yellow	paint	nomination
NPMI	0.243	0.219	0.216	0.327	0.322	0.281	0.331	0.266	0.260
Mean		0.142			0.177			0.167	
Chemistry									
	LDA			ProdLDA			CTM		
	electron	light	film	tof	gsh	battery	spectrometry	transmission	review
	spectroscopy	red	layer	flight	ros	cathode	mass	tem	chapter
	ray	green	thin	mass	oxidative	anode	ionization	sem	recent
	microscopy	blue	substrate	maldi	sod	cycling	identification	transform	advance
	scan	color	growth	esi	damage	impedance	tof	shell	summarize
NPMI	0.296	0.256	0.189	0.406	0.280	0.241	0.340	0.289	0.268
Mean		0.110			0.165			0.168	

Next, we examine the correlation of NPMI with human evaluations of topic quality via the word intrusion task.[150,151] In this task, domain experts are shown the top 5 words in random order with an intrusion word that has low probability of being in the current topic and high probability of belonging to another topic. (See SI for more information about this task.) The positive correlation between this metric and human perception has only been validated for LDA; more recently has been found that it may be less reliable than its axiomatic acceptance suggests, especially for neural topic models (here, ProdLDA and CTM).[149,151] Despite the widespread use of this metric in topic model development, we conduct this study to add to the availability of correlated metric with human judgement of which there are few in the topic modeling literature[149–151] and test if a correlation exists between this metric and domain-expert identified coherence. Pending the results of this task, it is possible that:

(i) correlations are less generalizable than previously thought and only applicable when certain criteria overlap, which would be consistent with recent findings[149] and/or,

(ii) NPMI is/is not correlated with the human judgments of topic coherence for the chemistry corpora.

We also examine the correlation of NPMI with a second, application focused task which we call search relevance. In this task, domain-experts are shown the first few sentences of a query abstract from the domain test corpus and asked to rate the relevance of a result abstract on a 3-point scale. This task is modeled after an alternative task used to evaluate topic coherence.[149,151] The result shown is sampled from the top *N* results returned by each model. This task is blind in that the evaluator does not know the ranking of the search result nor which model returns it. This task is to be conducted. (See SI for more information about the search relevance task.) It is possible that this task shows a negative correlation, which would be somewhat unsurprising given the results in Chang et al where it was shown, albeit in a different way that topic coherence was negatively correlated with the overall semantic structure.[150] Regardless of the correlation direction or lack thereof, this information is valuable for assessing the performance of topic models as a foundation for content-based information retrieval systems based on NPMI, which is commonly reported.

To decouple the search relevance task (a measure of the overall semantic structure of the topic model) from NPMI (a measure of the semantic structure of topics), we look at the Jensen-Shannon (JS) distances between the document representations (i.e., their probability distribution functions) from the test corpora. The JS distance is a measure of similarity between two probability distributions bounded by 0 and 1 with 0 meaning that the two are the same. Since these distributions are representations of the abstracts, we interpret the JS distance to be a similarity metric of the documents. First, abstracts are simplistically assigned to the one topic for which it

has the highest probability. Then, for each model, we calculate the mean pairwise JS distances of the representations assigned to the same topic. More concretely, if there are *a* different abstracts belonging to topic *t*, for each abstract, we calculate the $\binom{a}{2}$ different pairwise JS distances and then divide by *a* to obtain the mean pairwise JS distances of representations assigned to the t^{th} topic. For each abstract, we also calculate the mean pairwise JS distances of the representations that do not belong to that topic. To account for the fact that there are more that do not belong to topic *t* than those that do, we randomly sample *r* from the out of topic representations such that *r* is equal to *a*.

Figure 4.1 shows the resulting probability density functions estimates for each corpus-topic model combination with the more opaque distributions representing the in-topic probability density function and the more transparent distributions representing the out-of-topic probability density function estimates. For the general language corpus, the differences between the global maximum of the in and out of topic probability density functions appear to be greatest for ProdLDA and CTM; for the chemistry corpus, the greatest difference is found within CTM. Interestingly, for both corpora, the probability density function for the out of topic JS distances is sharpest for LDA. We suspect the difference between the in and out of topic distributions shown is likely correlated with the results of the to-be-done search relevance task on the domain corpus. Figure 4.2 shows the difference in the probability density function estimates of the in and out of topic JS distances shown in Figure 4.1 for the chemistry corpus. While the distribution for CTM is broader than that for ProdLDA and LDA, it tends to have higher probability density values for higher JS differences. We conduct two-sample, pairwise one-tailed KS at α=0.05 on the difference between the mean out of topic and in topic JS distances. The results of this test show that differences between the mean out of topic and in topic JS distances tend to be higher for CTM compared to ProdLDA and

LDA; and, that they tend to be higher for LDA than ProdLDA. (See SI for more information about null and alternative hypotheses tested and interpretation of results.) The results of this test are indicated by colored circles in Figure 2.

Next, we look at the correlation between the in-topic and out-of-topic JS distances along with their differences with the results of the to-be done search relevance task to see if this task can be automated. Depending on the results of the human evaluation tasks and its correlation with the calculated JS distances, these results can show one or both of the following:

(i) The JS distances of the in/out of topic abstracts or their differences are not/negatively/positively correlated with human judgment of search relevance.

(ii) The distribution of the JS distances of the in/out of topic abstracts or their differences is related to the results of the search relevance task.

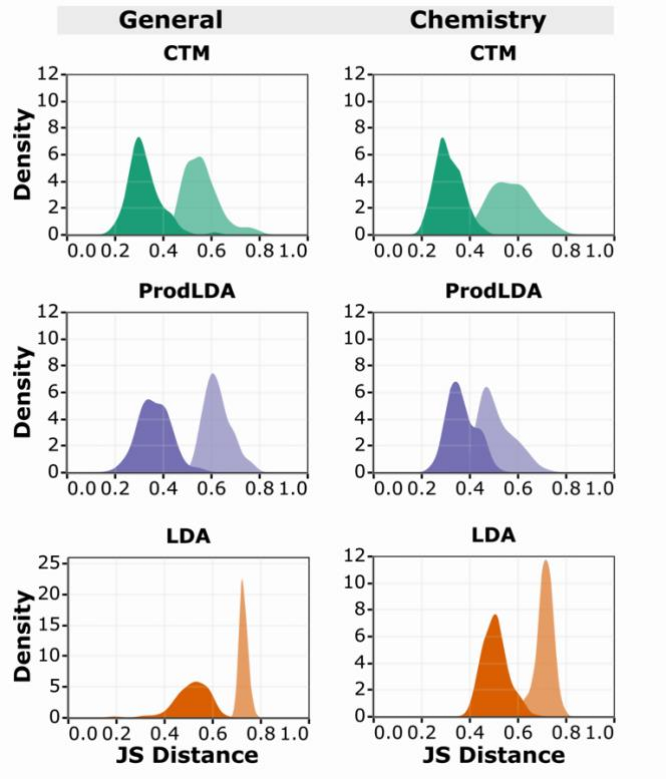

Figure 4.1. Probability density function estimates of the mean pairwise Jensen-Shannon distance between the document representation of an abstract belonging to the same model-generated topic (more opaque shade) and those belonging to different model-generated topics

(more transparent shade) across models for each corpus. Note the change in domain of the density values in LDA for the general language corpus.

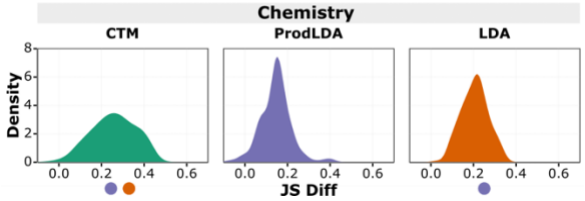

Figure 4.2. Difference in probability density function estimates of the Jensen-Shannon distance between the document representation of an abstract belonging to the same model-generated topic and those belonging to different model-generated topics for the chemistry corpus (JS diff). Results from the two-sample pairwise Kolmogorov-Smirnov tests at α=0.05; for example, the orange dot underneath CTM indicates that the values for JS diff tend to be higher than those for LDA.

4.5 Conclusion

This work evaluates the performance differences of LDA, ProdLDA, and CTM when applied to a scientific domain corpus as opposed to a general language one. The NPMI for topics generated by each model and corpus combination is calculated and it was found that NPMI changes the least when applied to a corpus rife with jargon. This metric is correlated with human judgments on topic coherence, the results and ensuing conclusion which are pending. The JS distance is also introduced as a domain-agnostic, application focused metric to assess their performance on domain corpora curation. We validate these results with a new human evaluation task, the results and subsequent conclusion which are pending.

As it sits at the head of the NLP extraction pipeline, topic modeling can have a variety of applications beyond being the foundation on which a content-based information retrieval system is built.[116,117] It can also be (re)applied to direct annotation efforts to train models for entity and relationship extraction, construction of knowledge graphs, and other downstream tasks to capture

the diversity of texts present in the corpus. This has direct implications on the generalizability of task specific architectures. Using them as opposed to tools based on co-citation and bibliographic coupling can enable the inclusion of both closely and distally related papers in building field-specific databases can also increase the rate at which new innovative concepts are adopted.[117,160]

4.6 Acknowledgements

This work is based on work supported by the U.S. Department of Energy under DE-EE0008492.

4.7 Supplementary Information

4.7.1 *Preprocessing Details*

For preprocessing, we use the default en-core-web-sm spaCy model.[161] We select a random subset of the chemistry and materials science abstracts from the S2ORC corpus and require that abstracts have at least 75 whitespace-separated tokens to ensure that we select abstracts that contain a sufficient amount information. After stop word removal, we remove abstracts with fewer than 50 tokens. We repeat this process until we acquire 1M abstracts which comprise the training set on which the models are trained. For the Wikipedia corpus, we follow the same protocol.

We use a general language stop word list provided by Dieng, et al[136] along with a list of symbols for both corpora. For the scientific domain corpus, we use an additional list of stop words comprising element names and their abbreviations. We tokenize using spaCy keeping the lowercased lemmatized terms if it is more than two letters. Following Hoyle et al, to build dictionary for each corpus, we remove tokens that appear in more than 90% of documents and fewer than $2(0.02|D|)^{1/\log 10}$ documents where $|D|$ is the corpus size which takes into account the power-law distribution of word frequency and scales to large corpora.[149,162]

We manually curate the test data for the scientific domain corpus to obtain a ground-truth categorization of abstracts. The 250 abstracts in the test set fall under five broad categories of papers: nanomaterials, biology, batteries, catalysts, and perovskites and are not included in the training of topic models. These abstracts are broken down into 'batches,' which contain field level topics. We outline descriptions below.

Nanomaterials

- Batch 0: MnO_2 in supercapacitors and their characterization
- Batch 1: usage of transition metal dichalcogenides in hydrogen evolution reactions
- Batch 2: characterization of single-walled carbon nanotubes; either purely computational or with experimental characterization
- Batch 3: development of CdS-based nanomaterials
- Batch 4: WS_2 and its characterization and applications either alone or with other transition metal dichalcogenides

Perovskites

- Batch 0: degradation mechanisms in $CH_3NH_3PbI_3$ perovskites and ways to prevent degradation
- Batch 1: trap-assisted recombination mechanistic studies
- Batch 2: Ruddlesden-Popper perovskites and their optoelectronic applications
- Batch 3: characterization of ferroelectric/multiferroic properties of perovskites
- Batch 4: excitons/biexcitons in cesium lead halide perovskite quantum dots and nanocrystals

Batteries

- Batch 0: understanding of or aiding in uniform formation of solid electrolyte interphase in lithium-ion batteries
- Batch 1: mitigating volume expansion in silicon used in batteries
- Batch 2: synthesis of new materials for cathodes of lithium-ion batteries

- Batch 3: development of new materials or modification of existing materials for solid state electrolytes and their characterization
- Batch 4: development and characterization of electrode materials for use in lithium-ion batteries

Catalysts

- Batch 0: activated carbon composite catalysts in various applications
- Batch 1: synthesis of new TiO_2 photocatalysts and their characterization
- Batch 2: nanoparticles used as catalysts or catalysts supports
- Batch 3: same as batch 2
- Batch 4: same as batch 2

Biology

- Batch 0: hypoxia inducible factor 1 and its role in cancer
- Batch 1: molecular level studies of amyloid beta peptide and its role in Alzheimer's
- Batch 2: estrogen, phytoestrogens, and estrogen receptors in humans and mammals
- Batch 3: glucocorticoids as it relates to stress and skeletal muscles
- Batch 4: antimycobacterial agents in tuberculosis

4.7.2 *Topic Models*

In all models examined in this paper, the common underlying components are as follow. Words are defined as a unit of discrete data from a vocabulary of size *V*. Each document is defined as a sequence of *n* words from the vocabulary and represented by a vector $\vec{\boldsymbol{w}} = (w_1, w_2, \ldots, w_N)$ where *N* is the number of words in a document. A corpus is a collection of documents, $D = \{\overrightarrow{w_1}, \overrightarrow{w_2}, \ldots, \overrightarrow{w_M}\}$ where M is the number of documents. Documents are assumed to be comprised of a mixture of a set of latent topics that is influenced by a parameter α. The words within each document are dependent on two variables: a topic from the mixture of topics comprising the

document and a parameter β. The parameters α and β are document-level and are sampled once per corpus within the model. The key difference between LDA and ProdLDA is that β is unnormalized whereas in LDA, β is constrained to a multinomial simplex.[5,6] For completeness, we note that there is an additional change in ProdLDA: the use of Autoencoding Variational Bayes. However, the use of a variational autoencoder replaces a model-specific inference method and is motivated by the development of a generalized inference method to accelerate topic model development (deriving a model-specific inference method is often difficult) and not differences in encoding documents for model input. As mentioned in the main text, CTM is builds on top of ProdLDA by incorporating document representations.[7]

4.7.3 *Selection of Hyperparameters*

For each model, different hyperparameters are required with the common one being the number of latent topics, N. For this, we utilize CorEx which aims to maximize the total correlation (multivariate mutual information) of the topics generated.[8] For each corpus, we train CorEx topic models on the 1M training set with N ranging from 10 to 170 in steps of 10. We pick the N that contains the maximal amount of total correlation explained after 5 replicates. The average total correlation along with error bars corresponding to 95% confidence intervals are shown in Figure 4.3.

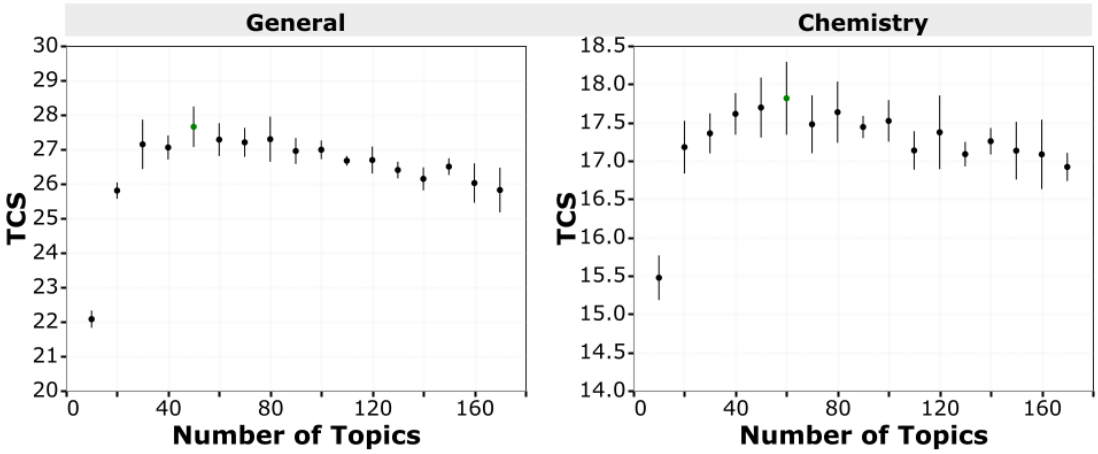

Figure 4.3. Total correlation explained by CorEx model as a function of number of topics for the general and Chemistry language corpora. The model explaining the maximum total correlation is 50 and 60 for the general and Chemistry corpora, respectively and is indicated by a green circle.

4.7.4 *Two-sample Kolmogorov-Smirnov (KS) Test*

The question a two-sample Kolmogorov-Smirnov (KS) test answers is whether two observed distributions come from the same distribution. It is a non-parametric test in that it does not assume that the data in question follows a specific distribution (e.g., a Gaussian distribution). It is a more general than a *t*-test in that it considers the underlying cumulative distribution functions and is thereby sensitive changes in variance, standard deviation, and modality of the two distributions in question.

The null hypothesis of a two-sample KS test is that the observed distribution of values comes from the same underlying CDFs $F(x) = G(x)$. There are three options for an alternative hypothesis; in this paper we use the alternative hypothesis $F(x) < G(x)$. This alternative hypothesis is interpreted by the following example: if x_i has the CDF $F(x)$ and x_j has the CDF $G(x)$ for all observations, then the values in x_i tend to be larger than those in x_j.

4.7.5 *Normalized Pointwise Mutual Information (NPMI)*

To calculate NPMI, we use the input training corpus to estimate the joint word probabilities with a co-occurrence window size of 10 applied to the top 10 words from the head of each topic distribution. Topic NPMIs were calculated with gensim. The resulting probability density function estimates for the topic NPMIs are shown in Figure 4.4. (See titled Probability Density Estimation for more information.)

In the main text, we make some qualitative observations about the topic NPMI values across the different models and corpora. Those observations along with the alternative hypothesis to test them are:

(i) Observation: the mean NPMI across topics in the general language corpus is lower for CTM compared to ProdLDA.
   a. In the alternative hypothesis, $F(x)$ was taken to be the underlying CDF of the calculated topic NPMI values for CTM, and $G(x)$ was taken to be that of ProdLDA.

(ii) Observation: the mean topic NPMI for ProdLDA is similar to that of CTM in the chemistry corpus.
   a. $F(x)$ was taken to be the underlying CDF of the calculated topic NPMI values for CTM, and $G(x)$ was taken to be that of ProdLDA. We test this hypothesis in both directions, i.e., in a second test, we switch what we take to be $F(x)$ and $G(x)$.

(iii) Observation: the mean NPMI is comparable between the general and domain corpus for CTM.
   a. $F(x)$ was taken to be the underlying CDF of the calculated topic NPMI values for the general corpus generated by CTM, and $G(x)$ was taken to be that of the topic NPMI values for the chemistry corpus generated by CTM. We test this hypothesis in both directions.

(iv) We also test if the underlying CDFs of the topic NMPIs within the remaining models (LDA and ProdLDA) across the two corpora are different.
   a. See (iii) a. for an example.

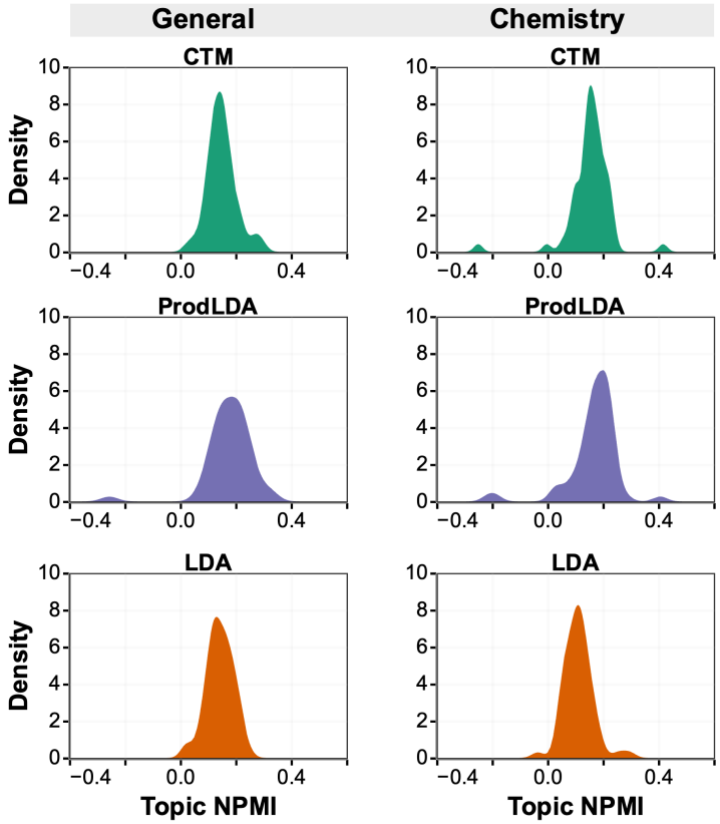

Figure 4.4. Probability density function estimates of the topic NPMIs across models for each corpus.

4.7.6 *Jensen Shannon (JS) Distance Differences*

We test all possible pairwise combinations regardless of the observed global maxima of the probability density functions shown in Figure 2 for the main text. If $F(x)$ is the underlying CDF for CTM, we take $G(x)$ to be the underlying CDF for LDA and ProdLDA. If $F(x)$ is the underlying CDF for ProdLDA, we take $G(x)$ to be the underlying CDF for CTM and LDA despite its global maximum being less than that of LDA and CTM. The results of this are shown in Figure 2 and described in the main text.

4.7.7 *Human Evaluation Tasks*

We conduct a study involving human evaluation of topic quality and search relevance to probe the semantic quality of topics and its overall latent structure, respectively. We implement the following quality control measures for these tasks: we require that domain expert annotators have at least 2+ years of graduate school experience for both the domain and general language task and we show examples and descriptions of each task before asking annotators to perform the task.

For the word intrusion task, we first show domain experts obvious examples of intrusion word.[149] These examples include "baby, crib, diaper, beer, pacifier, cry" as shown in Hoyle et al. with the explanation that the example word "beer" is least related to infants.[149] We show a more difficult example, as shown in Hoyle et al with "hard drive, motherboard, video card, processor, RAM, USB key" with the example word USB key being the least related to the rest which are found inside a computer.

For the search relevance task, we show annotators the first few complete sentences for the query abstract, the first few complete sentences for the result abstract, and ask them to rate the relevance of the result abstract with the following descriptions below.

- Very related: the two abstracts are clearly related to each other, and it would be easy to describe they are related. A researcher doing research on the content in the query abstract will definitely find the search abstract useful.
- Somewhat related: the two abstracts are loosely related to each other, but it is difficult to tell. A researcher doing research on the content in the query abstract might find the search abstract useful.

- Not very related: the two abstracts do not share an obvious relationship to each other. A researcher doing research on the content in the query abstract will not find the search abstract useful.

4.7.8 *Probability Density Function Estimation*

We use Gaussian kernels for probability density function estimation with Scott's Rule to estimate bandwidth. This method is known to work best for unimodal distributions, with a tendency to over smooth multimodal ones. While we see some evidence of the multimodality of the estimates obtained, we are primarily interested in using these estimates to visualize the distribution of the topic NPMIs, in topic and out of topic JS distances and the difference between them. We pick this over box and whisker plots traditionally shown in the topic modeling literature because they provide a sense of how the data are distributed. The probability density function estimates are not used for the KS tests nor are they correlated with results from the search relevance task. For these two tasks, we use the calculated NPMIs and raw difference in the JS distances.

Chapter 5. OUTLOOK AND FUTURE DIRECTIONS

This dissertation outlines projects that are concrete demonstrations of how statistical learning and natural language processing tools made available via a high-level programming language, can be incorporated into the entire research pipeline, albeit in different fields. The work with topic models shows how natural language processing can be utilized at the beginning of the research pipeline to gather a diverse, field-specific corpus for data-driven materials research. An example of how instrument class-specific Python modules and packages can be incorporated into workflows to inform how experiments are conducted to minimize undesirable signal contributions was described in the work with Band-Excitation Piezoforce Microscopy. In another demonstration at the intersection with atomic force microscopy, the work with Photoinduced Force Microscopy is an example of how statistical learning and image registration can be used to fuse the results of two different methods to form an additional data set that provides new domain insights not present in analysis of the inputs individually.

While existing tools can be incorporated as shown in the examples described herein, they are not without limitations. For example, applying dimensionality reduction methods to infrared spectra can yield representation spectra that are not sensible to a domain expert. As a second example, some topic models can be integrated more easily into a domain setting than others, but it is difficult to discern those that can from those that cannot. These limitations point towards the need for integrating domain information into the development of statistical learning methods and natural language processing methods.

BIBLIOGRAPHY

(1) Shannon, C. E. A Mathematical Theory of Communication. 55.

(2) Maes, F.; Collignon, A.; Vandermeulen, D.; Marchal, G.; Suetens, P. Multimodality Image Registration by Maximization of Mutual Information. *IEEE Trans. Med. Imaging* **1997**, *16* (2), 187–198. https://doi.org/10.1109/42.563664.

(3) Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **1962**, *33* (3), 1065–1076. https://doi.org/10.1214/aoms/1177704472.

(4) Kullback, S.; Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22* (1), 79–86. https://doi.org/10.1214/aoms/1177729694.

(5) Blei, D. M. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, 993–1022.

(6) Jaakkola, T. S. A Variational Approach to Bayesian Logistic Regression Models and Their Extensions. 12.

(7) Casella, G.; Robert, C. P. Rao-Biackwellization of Sampling Schemes. 27.

(8) Moerman, D.; Eperon, G. E.; Precht, J. T.; Ginger, D. S. Correlating Photoluminescence Heterogeneity with Local Electronic Properties in Methylammonium Lead Tribromide Perovskite Thin Films. *Chem. Mater.* **2017**, *29* (13), 5484–5492. https://doi.org/10.1021/acs.chemmater.7b00235.

(9) Baum, F.; Pretto, T.; Köche, A.; Santos, M. J. L. Machine Learning Tools to Predict Hot Injection Syntheses Outcomes for II–VI and IV–VI Quantum Dots. *J. Phys. Chem. C* **2020**, *124* (44), 24298–24305. https://doi.org/10.1021/acs.jpcc.0c05993.

(10) Samai, S.; Choi, T. L. Y.; Guye, K. N.; Yan, Y.; Ginger, D. S. Plasmonic Nanoparticle Dimers with Reversibly Photoswitchable Interparticle Distances Linked by DNA. *J. Phys. Chem. C* **2018**, *122* (25), 13363–13370. https://doi.org/10.1021/acs.jpcc.7b10181.

(11) Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chem. Mater.* **2017**, *29* (21), 9436–9444. https://doi.org/10.1021/acs.chemmater.7b03500.

(12) *National Science Foundation Research Traineeship (NRT) Program (nsf21536) | NSF - National Science Foundation*. https://www.nsf.gov/pubs/2021/nsf21536/nsf21536.htm (accessed 2022-06-28).

(13) Balke, N.; Bonnell, D.; Ginger, D. S.; Kemerink, M. Scanning Probes for New Energy Materials: Probing Local Structure and Function. *MRS Bull.* **2012**, *37* (7), 633–637. https://doi.org/10.1557/mrs.2012.141.

(14) Tennyson, E. M.; Gong, C.; Leite, M. S. Imaging Energy Harvesting and Storage Systems at the Nanoscale. *ACS Energy Lett.* **2017**, *2* (12), 2761–2777. https://doi.org/10.1021/acsenergylett.7b00944.

(15) Kalinin, S. V.; Sumpter, B. G.; Archibald, R. K. Big-Deep-Smart Data in Imaging for Guiding Materials Design. *Nat. Mater.* **2015**, *14* (10), 973–980. https://doi.org/10.1038/nmat4395.

(16) Kalinin, S. V.; Strelcov, E.; Belianinov, A.; Somnath, S.; Vasudevan, R. K.; Lingerfelt, E. J.; Archibald, R. K.; Chen, C.; Proksch, R.; Laanait, N.; Jesse, S. Big, Deep, and Smart Data in Scanning Probe Microscopy. *ACS Nano* **2016**, 9068–9086. https://doi.org/10.1021/acsnano.6b04212.

(17) Giridharagopal, R.; Rayermann, G. E.; Shao, G.; Moore, D. T.; Reid, O. G.; Tillack, A. F.; Masiello, David. J.; Ginger, D. S. Submicrosecond Time Resolution Atomic Force

Microscopy for Probing Nanoscale Dynamics. *Nano Lett.* **2012**, *12* (2), 893–898. https://doi.org/10.1021/nl203956q.

(18) Shao, G.; Glaz, M. S.; Ma, F.; Ju, H.; Ginger, D. S. Intensity-Modulated Scanning Kelvin Probe Microscopy for Probing Recombination in Organic Photovoltaics. *ACS Nano* **2014**, *8* (10), 10799–10807. https://doi.org/10.1021/nn5045867.

(19) Salvador, M.; Vorpahl, S. M.; Xin, H.; Williamson, W.; Shao, G.; Karatay, D. U.; Hillhouse, H. W.; Ginger, D. S. Nanoscale Surface Potential Variation Correlates with Local S/Se Ratio in Solution-Processed CZTSSe Solar Cells. *Nano Lett.* **2014**, *14* (12), 6926–6930. https://doi.org/10.1021/nl503068h.

(20) Jesse, S.; Rodriguez, B. J.; Choudhury, S.; Baddorf, A. P.; Vrejoiu, I.; Hesse, D.; Alexe, M.; Eliseev, E. A.; Morozovska, A. N.; Zhang, J.; Chen, L.-Q.; Kalinin, S. V. Direct Imaging of the Spatial and Energy Distribution of Nucleation Centres in Ferroelectric Materials. *Nat. Mater.* **2008**, *7* (3), 209–215. https://doi.org/10.1038/nmat2114.

(21) Coffey, D. C.; Reid, O. G.; Rodovsky, D. B.; Bartholomew, G. P.; Ginger, D. S. Mapping Local Photocurrents in Polymer/Fullerene Solar Cells with Photoconductive Atomic Force Microscopy. *Nano Lett.* **2007**, *7* (3), 738–744. https://doi.org/10.1021/nl062989e.

(22) Stöckle, R. M.; Suh, Y. D.; Deckert, V.; Zenobi, R. Nanoscale Chemical Analysis by Tip-Enhanced Raman Spectroscopy. *Chem. Phys. Lett.* **2000**, *318* (1–3), 131–136. https://doi.org/10.1016/S0009-2614(99)01451-7.

(23) Dunn, R. C. Near-Field Scanning Optical Microscopy. *Chem. Rev.* **1999**, *99* (10), 2891–2928. https://doi.org/10.1021/cr980130e.

(24) Bao, W.; Melli, M.; Caselli, N.; Riboli, F.; Wiersma, D. S.; Staffaroni, M.; Choo, H.; Ogletree, D. F.; Aloni, S.; Bokor, J.; Cabrini, S.; Intonti, F.; Salmeron, M. B.; Yablonovitch, E.; Schuck, P. J.; Weber-Bargioni, A. Mapping Local Charge Recombination Heterogeneity by Multidimensional Nanospectroscopic Imaging. *Science* **2012**, *338* (6112), 1317–1321. https://doi.org/10.1126/science.1227977.

(25) Schmid, T.; Opilik, L.; Blum, C.; Zenobi, R. Nanoscale Chemical Imaging Using Tip-Enhanced Raman Spectroscopy: A Critical Review. *Angew. Chem. Int. Ed.* **2013**, *52* (23), 5940–5954. https://doi.org/10.1002/anie.201203849.

(26) Jahng, J.; Fishman, D. A.; Park, S.; Nowak, D. B.; Morrison, W. A.; Wickramasinghe, H. K.; Potma, E. O. Linear and Nonlinear Optical Spectroscopy at the Nanoscale with Photoinduced Force Microscopy. *Acc. Chem. Res.* **2015**, *48* (10), 2671–2679. https://doi.org/10.1021/acs.accounts.5b00327.

(27) Nowak, D.; Morrison, W.; Wickramasinghe, H. K.; Jahng, J.; Potma, E.; Wan, L.; Ruiz, R.; Albrecht, T. R.; Schmidt, K.; Frommer, J.; Sanders, D. P.; Park, S. Nanoscale Chemical Imaging by Photoinduced Force Microscopy. *Sci. Adv.* **2016**, *2* (3), e1501571. https://doi.org/10.1126/sciadv.1501571.

(28) Dazzi, A.; Prazeres, R.; Glotin, F.; Ortega, J. M. Local Infrared Microspectroscopy with Subwavelength Spatial Resolution with an Atomic Force Microscope Tip Used as a Photothermal Sensor. *Opt. Lett.* **2005**, *30* (18), 2388–2390. https://doi.org/10.1364/OL.30.002388.

(29) Dazzi, A.; Prater, C. B.; Hu, Q.; Chase, D. B.; Rabolt, J. F.; Marcott, C. AFM-IR: Combining Atomic Force Microscopy and Infrared Spectroscopy for Nanoscale Chemical Characterization. *Appl. Spectrosc.* **2012**, *66* (12), 1365–1384.

(30) Centrone, A. Infrared Imaging and Spectroscopy Beyond the Diffraction Limit. *Annu. Rev. Anal. Chem.* **2015**, *8* (1), 101–126. https://doi.org/10.1146/annurev-anchem-071114-040435.
(31) Chaumet, P. C.; Nieto-Vesperinas, M. Time-Averaged Total Force on a Dipolar Sphere in an Electromagnetic Field. *Opt. Lett.* **2000**, *25* (15), 1065–1067. https://doi.org/10.1364/OL.25.001065.
(32) Dholakia, K.; Zemánek, P. Colloquium: Gripped by Light: Optical Binding. *Rev. Mod. Phys.* **2010**, *82* (2), 1767–1791. https://doi.org/10.1103/RevModPhys.82.1767.
(33) Arias-González, J. R.; Nieto-Vesperinas, M. Optical Forces on Small Particles: Attractive and Repulsive Nature and Plasmon-Resonance Conditions. *J. Opt. Soc. Am. A* **2003**, *20* (7), 1201–1209. https://doi.org/10.1364/JOSAA.20.001201.
(34) Jahng, J.; Brocious, J.; Fishman, D. A.; Huang, F.; Li, X.; Tamma, V. A.; Wickramasinghe, H. K.; Potma, E. O. Gradient and Scattering Forces in Photoinduced Force Microscopy. *Phys. Rev. B* **2014**, *90* (15), 155417. https://doi.org/10.1103/PhysRevB.90.155417.
(35) Ladani, F. T.; Potma, E. O. Dyadic Green's Function Formalism for Photoinduced Forces in Tip-Sample Nanojunctions. *Phys. Rev. B* **2017**, *95* (20), 205440. https://doi.org/10.1103/PhysRevB.95.205440.
(36) Jahng, J.; Park, S.; Morrison, W. A.; Kwon, H.; Nowak, D.; Potma, E. O.; Lee, E. S. Nanoscale Spectroscopic Studies of Two Different Physical Origins of the Tip-Enhanced Force: Dipole and Thermal. *arXiv:1711.02479v2 [physics.optics]* **2017**.
(37) Ovchinnikova, O. S.; Tai, T.; Bocharova, V.; Okatan, M. B.; Belianinov, A.; Kertesz, V.; Jesse, S.; Van Berkel, G. J. Co-Registered Topographical, Band Excitation Nanomechanical, and Mass Spectral Imaging Using a Combined Atomic Force Microscopy/Mass Spectrometry Platform. *ACS Nano* **2015**, *9* (4), 4260–4269. https://doi.org/10.1021/acsnano.5b00659.
(38) Jesse, S.; Kalinin, S. V. Principal Component and Spatial Correlation Analysis of Spectroscopic-Imaging Data in Scanning Probe Microscopy. *Nanotechnology* **2009**, *20* (8), 085714. https://doi.org/10.1088/0957-4484/20/8/085714.
(39) Belianinov, A.; Kalinin, S. V.; Jesse, S. Complete Information Acquisition in Dynamic Force Microscopy. *Nat. Commun.* **2015**, *6*, 6550. https://doi.org/10.1038/ncomms7550.
(40) Somnath, S.; Collins, L.; Matheson, M. A.; Sukumar, S. R.; Kalinin, S. V.; Jesse, S. Imaging via Complete Cantilever Dynamic Detection: General Dynamic Mode Imaging and Spectroscopy in Scanning Probe Microscopy. *Nanotechnology* **2016**, *27* (41), 414003. https://doi.org/10.1088/0957-4484/27/41/414003.
(41) Collins, L.; Belianinov, A.; Somnath, S.; Balke, N.; Kalinin, S. V.; Jesse, S. Full Data Acquisition in Kelvin Probe Force Microscopy: Mapping Dynamic Electric Phenomena in Real Space. *Sci. Rep.* **2016**, *6*, 30557. https://doi.org/10.1038/srep30557.
(42) Strelcov, E.; Belianinov, A.; Hsieh, Y.-H.; Jesse, S.; Baddorf, A. P.; Chu, Y.-H.; Kalinin, S. V. Deep Data Analysis of Conductive Phenomena on Complex Oxide Interfaces: Physics from Data Mining. *ACS Nano* **2014**, *8* (6), 6449–6457. https://doi.org/10.1021/nn502029b.
(43) Strelcov, E.; Belianinov, A.; Hsieh, Y.-H.; Chu, Y.-H.; Kalinin, S. V. Constraining Data Mining with Physical Models: Voltage- and Oxygen Pressure-Dependent Transport in Multiferroic Nanostructures. *Nano Lett.* **2015**, *15* (10), 6650–6657. https://doi.org/10.1021/acs.nanolett.5b02472.
(44) Quilettes, D. W. de; Vorpahl, S. M.; Stranks, S. D.; Nagaoka, H.; Eperon, G. E.; Ziffer, M. E.; Snaith, H. J.; Ginger, D. S. Impact of Microstructure on Local Carrier Lifetime in

Perovskite Solar Cells. *Science* **2015**, *348* (6235), 683–686. https://doi.org/10.1126/science.aaa5333.
(45) Kergoat, L.; Battaglini, N.; Miozzo, L.; Piro, B.; Pham, M.-C.; Yassar, A.; Horowitz, G. Use of Poly(3-Hexylthiophene)/Poly(Methyl Methacrylate) (P3HT/PMMA) Blends to Improve the Performance of Water-Gated Organic Field-Effect Transistors. *Org. Electron.* **2011**, *12* (7), 1253–1257. https://doi.org/10.1016/j.orgel.2011.04.006.
(46) Dirlikov, S.; Koenig, J. L. Infrared Spectra of Poly(Methyl Methacrylate) Labeled with Oxygen-18. *Appl. Spectrosc.* **1979**, *33* (6), 551–555.
(47) Hotta, S.; Rughooputh, S. D. D. V.; Heeger, A. J.; Wudl, F. Spectroscopic Studies of Soluble Poly(3-Alkylthienylenes). *Macromolecules* **1987**, *20* (1), 212–215. https://doi.org/10.1021/ma00167a038.
(48) Shaul Mukamel. Quantum Electrodynamics. In *Principles of Nonlinear Optical Spectroscopy*; pp 94–96.
(49) Yang, H. U.; Raschke, M. B. Resonant Optical Gradient Force Interaction for Nano-Imaging and -Spectroscopy. *New J. Phys.* **2016**, *18* (5), 053042. https://doi.org/10.1088/1367-2630/18/5/053042.
(50) Gu, K. L.; Zhou, Y.; Morrison, W. A.; Park, K.; Park, S.; Bao, Z. Nanoscale Domain Imaging of All-Polymer Organic Solar Cells by Photo-Induced Force Microscopy. *ACS Nano* **2018**, *12* (2), 1473–1481. https://doi.org/10.1021/acsnano.7b07865.
(51) Garyfallidis, E.; Brett, M.; Amirbekian, B.; Rokem, A.; Van Der Walt, S.; Descoteaux, M.; Nimmo-Smith, I. Dipy, a Library for the Analysis of Diffusion MRI Data. *Front. Neuroinformatics* **2014**, *8*, 1–17. https://doi.org/10.3389/fninf.2014.00008.
(52) kongjy. *si_notebook.* https://github.com/kongjy/MultimodalAFMonPMMAP3HT_processing.
(53) Pingree, L. S. C.; Reid, O. G.; Ginger, D. S. Imaging the Evolution of Nanoscale Photocurrent Collection and Transport Networks during Annealing of Polythiophene/Fullerene Solar Cells. *Nano Lett.* **2009**, *9* (8), 2946–2952. https://doi.org/10.1021/nl901358v.
(54) Kergoat, L.; Battaglini, N.; Miozzo, L.; Piro, B.; Pham, M.-C.; Yassar, A.; Horowitz, G. Use of Poly(3-Hexylthiophene)/Poly(Methyl Methacrylate) (P3HT/PMMA) Blends to Improve the Performance of Water-Gated Organic Field-Effect Transistors. *Org. Electron.* **2011**, *12* (7), 1253–1257. https://doi.org/10.1016/j.orgel.2011.04.006.
(55) Shlens, J. A Tutorial on Principal Component Analysis. *arXiv:1404.1100v1 [cs.LG]* **2014**.
(56) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(57) Belianinov, A.; He, Q.; Kravchenko, M.; Jesse, S.; Borisevich, A.; Kalinin, S. V. Identification of Phases, Symmetries and Defects through Local Crystallography. *Nat. Commun.* **2015**, *6*, 7801. https://doi.org/10.1038/ncomms8801.
(58) Collins, L.; Ahmadi, M.; Wu, T.; Hu, B.; Kalinin, S. V.; Jesse, S. Breaking the Time Barrier in Kelvin Probe Force Microscopy: Fast Free Force Reconstruction Using the G-Mode Platform. *ACS Nano* **2017**, *11* (9), 8717–8729. https://doi.org/10.1021/acsnano.7b02114.
(59) Nascimento, J. M. P.; Dias, J. M. B. Vertex Component Analysis: A Fast Algorithm to Unmix Hyperspectral Data. *Ieee Trans. Geosci. Remote Sens.* **2005**, 898–910.

(60) Winter, M. E. N-FINDR: An Algorithm for Fast Autonomous Spectral End-Member Determination in Hyperspectral Data; International Society for Optics and Photonics, 1999; Vol. 3753, pp 266–276. https://doi.org/10.1117/12.366289.
(61) Chang, C.-I.; Plaza, A. A Fast Iterative Algorithm for Implementation of Pixel Purity Index. *IEEE Geosci. Remote Sens. Lett.* **2006**, *3* (1), 63–67. https://doi.org/10.1109/LGRS.2005.856701.
(62) Ren, H.; Chang, C.-I. Automatic Spectral Target Recognition in Hyperspectral Imagery. *IEEE Trans. Aerosp. Electron. Syst.* **2003**, *39* (4), 1232–1249. https://doi.org/10.1109/TAES.2003.1261124.
(63) Erichson, N. B.; Mendible, A.; Wihlborn, S.; Kutz, J. N. Randomized Nonnegative Matrix Factorization. *Pattern Recognit. Lett.* **2018**, *104*, 1–7. https://doi.org/10.1016/j.patrec.2018.01.007.
(64) Hyvärinen, A.; Oja, E. Independent Component Analysis: Algorithms and Applications. *Neural Netw. Off. J. Int. Neural Netw. Soc.* **2000**, *13* (4–5), 411–430.
(65) Paatero, P.; Tapper, U. Positive Matrix Factorization: A Non-Negative Factor Model with Optimal Utilization of Error Estimates of Data Values. *Environmetrics* **1994**, *5* (2), 111–126. https://doi.org/10.1002/env.3170050203.
(66) Wood, D.; Hancox, I.; Jones, T. S.; Wilson, N. R. Quantitative Nanoscale Mapping with Temperature Dependence of the Mechanical and Electrical Properties of Poly(3-Hexylthiophene) by Conductive Atomic Force Microscopy. *J. Phys. Chem. C* **2015**, *119* (21), 11459–11467. https://doi.org/10.1021/acs.jpcc.5b02197.
(67) Jiang, X.; Patil, R.; Harima, Y.; Ohshita, J.; Kunai, A. Influences of Self-Assembled Structure on Mobilities of Charge Carriers in π-Conjugated Polymers. *J. Phys. Chem. B* **2005**, *109* (1), 221–229. https://doi.org/10.1021/jp0460994.
(68) Mauer, R.; Kastler, M.; Laquai, F. The Impact of Polymer Regioregularity on Charge Transport and Efficiency of P3HT:PCBM Photovoltaic Devices. *Adv. Funct. Mater.* **2010**, *20* (13), 2085–2092. https://doi.org/10.1002/adfm.201000320.
(69) Murdick, R. A.; Morrison, W.; Nowak, D.; Albrecht, T. R.; Jahng, J.; Park, S. Photoinduced Force Microscopy: A Technique for Hyperspectral Nanochemical Mapping. *Jpn. J. Appl. Phys.* **2017**, *56* (8S1), 08LA04. https://doi.org/10.7567/JJAP.56.08LA04.
(70) Friedman, J. H. Multivariate Adaptive Regression Splines. *Ann. Stat.* **1991**, *19* (1), 1–67.
(71) Christian P. Robert. A Defense of the Bayesian Choice. In *The Bayesian Choice*; Springer Texts in Statistics; Springer, New York, NY, 2007; pp 507–530. https://doi.org/10.1007/0-387-71599-1_11.
(72) *Affine Registration in 3D.* http://nipy.org/dipy/examples_built/affine_registration_3d.html (accessed 2018-02-02).
(73) Richard O. Duda; Peter E. Hart. Parzen Windows. In *Pattern Classification and Scene Analysis*; Wiley: New York, NY, 1973; pp 88–95.
(74) Vajda, I. *Theory of Statistical Inference and Information*; Theory and Decision Library B; Springer Netherlands, 1989.
(75) Lee, D. D.; Seung, H. S. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature* **1999**, *401* (6755), 788–791. https://doi.org/10.1038/44565.
(76) Nascimento, J. M. P.; Dias, J. M. B. Vertex Component Analysis: A Fast Algorithm to Unmix Hyperspectral Data. *Ieee Trans Geosci Rem Sens* **2004**, *43*, 898–910.
(77) A.Lagrange. *VCA: Vertex Component Analysis in Python*; 2018.

(78) Chang, C.-I.; Plaza, A. A Fast Iterative Algorithm for Implementation of Pixel Purity Index. *IEEE Geosci. Remote Sens. Lett.* **2006**, *3* (1), 63–67. https://doi.org/10.1109/LGRS.2005.856701.
(79) James, G.; Witten, D.; Hastie, T.; Tibshirani, R. Linear Regression. In *An Introduction to Statistical Learning*; Springer Texts in Statistics; Springer, New York, NY, 2013; pp 59–126. https://doi.org/10.1007/978-1-4614-7138-7_3.
(80) Kolosov, O.; Gruverman, A.; Hatano, J.; Takahashi, K.; Tokumoto, H. Nanoscale Visualization and Control of Ferroelectric Domains by Atomic Force Microscopy. *Phys. Rev. Lett.* **1995**, *74* (21), 4309–4312. https://doi.org/10.1103/PhysRevLett.74.4309.
(81) Gruverman, A.; Rodriguez, B. J.; Kalinin, S. V. Nanoscale Electromechanical and Mechanical Imaging of Butterfly Wings by Scanning Probe Microscopy. *J. Scanning Probe Microsc.* **2006**. https://doi.org/info:doi/10.1166/jspm.2006.008.
(82) Coondoo, I.; Panwar, N.; Bdikin, I.; Puli, V. S.; Katiyar, R. S.; Kholkin, A. L. Structural, Morphological and Piezoresponse Studies of Pr and Sc Co-Substituted BiFeO3ceramics. *J. Phys. Appl. Phys.* **2012**, *45* (5), 055302. https://doi.org/10.1088/0022-3727/45/5/055302.
(83) Rodriguez, B. J.; Gao, X. S.; Liu, L. F.; Lee, W.; Naumov, I. I.; Bratkovsky, A. M.; Hesse, D.; Alexe, M. Vortex Polarization States in Nanoscale Ferroelectric Arrays. *Nano Lett.* **2009**, *9* (3), 1127–1131. https://doi.org/10.1021/nl8036646.
(84) Kholkin, A. L.; Kalinin, S. V.; Roelofs, A.; Gruverman, A. Review of Ferroelectric Domain Imaging by Piezoresponse Force Microscopy. In *Scanning Probe Microscopy: Electrical and Electromechanical Phenomena at the Nanoscale*; Kalinin, S., Gruverman, A., Eds.; Springer: New York, NY, 2007; pp 173–214. https://doi.org/10.1007/978-0-387-28668-6_7.
(85) Vasudevan, R.; Jesse, S.; Kim, Y.; Kumar, A.; Kalinin, S. Spectroscopic Imaging in Piezoresponse Force Microscopy: New Opportunities for Studying Polarization Dynamics in Ferroelectrics and Multiferroics. *Mrs Commun.* **2012**, *2*, 61–73. https://doi.org/10.1557/mrc.2012.15.
(86) Rodriguez, B. J.; Chu, Y. H.; Ramesh, R.; Kalinin, S. V. Ferroelectric Domain Wall Pinning at a Bicrystal Grain Boundary in Bismuth Ferrite. *Appl. Phys. Lett.* **2008**, *93* (14), 142901. https://doi.org/10.1063/1.2993327.
(87) Vasudevan, R. K.; Balke, N.; Maksymovych, P.; Jesse, S.; Kalinin, S. V. Ferroelectric or Non-Ferroelectric: Why so Many Materials Exhibit "Ferroelectricity" on the Nanoscale. *Appl. Phys. Rev.* **2017**, *4* (2), 021302. https://doi.org/10.1063/1.4979015.
(88) Balke, N.; Maksymovych, P.; Jesse, S.; Herklotz, A.; Tselev, A.; Eom, C.-B.; Kravchenko, I. I.; Yu, P.; Kalinin, S. V. Differentiating Ferroelectric and Nonferroelectric Electromechanical Effects with Scanning Probe Microscopy. *ACS Nano* **2015**, *9* (6), 6484–6492. https://doi.org/10.1021/acsnano.5b02227.
(89) Vorpahl, S. M.; Giridharagopal, R.; Eperon, G. E.; Hermes, I. M.; Weber, S. A. L.; Ginger, D. S. Orientation of Ferroelectric Domains and Disappearance upon Heating Methylammonium Lead Triiodide Perovskite from Tetragonal to Cubic Phase. *ACS Appl. Energy Mater.* **2018**, *1* (4), 1534–1539. https://doi.org/10.1021/acsaem.7b00330.
(90) Röhm, H.; Leonhard, T.; Hoffmann, M. J.; Colsmann, A. Ferroelectric Domains in Methylammonium Lead Iodide Perovskite Thin-Films. *Energy Environ. Sci.* **2017**, *10* (4), 950–955. https://doi.org/10.1039/C7EE00420F.
(91) Liu, Y.; Collins, L.; Proksch, R.; Kim, S.; Watson, B. R.; Doughty, B.; Calhoun, T. R.; Ahmadi, M.; Ievlev, A. V.; Jesse, S.; Retterer, S. T.; Belianinov, A.; Xiao, K.; Huang, J.;

Sumpter, B. G.; Kalinin, S. V.; Hu, B.; Ovchinnikova, O. S. Chemical Nature of Ferroelastic Twin Domains in CH 3 NH 3 PbI 3 Perovskite. *Nat. Mater.* **2018**, *17* (11), 1013–1019. https://doi.org/10.1038/s41563-018-0152-z.

(92) Schulz, A. D.; Röhm, H.; Leonhard, T.; Wagner, S.; Hoffmann, M. J.; Colsmann, A. On the Ferroelectricity of CH 3 NH 3 PbI 3 Perovskites. *Nat. Mater.* **2019**, *18* (10), 1050–1050. https://doi.org/10.1038/s41563-019-0480-7.

(93) Liu, Y.; Collins, L.; Proksch, R.; Kim, S.; Watson, B. R.; Doughty, B.; Calhoun, T. R.; Ahmadi, M.; Ievlev, A. V.; Jesse, S.; Retterer, S. T.; Belianinov, A.; Xiao, K.; Huang, J.; Sumpter, B. G.; Kalinin, S. V.; Hu, B.; Ovchinnikova, O. S. Reply to: On the Ferroelectricity of CH 3 NH 3 PbI 3 Perovskites. *Nat. Mater.* **2019**, *18* (10), 1051–1053. https://doi.org/10.1038/s41563-019-0481-6.

(94) Balke, N.; Jesse, S.; Carmichael, B.; Okatan, M. B.; Kravchenko, I. I.; Kalinin, S. V.; Tselev, A. Quantification of In-Contact Probe-Sample Electrostatic Forces with Dynamic Atomic Force Microscopy. *Nanotechnology* **2017**, *28* (6), 065704. https://doi.org/10.1088/1361-6528/aa5370.

(95) Seol, D.; Kang, S.; Sun, C.; Kim, Y. Significance of Electrostatic Interactions Due to Surface Potential in Piezoresponse Force Microscopy. *Ultramicroscopy* **2019**, *207*, 112839. https://doi.org/10.1016/j.ultramic.2019.112839.

(96) Balke, N.; Maksymovych, P.; Jesse, S.; Kravchenko, I. I.; Li, Q.; Kalinin, S. V. Exploring Local Electrostatic Effects with Scanning Probe Microscopy: Implications for Piezoresponse Force Microscopy and Triboelectricity. *ACS Nano* **2014**, *8* (10), 10229–10236. https://doi.org/10.1021/nn505176a.

(97) Yang, S. M.; Morozovska, A. N.; Kumar, R.; Eliseev, E. A.; Cao, Y.; Mazet, L.; Balke, N.; Jesse, S.; Vasudevan, R. K.; Dubourdieu, C.; Kalinin, S. V. Mixed Electrochemical–Ferroelectric States in Nanoscale Ferroelectrics. *Nat. Phys.* **2017**, *13* (8), 812–818. https://doi.org/10.1038/nphys4103.

(98) Kumar, A.; Ehara, Y.; Wada, A.; Funakubo, H.; Griggio, F.; Trolier-McKinstry, S.; Jesse, S.; Kalinin, S. V. Dynamic Piezoresponse Force Microscopy: Spatially Resolved Probing of Polarization Dynamics in Time and Voltage Domains. *J. Appl. Phys.* **2012**, *112* (5), 052021. https://doi.org/10.1063/1.4746080.

(99) Shvartsman, V. V.; Kholkin, A. L.; Tyunina, M.; Levoska, J. Relaxation of Induced Polar State in Relaxor PbMg1⁄3Nb2⁄3O3 Thin Films Studied by Piezoresponse Force Microscopy. *Appl. Phys. Lett.* **2005**, *86* (22), 222907. https://doi.org/10.1063/1.1942635.

(100) Kumar, A.; Ovchinnikov, O. S.; Funakubo, H.; Jesse, S.; Kalinin, S. V. Real-Space Mapping of Dynamic Phenomena during Hysteresis Loop Measurements: Dynamic Switching Spectroscopy Piezoresponse Force Microscopy. *Appl. Phys. Lett.* **2011**, *98* (20), 202903. https://doi.org/10.1063/1.3590919.

(101) Collins, L.; Jesse, S.; Kilpatrick, J. I.; Tselev, A.; Varenyk, O.; Okatan, M. B.; Weber, S. A. L.; Kumar, A.; Balke, N.; Kalinin, S. V.; Rodriguez, B. J. Probing Charge Screening Dynamics and Electrochemical Processes at the Solid–Liquid Interface with Electrochemical Force Microscopy. *Nat. Commun.* **2014**, *5* (1), 1–8. https://doi.org/10.1038/ncomms4871.

(102) Kim, Y.; Bae, C.; Ryu, K.; Ko, H.; Kim, Y. K.; Hong, S.; Shin, H. Origin of Surface Potential Change during Ferroelectric Switching in Epitaxial PbTiO3 Thin Films Studied by Scanning Force Microscopy. *Appl. Phys. Lett.* **2009**, *94* (3), 032907. https://doi.org/10.1063/1.3046786.

(103) Collins, L.; Liu, Y.; Ovchinnikova, O. S.; Proksch, R. Quantitative Electromechanical Atomic Force Microscopy. *ACS Nano* **2019**, *13* (7), 8055–8066. https://doi.org/10.1021/acsnano.9b02883.

(104) Kiselev, D. A.; Zhukov, R. N.; Bykov, A. S.; Malinkovich, M. D.; Parkhomenko, Y. N. Growth and Investigation of LiNbO3 Thin Films at Nanoscale by Scanning Force Microscopy. In *PIERS PRoceedings*; 2012.

(105) Hong, S.; Tong, S.; Park, W. I.; Hiranaga, Y.; Cho, Y.; Roelofs, A. Charge Gradient Microscopy. *Proc. Natl. Acad. Sci.* **2014**, *111* (18), 6566–6569. https://doi.org/10.1073/pnas.1324178111.

(106) Liu, X.; Kitamura, K.; Terabe, K. Surface Potential Imaging of Nanoscale LiNbO3 Domains Investigated by Electrostatic Force Microscopy. *Appl. Phys. Lett.* **2006**, *89* (13), 132905. https://doi.org/10.1063/1.2358115.

(107) Jesse, S.; Kalinin, S. V. Band Excitation in Scanning Probe Microscopy: Sines of Change. *J. Phys. Appl. Phys.* **2011**, *44* (46), 464006. https://doi.org/10.1088/0022-3727/44/46/464006.

(108) Labuda, A.; Proksch, R. Quantitative Measurements of Electromechanical Response with a Combined Optical Beam and Interferometric Atomic Force Microscope. *Appl. Phys. Lett.* **2015**, *106* (25), 253103. https://doi.org/10.1063/1.4922210.

(109) Li, Q.; Cao, Y.; Yu, P.; Vasudevan, R. K.; Laanait, N.; Tselev, A.; Xue, F.; Chen, L. Q.; Maksymovych, P.; Kalinin, S. V.; Balke, N. Giant Elastic Tunability in Strained BiFeO 3 near an Electrically Induced Phase Transition. *Nat. Commun.* **2015**, *6* (1), 1–9. https://doi.org/10.1038/ncomms9985.

(110) Vasudevan, R. K.; Khassaf, H.; Cao, Y.; Zhang, S.; Tselev, A.; Carmichael, B.; Okatan, M. B.; Jesse, S.; Chen, L.-Q.; Alpay, S. P.; Kalinin, S. V.; Bassiri-Gharb, N. Acoustic Detection: Acoustic Detection of Phase Transitions at the Nanoscale (Adv. Funct. Mater. 4/2016). *Adv. Funct. Mater.* **2016**, *26* (4), 470–470. https://doi.org/10.1002/adfm.201670023.

(111) Sakamoto, Y.; Ishiguro, M.; Kitagawa, G. *Akaike Information Criterion Statistics*; Tokyo : KTK Scientific Publishers ; Dordrecht ; Boston : D. Reidel ; Hingham, MA : Sold and distributed in the U.S.A. and Canada by Kluwer Academic Publishers, 1986.

(112) Ding, J.; Strelcov, E.; Kalinin, S. V.; Bassiri-Gharb, N. Spatially Resolved Probing of Electrochemical Reactions via Energy Discovery Platforms. *Nano Lett.* **2015**, *15* (6), 3669–3676. https://doi.org/10.1021/acs.nanolett.5b01613.

(113) Strelcov, E.; Ievlev, A. V.; Jesse, S.; Kravchenko, I. I.; Shur, V. Y.; Kalinin, S. V. Direct Probing of Charge Injection and Polarization-Controlled Ionic Mobility on Ferroelectric LiNbO(3) Surfaces. *Adv. Mater. Deerfield Beach Fla* **2014**, *26* (6), 958–963. https://doi.org/10.1002/adma.201304002.

(114) Balke, N.; Jesse, S.; Yu, P.; Carmichael, B.; Kalinin, S. V.; Tselev, A. Quantification of Surface Displacements and Electromechanical Phenomena via Dynamic Atomic Force Microscopy. *Nanotechnology* **2016**, *27* (42), 425707. https://doi.org/10.1088/0957-4484/27/42/425707.

(115) Strelcov, E.; Kim, Y.; Yang, J. C.; Chu, Y. H.; Yu, P.; Lu, X.; Jesse, S.; Kalinin, S. V. Role of Measurement Voltage on Hysteresis Loop Shape in Piezoresponse Force Microscopy. *Appl. Phys. Lett.* **2012**, *101* (19), 192902. https://doi.org/10.1063/1.4764939.

(116) Krallinger, M.; Rabal, O.; Lourenço, A.; Oyarzabal, J.; Valencia, A. Information Retrieval and Text Mining Technologies for Chemistry. *Chem. Rev.* **2017**, *117* (12), 7673–7761. https://doi.org/10.1021/acs.chemrev.6b00851.

(117) Olivetti, E. A.; Cole, J. M.; Kim, E.; Kononova, O.; Ceder, G.; Han, T. Y.-J.; Hiszpanski, A. M. Data-Driven Materials Research Enabled by Natural Language Processing and Information Extraction. *Appl. Phys. Rev.* **2020**, *7* (4), 041317. https://doi.org/10.1063/5.0021106.

(118) Lo, K.; Wang, L. L.; Neumann, M.; Kinney, R.; Weld, D. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Online, 2020; pp 4969–4983. https://doi.org/10.18653/v1/2020.acl-main.447.

(119) Foppiano, L.; Dieb, S.; Suzuki, A.; de Castro, P. B.; Iwasaki, S.; Uzuki, A.; Echevarria, M. G. E.; Meng, Y.; Terashima, K.; Romary, L.; Takano, Y.; Ishii, M. SuperMat: Construction of a Linked Annotated Dataset from Superconductors-Related Publications. **2021**. https://doi.org/10.1080/27660400.2021.1918396.

(120) Guo, J.; Ibanez-Lopez, A. S.; Gao, H.; Quach, V.; Coley, C. W.; Jensen, K. F.; Barzilay, R. Automated Chemical Reaction Extraction from Scientific Literature. *J. Chem. Inf. Model.* **2022**, *62* (9), 2035–2045. https://doi.org/10.1021/acs.jcim.1c00284.

(121) Yamaguchi, K.; Asahi, R.; Sasaki, Y. SC-CoMIcs: A Superconductivity Corpus for Materials Informatics. In *Proceedings of the 12th Language Resources and Evaluation Conference*; European Language Resources Association: Marseille, France, 2020; pp 6753–6760.

(122) Dieb, T. M.; Yoshioka, M.; Hara, S. NaDev: An Annotated Corpus to Support Information Extraction from Research Papers on Nanocrystal Devices. *J. Inf. Process.* **2016**, *24* (3), 554–564. https://doi.org/10.2197/ipsjjip.24.554.

(123) Georgescu, A. B.; Ren, P.; Toland, A. R.; Zhang, S.; Miller, K. D.; Apley, D. W.; Olivetti, E. A.; Wagner, N.; Rondinelli, J. M. Database, Features, and Machine Learning Model to Identify Thermally Driven Metal-Insulator Transition Compounds. *Chem. Mater.* **2021**, *33* (14), 5591–5605. https://doi.org/10.1021/acs.chemmater.1c00905.

(124) Hiszpanski, A. M.; Gallagher, B.; Chellappan, K.; Li, P.; Liu, S.; Kim, H.; Han, J.; Kailkhura, B.; Buttler, D. J.; Han, T. Y.-J. Nanomaterial Synthesis Insights from Machine Learning of Scientific Articles by Extracting, Structuring, and Visualizing Knowledge. *J. Chem. Inf. Model.* **2020**, *60* (6), 2876–2887. https://doi.org/10.1021/acs.jcim.0c00199.

(125) Kononova, O.; Huo, H.; He, T.; Rong, Z.; Botari, T.; Sun, W.; Tshitoyan, V.; Ceder, G. Text-Mined Dataset of Inorganic Materials Synthesis Recipes. *Sci. Data* **2019**, *6* (1), 203. https://doi.org/10.1038/s41597-019-0224-1.

(126) Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K. A.; Ceder, G.; Jain, A. Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature. *Nature* **2019**, *571* (7763), 95–98. https://doi.org/10.1038/s41586-019-1335-8.

(127) Huo, H.; Rong, Z.; Kononova, O.; Sun, W.; Botari, T.; He, T.; Tshitoyan, V.; Ceder, G. Semi-Supervised Machine-Learning Classification of Materials Synthesis Procedures. *Npj Comput. Mater.* **2019**, *5* (1), 1–7. https://doi.org/10.1038/s41524-019-0204-1.

(128) Schneider, N.; Fechner, N.; Landrum, G. A.; Stiefl, N. Chemical Topic Modeling: Exploring Molecular Data Sets Using a Common Text-Mining Approach. *J. Chem. Inf. Model.* **2017**, *57* (8), 1816–1831. https://doi.org/10.1021/acs.jcim.7b00249.

(129) Schwenker, E.; Jiang, W.; Spreadbury, T.; Ferrier, N.; Cossairt, O.; Chan, M. K. Y. *EXSCLAIM! -- An Automated Pipeline for the Construction of Labeled Materials Imaging Datasets from Literature*; arXiv:2103.10631; arXiv, 2021. https://doi.org/10.48550/arXiv.2103.10631.

(130) van der Hooft, J. J. J.; Wandy, J.; Barrett, M. P.; Burgess, K. E. V.; Rogers, S. Topic Modeling for Untargeted Substructure Exploration in Metabolomics. *Proc. Natl. Acad. Sci.* **2016**, *113* (48), 13738–13743. https://doi.org/10.1073/pnas.1608041113.

(131) Fortunato, S.; Bergstrom, C. T.; Börner, K.; Evans, J. A.; Helbing, D.; Milojević, S.; Petersen, A. M.; Radicchi, F.; Sinatra, R.; Uzzi, B.; Vespignani, A.; Waltman, L.; Wang, D.; Barabási, A.-L. Science of Science. *Science* **2018**, *359* (6379), eaao0185. https://doi.org/10.1126/science.aao0185.

(132) Griffiths, T. L.; Steyvers, M. Finding Scientific Topics. *Proc. Natl. Acad. Sci.* **2004**, *101* (suppl_1), 5228–5235. https://doi.org/10.1073/pnas.0307752101.

(133) Manning, C. D. *Introduction to Information Retrieval*; Syngress Publishing, 2008.

(134) Srivastava, A.; Sutton, C. Autoencoding Variational Inference For Topic Models. **2017**. https://doi.org/10.48550/arXiv.1703.01488.

(135) Bianchi, F.; Terragni, S.; Hovy, D. Pre-Training Is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. **2020**. https://doi.org/10.48550/arXiv.2004.03974.

(136) Dieng, A. B.; Ruiz, F. J. R.; Blei, D. M. Topic Modeling in Embedding Spaces. arXiv July 7, 2019.

(137) Dieng, A. B.; Ruiz, F. J. R.; Blei, D. M. The Dynamic Embedded Topic Model. arXiv October 10, 2019. https://doi.org/10.48550/arXiv.1907.05545.

(138) Swain, M. C.; Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* **2016**, *56* (10), 1894–1904. https://doi.org/10.1021/acs.jcim.6b00207.

(139) Isazawa, T.; Cole, J. M. Single Model for Organic and Inorganic Chemical Named Entity Recognition in ChemDataExtractor. *J. Chem. Inf. Model.* **2022**, *62* (5), 1207–1213. https://doi.org/10.1021/acs.jcim.1c01199.

(140) Leaman, R.; Wei, C.-H.; Lu, Z. TmChem: A High Performance Approach for Chemical Named Entity Recognition and Normalization. *J. Cheminformatics* **2015**, *7* (1), S3. https://doi.org/10.1186/1758-2946-7-S1-S3.

(141) Wilary, D. M.; Cole, J. M. ReactionDataExtractor: A Tool for Automated Extraction of Information from Chemical Reaction Schemes. *J. Chem. Inf. Model.* **2021**, *61* (10), 4962–4974. https://doi.org/10.1021/acs.jcim.1c01017.

(142) Mavračić, J.; Court, C. J.; Isazawa, T.; Elliott, S. R.; Cole, J. M. ChemDataExtractor 2.0: Autopopulated Ontologies for Materials Science. *J. Chem. Inf. Model.* **2021**, *61* (9), 4280–4289. https://doi.org/10.1021/acs.jcim.1c00446.

(143) Beard, E. J.; Cole, J. M. ChemSchematicResolver: A Toolkit to Decode 2D Chemical Diagrams with Labels and R-Groups into Annotated Chemical Named Entities. *J. Chem. Inf. Model.* **2020**, *60* (4), 2059–2072. https://doi.org/10.1021/acs.jcim.0c00042.

(144) Neumann, M.; King, D.; Beltagy, I.; Ammar, W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*; Association for Computational Linguistics: Florence, Italy, 2019; pp 319–327. https://doi.org/10.18653/v1/W19-5034.

(145) Cohen, K. B.; Lanfranchi, A.; Choi, M. J.; Bada, M.; Baumgartner, W. A.; Panteleyeva, N.; Verspoor, K.; Palmer, M.; Hunter, L. E. Coreference Annotation and Resolution in the Colorado Richly Annotated Full Text (CRAFT) Corpus of Biomedical Journal Articles. *BMC Bioinformatics* **2017**, *18* (1), 372. https://doi.org/10.1186/s12859-017-1775-9.

(146) Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. *ArXiv190310676 Cs* **2019**.

(147) Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthc.* **2022**, *3* (1), 1–23. https://doi.org/10.1145/3458754.

(148) Huang, S.; Cole, J. M. BatteryBERT: A Pretrained Language Model for Battery Database Enhancement. *J. Chem. Inf. Model.* **2022**. https://doi.org/10.1021/acs.jcim.2c00035.

(149) Hoyle, A.; Goel, P.; Peskov, D.; Hian-Cheong, A.; Boyd-Graber, J.; Resnik, P. Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence. 24.

(150) Chang, J.; Gerrish, S.; Wang, C.; Boyd-graber, J.; Blei, D. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2009; Vol. 22.

(151) Lau, J. H.; Newman, D.; Baldwin, T. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*; Association for Computational Linguistics: Gothenburg, Sweden, 2014; pp 530–539. https://doi.org/10.3115/v1/E14-1056.

(152) Wikimedia Foundation. Wikimedia Downloads.

(153) Bingham, E.; Chen, J. P.; Jankowiak, M.; Obermeyer, F.; Pradhan, N.; Karaletsos, T.; Singh, R.; Szerlip, P.; Horsfall, P.; Goodman, N. D. Pyro: Deep Universal Probabilistic Programming. *J. Mach. Learn. Res.* **2019**, *20* (1), 973–978.

(154) Řehůřek, R.; Sojka, P. Software Framework for Topic Modelling with Large Corpora; 2010; pp 45–50. https://doi.org/10.13140/2.1.2393.1847.

(155) Hinton, G. E. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Comput.* **2002**, *14* (8), 1771–1800. https://doi.org/10.1162/089976602760128018.

(156) Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. arXiv August 27, 2019. https://doi.org/10.48550/arXiv.1908.10084.

(157) Gallagher, R. J.; Reing, K.; Kale, D.; Ver Steeg, G. Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 529–542. https://doi.org/10.1162/tacl_a_00078.

(158) Bouma, G. Normalized (Pointwise) Mutual Information in Collocation Extraction. *undefined* **2009**.

(159) Ding, R.; Nallapati, R.; Xiang, B. Coherence-Aware Neural Topic Modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Brussels, Belgium, 2018; pp 830–836. https://doi.org/10.18653/v1/D18-1096.

(160) Hofstra, B.; Kulkarni, V. V.; Munoz-Najar Galvez, S.; He, B.; Jurafsky, D.; McFarland, D. A. The Diversity–Innovation Paradox in Science. *Proc. Natl. Acad. Sci.* **2020**, *117* (17), 9284–9291. https://doi.org/10.1073/pnas.1915378117.

(161) Honnibal, M.; Montani, I.; Van Landeghem, S.; Boyd, A. SpaCy: Industrial-Strength Natural Language Processing in Python.

(162) Zipf, G. K. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*; Addison-Wesley, 1949.