

High Dimensional Data Analysis Of Microscopy Images

Project Summary:

Technological and computing advances have resulted in the ability to obtain new microscopy data with unprecedented resolution over multiple dimensions and over large data sets. This project is aimed at developing techniques to analyze two and three dimensional images that are obtained by scanning tunneling microscopy. The project will develop high-dimensional data analysis tools to identify basic motifs in microscopy images in a statistically rigorous fashion. The motifs identified by image analysis will then be used to feed back into theoretical analyses of the electronic structure of nanoscale materials. The analysis techniques developed as part of this project will also be packaged and made available for use for any microscopy technique.

Basics of STM and STS

Scanning Tunneling Microscopy (STM) is a widely used technique to probe new materials at the nanoscale. The STM sensor relies on the principle of electron tunneling – when two materials are brought extremely closely together (typically 0.5 nanometers apart), electrons can jump from one material to the other as described by the laws of quantum mechanics. If a voltage difference exists

between the two materials, a net flow of electrons from one material to the other occurs, resulting in an electrical current that is detected in the experiment. In practice, one of the two materials is the sample of interest, and the other material is an atomically sharp probe tip that is computer and electronically controlled. By rastering the tip across the material and measuring the flow of electrons at each point, we can obtain a map of the surface of the material down to the sub-atomic scale. Shown in figure 1 are typical STM images obtained in PI Pasupathy's laboratory. Typically, each image contains 256 or 512 pixels per side.

While the basic STM technique described above was invented in the 1980's and 1990's, over the past ~15 years the associated technique of scanning tunneling spectroscopy (STS) has been developed rapidly [1]. In STS, the probe is located at a fixed point in space, and the tunneling current is measured for several different values of voltage. The tip is then rastered across the surface as in STM and the data collected at each point. At the end of the measurement, we have obtained a three-dimensional dataset – one dimension being the voltage values and the other two being the spatial dimensions as before. A typical dataset could have size

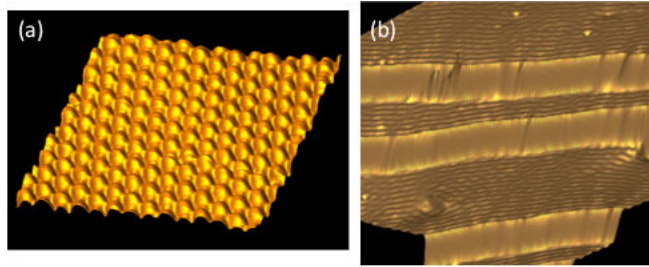


Figure 1: STM images of (a) atoms of gold on a single crystal and (b) surface of copper showing atomic steps and electronic waves

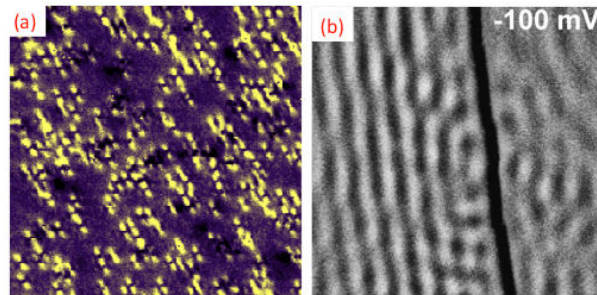


Figure 2: STS images of (a) anisotropic states in NaFeAs (b) structure around a step edge in copper

100 X 512 X 512. From such a dataset, we can generate 100 images in space, one for each of the voltages at which the measurement was conducted. Typical STS images are shown in figure 2.

The density of states and inhomogeneity

The contrast in STS images is directly related to the electronic properties of a nanoscale material. At a basic level, a high value at a given point in space implies that more electrons are tunneling into the material at that given energy. In the language of quantum mechanics, this implies that the density of electronic states available for tunneling to occur is high at that point. The density of states (DOS) determines many basic properties such as electrical resistance, color, reactivity, and so on.

For a perfect crystal, the DOS of the crystal will be uniform in space. For such a material, the STS maps at any voltage will have absolutely no contrast. Real materials on the other hand have disorder on the nanoscale. Such disorder can come from crystal defects such as missing atoms; they can be intentionally placed in the crystal in order to achieve certain desired properties of the nanoscale material; or they can occur in the context of chemistry where the nanomaterial being studied is not a perfect crystal structure. In each of these cases, the STS images will have contrast near the disorder that is present in the crystal. Typical examples are shown in figure 2. The contrast in the STS images is a measure of the disorder in the DOS at the nanoscale. The disorder in the DOS in turn affects fundamental properties of the material – for example, the mobility and resistance of a transistor is strongly influenced by the disorder in the DOS. Further, understanding the disorder in a material can give us basic scientific information on the nature of the electronic states of the material, one of the key goals of materials science.

STS images in nanoscience

In order to get the most useful information out of the STS images, one would like to know the following: given the different types of disorder present, how exactly do each of them influence the DOS in their vicinity? One way to do this would be to make a material that has one type of disorder that is very dilute in space. We could then locate one of the disordered points and perform STS imaging in its vicinity. An example of this is shown in figure 3. In this instance, the crystal is a two-dimensional sheet of graphene, and disorder is created by replacing a few ($\ll 1\%$) of the carbon atoms in the graphene by nitrogen [2]. We can control the number of nitrogen atoms present by synthesis, and so can isolate one nitrogen atom and examine its properties in STS images.

The case of nitrogen-doped graphene described above is very special and not representative of most nanoscale materials. In most materials, the levels of disorder are much higher, and several types of disorder can be present at the same time. One of the chief reasons for this is that in order to produce a na-

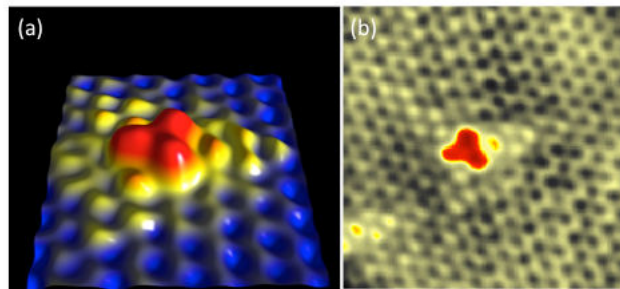


Figure 3: (a) STM and (b) STS images of a single nitrogen dopant in a graphene sheet

nomaterial with certain desired properties, it is often necessary to intentionally add disorder or non-crystallinity at a high level. A classic example is the high-temperature superconductor BSCCO. When “perfect” crystalline BSCCO is produced, the material is not a superconductor at all. To make it a good superconductor, one has to add about 20% excess oxygen into the material [3]. Studying the “perfect” crystal will be useless, since it does not have the desired superconducting property. Modern materials of interest to the nanoscience community are almost always in the state of high disorder. Such high amounts of disorder makes the STS images one obtains on these materials highly complex. Three examples are shown in figure 4 – charge density wave ordered NbSe₂ (unpublished), electronic nematic NaFeAs [1] and quantum dot array of CdSe (unpublished). Unlike nitrogen doped graphene, it is no longer possible to identify individual defects and their effect on the DOS from these images.

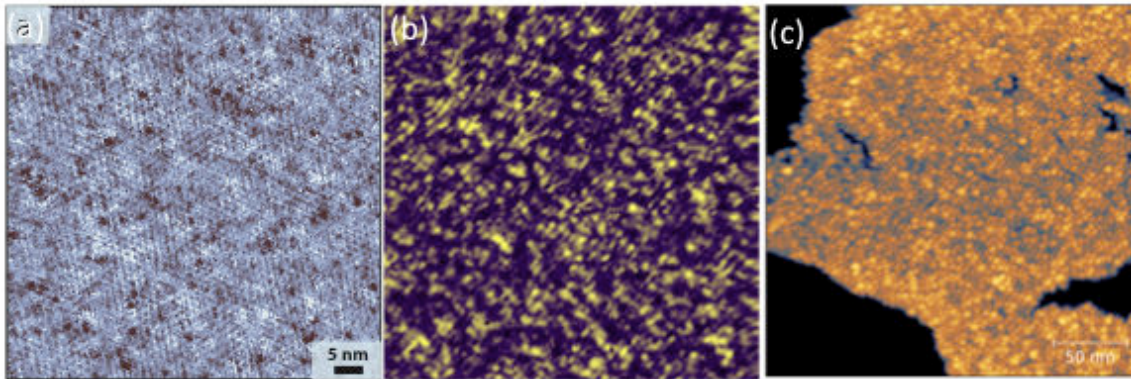


Figure 4: STS images with a high degree of disorder (a) NbSe₂ (b) Na-CoFeAs (c) CdSe quantum dots

Current state of the art in STS image analysis

The question arises as to how to extract the maximum meaningful information present in STS images such as the ones shown in figure 4. Since one cannot identify individual defects, one has to use statistical tools to extract meaningful information. Currently, the state-of-the-art method to analyze STS images is based on the Fourier transform (called FT-STS) [4]. A single defect produces a DOS pattern around it in space, and the FT of this pattern gives us information on the wavelengths present in the DOS pattern. If one imagines that a complex STS image such as the ones shown in figure 5 are made up of identical copies of individual patterns distributed in space, a FT of the entire image will be able to identify these wavelengths. This is shown for two materials in figure 5. Further theoretical analysis of the patterns then involves constructing models to explain the wavelengths observed in the FT images [5].

The FT-STS procedure suffers from several significant drawbacks. The first is the so-called phase problem. When one takes the FT of patterns that are located randomly in space, each of the patterns when transformed picks up a phase factor $\exp(ik \cdot r)$. When one adds together the many patterns that arise from each of the disorder points in the material, we obtain at each wavelength a sum of a large number of phases. This sum of phases is in general a number that fluctuates strongly and can take any magnitude from $-N$ to N at each point, where N is the total number of impurities. This results in a large noise in the FT-STS images, as can be seen in figure 5. A related problem is that in the limit where N

is large, the magnitude of the phase factor averages out to zero! This gives rise to the very undesirable situation that the more data we obtain in real space (for large data sets), the useful information we obtain from FT-STS is not improved.

A second and equally important drawback of the FT-STS procedure is that it completely fails when multiple types of disorder are present in the material of interest. In such a situation, the FT-STS images mix together the patterns from each type of disorder, rendering the entire analysis useless.

Finally, the FT-STS procedure also fails when the basic motif in real space cannot be reduced to a few wavevectors in Fourier space. This is especially relevant in other microscopy techniques – for example, an optical image of a dense array of fluorescent proteins cannot be reduced simply to a few wavelengths in the image. In this case, one needs to perform the entire analysis in real space.

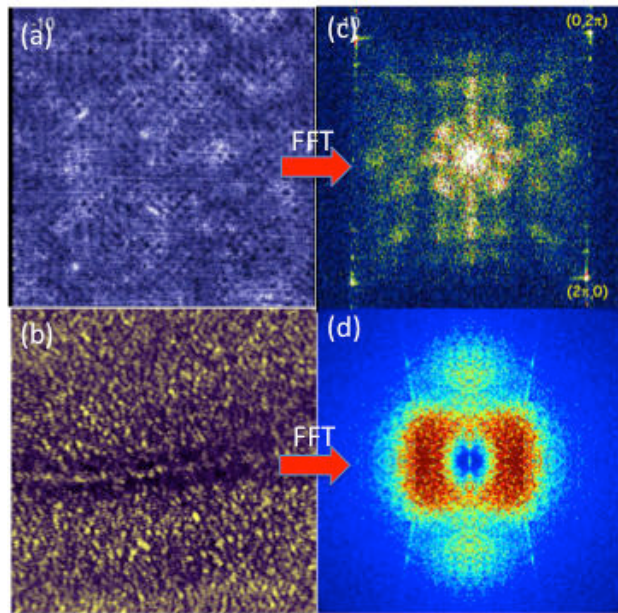


Figure 5: Real space STS images of (a) BSCCO and (b) NaFeAs and their Fourier transform FT-STS images in (c) and (d) respectively.

An ideal analysis

An ideal analysis of the images will be able to extract the DOS pattern associated with each type of disorder at each energy by directly working with real space images. We will rely on the fact that the number of types of disorder is limited to a few per image. The key questions we would like to answer are:

- (a) What is the transformation that best represents the real space data (by best we imply that it represents the data with the minimum number of parameters possible)
- (b) Can we start with theoretical models of patterns, and optimize to figure out the best defect patterns that fit the spatial data?
- (c) Can we use the information gained at one energy to learn how to fit the data at other energies better?

High-dimensional Data analysis

We will leverage recent theoretical and algorithmic developments in data science — specifically, in the area of *high-dimensional data analysis*. This area seeks to develop efficient computational tools that can uncover simple structure in large, noisy datasets, as well as fundamental theory to clarify when this is possible.

In particular, the problem of identifying the density of states pattern in the region of a defect maps almost perfectly onto a data analysis problem called **dictionary learning [6]**. In a nutshell, the dictionary learning problem is as follows: *Given one or more input signals, identify the minimum number of basic building blocks (words in the dictionary)*

needed to concisely approximate the input. That is, instead of designing a data representation analytically (as with Fourier or wavelet approaches), one attempts to learn a representation *from the data itself* (see figure 6).

In applications in consumer image and video processing, these learned transformations substantially outperform Fourier and wavelet approaches

(see figure 7) [7-9]. The reason is that they can adapt to structure present in the specific input signal of interest, whereas Fourier approaches must cope with very broad, and sometimes unrealistic classes of signals (e.g., all low-pass signals). Tools developed by Wright’s group have been used to determine several such dictionaries in various contexts.

In the context of STS, the dictionary to be learned will consist of the DOS patterns associated with different types of defects in the material — one per type of disorder. Compared to FT-STs, a data-driven approach should have the following advantages: (i) ability to directly learn across voltage scales in an integrated fashion, and hence produce more accurate estimates, even when only one type of disorder is present, (ii) ability to cope with multiple types of disorder (a situation in which traditional STS algorithms completely break down), and (iii) ability to seamlessly integrate scientific knowledge on the functional form of the disorder.

We will develop our approach in two stages. In the first stage, we will attempt to learn the DOS patterns in a strictly data driven fashion. For this, we will leverage recent advances in dictionary learning and sparse optimization developed in PI Wright’s group. In the second stage, we will attempt to balance between allowing the data to speak for itself and enforcing priors on the structure of the DOS patterns.

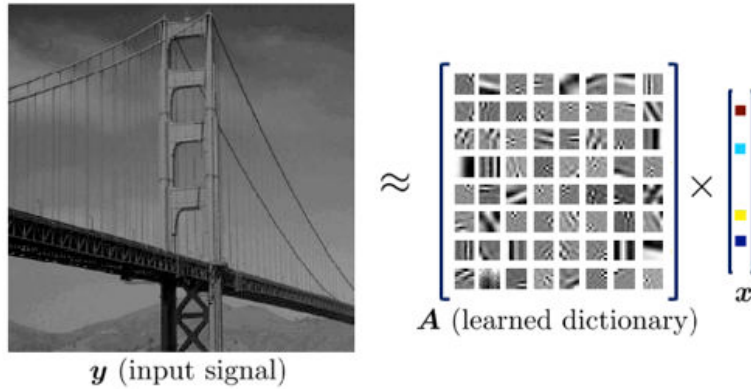


Figure 6: Learning to represent an image: an image is decomposed as a superposition of a few basis signals (the dictionary), chosen to make the representation as concise as possible.

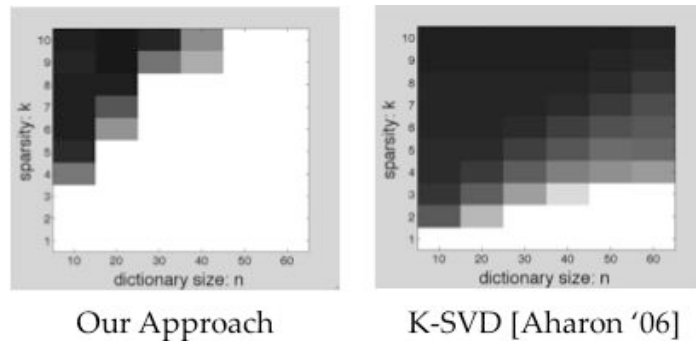


Figure 7: Phase transitions in dictionary recovery, using algorithm based on PI Wright’s research. White indicates perfect recovery. Our approach solves all sufficiently well-structured problems exactly.

Data Science challenges and technical approach

While learned signal representations have had tremendous impact in academic and industrial signal processing, thus far their impact on the natural sciences has been extremely limited. This is partially due to traditional communication and funding barriers to work at the interface of the data and natural sciences, which we will elaborate on below. However, it is also due to the difficulty of the data science problem itself. For consumer applications, one may be satisfied with visually appealing images. For applications in the natural sciences, however, it is essential that the computational tool be well-structured, and with clearly delineated working conditions, since they need to be part of a chain that produces reliable scientific knowledge.

This is especially challenging for dictionary learning. The most natural formulation is highly nonconvex, and the problem is NP-hard in the worst case [10]. Until very recently, no theory was available to explain which instances could be solved. Recently, PI Wright, in collaboration with Wang and Spielman, demonstrated the first efficient algorithm for dictionary learning with a theoretical performance guarantee [10]. This algorithm is based on the observation that if the number of dictionary elements is not too large, it is possible to reformulate the problem as a sequence of linear programs, which can be solved very efficiently using mature numerical techniques [11].

Wright's group has also done highly impactful work on related data sciences problems such as robust matrix and tensor recovery [12, 13]. Because the observed data cube can be viewed as a large, noisy three-way tensor, these techniques may also be useful for improving the estimates produced by our algorithms. The PI's group has a strong track record of producing effective, useful computational tools with clear theoretical foundations. In addition to the specific technical advances outlined above, another impact of this project will be the introduction of modern data-driven signal processing tools into the microscopy community.

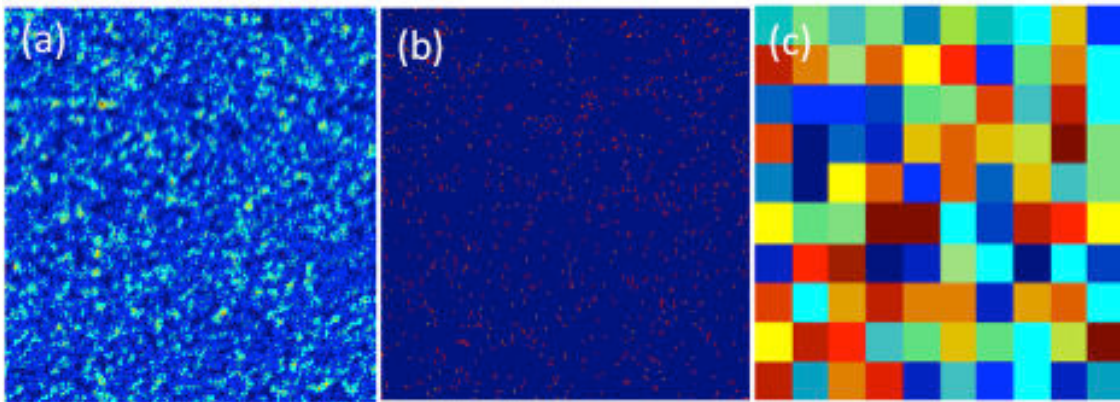


Figure 8: Analysis of STS image using high-dimensional data analysis. (a) Original STS image (b) Location of points of defects by optimization (c) optimized DOS structure from analysis

Progress in 16 months

The two PI's have already identified two outstanding graduate students who are midway through their doctoral programs. Both are already completely familiar with the techniques in natural science and data science respectively. Large amounts of microscopy

data already exist in PI Pasupathy's group, and basic analysis tools have already been developed in PI Wright's group. Motivated by this grant opportunity, the two groups have already begun an analysis of STS images from Cheung/Pasupathy using existing techniques developed by Sun/Wright. An example of this analysis is shown in figure 8. We are able to optimize for the location of scattering centers in an individual STS image, and extract the DOS pattern associated with the defect (assumed to be only one type in the current iteration).

A new themed cluster - extension to other microscopy techniques

The problem described above is of great generality in modern microscopy tools across the natural sciences. In the field of materials science and chemistry, there exist several tools that investigate materials in space as well as one or more additional dimensions, usually energy or time. Typical examples are transmission electron microscopes, atomic force microscopes, and so on. All of these tools produce data that is similar in concept to the STS imaging described above. Another key technique that is directly relevant is optical microscopy, especially as applied to biophysics, biochemistry and astronomy. Modern optical tools such as fluorescent labeling, near field spectroscopy and so on also generate complex, multidimensional dimensional data sets. In the case of chemical reagents and biological materials, the generators of optical signals can be close together, resulting in images that are functionally equivalent to STS images described earlier. This is also true of astronomical images in many cases. The data science techniques developed here can be directly used to address these problems as well.

We expect that several faculty in the natural sciences at Columbia and beyond will be directly able to use the data science techniques developed as part of this project. One of the most important outputs of the project will be a numerical toolbox, which is made freely available to interested researchers at Columbia and in the larger scientific community. We believe that the problem is of high significance across the natural sciences and can support a new themed cluster in the IDSE.

Why traditional funding will not fund this project currently

To be successful in this project, we require a hands-on application of cutting edge data science to modern microscopy imaging data, with active feedback between the two. This interdisciplinary requirement makes it difficult to fund using the traditional funding sources, either in data science or in natural science. The traditional funding sources in natural sciences (NSF, DOE) are willing to fund cutting edge microscopy projects. However, funding is limited to traditional natural scientists for experiments, and not for analysis. Funding also exists for theoreticians in the natural sciences, but by and large they are not familiar with advanced techniques in data science. On the other hand, funding in the computational and statistical sciences is typically directed either towards methodological work, or traditional consumer-oriented imaging applications.

References

1. Rosenthal, E.P., et al., *Visualization of electron nematicity and unidirectional antiferroic fluctuations at high temperatures in NaFeAs*. Nature Physics, 2014. **10**(3): p. 225-232.
2. Zhao, L.Y., et al., *Visualizing Individual Nitrogen Dopants in Monolayer Graphene*. Science, 2011. **333**(6045): p. 999-1003.
3. McElroy, K., et al., *Atomic-scale sources and mechanism of nanoscale electronic disorder in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$* . Science, 2005. **309**(5737): p. 1048-1052.
4. McElroy, K., et al., *Relating atomic-scale electronic phenomena to wave-like quasiparticle states in superconducting $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$* . Nature, 2003. **422**(6932): p. 592-596.
5. Mesaros, A., et al., *Topological Defects Coupling Smectic Modulations to Intra-Unit-Cell Nematicity in Cuprates*. Science, 2011. **333**(6041): p. 426-430.
6. Aharon, M., M. Elad, and A. Bruckstein, *K-SVD: Design of dictionaries for sparse representation*. SPARSE, 2005.
7. Elad, M. and M. Aharon, *Image Denoising via Sparse and Redundant Representations over Learned Dictionaries*. IEEE Transactions on Image Processing, 2006. **15**(12): p. 3736-3745.
8. Mairal, J., M. Elad, and G. Sapiro, *Sparse Representation for Color Image Restoration*. IEEE Transactions on Image Processing, 2008. **17**(1).
9. Yang, J., et al., *Image Super-Resolution via Sparse Representation*. IEEE Transactions on Image Processing, 2010. **19**(11): p. 2861-2873.
10. Spielman, D.A., H. Wang, and J. Wright. *Exact recovery of sparsely-used dictionaries*. in *Proceedings of the 25th Annual Conference on Learning Theory*. 2012.
11. Nesterov, Y., *Introductory lectures on convex optimization : a basic course*. 2004: Kluwer.
12. Candes, E., et al., *Robust Principal Component Analysis?* Journal of the ACM, 2011. **58**(3).
13. Mu, C., et al. *Square deal: New lower bounds and improved relaxations for tensor recovery*. in *International Conference on Machine Learning*. 2014.