MATDAT18: Materials and Data Science Hackathon

Team Composition (2 people max.)

Name	Department	Institution	Email
Henry D. Castillo (student)	Chemistry	Indiana University	henrcast@indiana.edu
Steven L. Tait (Associate Prof.)	Chemistry	Indiana University	tait@indiana.edu

Project Title

Machine Learning for Identification and Automated Analysis of 2D Supramolecular Self-Assemblies at Surfaces from Scanning Tunneling Microscopy Data

Project Synopsis (approx. 100 words)

Molecular self-assembly at interfaces produce highly-ordered materials and are studied to gain fundamental understanding of intermolecular interactions and materials design. We analyze these systems with scanning tunneling microscopy to obtain real space imaging of molecular packing at surfaces. The supramolecular structure is derived from STM data by visual inspection, then sketching out packing models. Complex systems can be polymorphic and studies often require statistical analysis from many images, so human analysis is inefficient and often inaccurate. Machine learning offers an opportunity to generate packing models, reconcile those with molecular structure, and provide statistical analysis of their consistency with the experimental data.

Identified Data-Science Collaborative Need (approx. 100 words)

We would like to develop algorithms to: (1) recognize and measure periodic patterns in STM images, (2) develop feasible molecular packing models consistent with those patterns, (3) reconcile the packing structure with the molecular structure, (4) identify structural domain boundaries, and (5) compile statistical data from large sets of images on relative populations of packing polymorphs. Machine learning algorithms will be employed to improve packing model prediction. Bayesian data analysis will assess the quality of the models. These analyses will require identification of real molecular structures from data with noise and variation in data quality, which can be quantified in an uncertainty analysis.

Data Origin and Access (*data must be available and sharable with data science teams* – please address: data source/origin, access privileges, sharing privileges)

Data input will be (1) STM image files, stored as 2D matrices of pixel (height) values, (2) molecular models, and (3) charge distribution models of molecules.
The proposers have full access to the data and these can be shared easily through cloud storage (box/dropbox) or USB thumb drive transfer. A data set for a single experiment is on the order of 50-100 MB. We will be able to have 10-20 such data sets available at the hackathon and more available by cloud.

Project Description (approx. 1.5 pages, plus figures and references; please describe data size, form, dimensionality, uncertainties, number of examples, etc.)

Self-assembly involves the spontaneous ordering of molecular components via non-covalent (supramolecular) interactions to produce complex materials.[1] Nature is filled with such structures (e.g., DNA, cell walls, crystals) that come together through various supramolecular interactions, including van der Waals (vdW) interactions, hydrogen bonding, and other electrostatic interactions. Compared with biological systems, synthetic (man-made) systems are much simpler, but rapid progress is being made in advancing complexity and predictive modeling of such structures.

Our experiments with two-dimensional (2D) supramolecular self-assembly using molecular resolution microscopy allow key insight into molecular packing, but need computer-aided data reduction and packing model analysis. Self-assembly at solid surfaces is of great interest because the surface not only templates self-assembly to form materials that are not possible in solution but also enables scanning tunneling microscopy (STM) analysis which can visualize individual molecular components (Figure 1).[2-4] These studies also lead toward next generation materials as the combination of surface support and STM analysis provides a means to develop nanomaterials from bottom-up strategies with applications ranging from electronics, photovoltaics, sensors, and separations.[2, 5-10]

Strategic and predictive molecular design for functional supramolecular layers is prerequisite to such bottom-up strategies and relies on a thorough understanding of the complex interplay of non-covalent interactions between molecular components. Through the use of STM, much has been learned by us and many other groups about the interplay of supramolecular interactions and other parameters (e.g., concentration and surface binding) on the structure and properties of self-assembled materials.[2, 8-9]

Task 1. Machine learning algorithm to identify periodic structures in STM images

From STM imaging, spatial information and electronic features can be obtained. In order to translate this data into a molecular model, STM images are visually inspected to look for repeating patterns (see examples in Figure 1). We envision an automated machine learning algorithm that could search large sets (usually 50 images in an experiment taken over the course of a day) for repeating patterns that might indicate molecular ordering at a surface. For single image analysis, doing this analysis “by hand” is sufficient (Figure a-b), but in more complex systems, particularly those with polymorphic properties (Figure c-d), human analysis is a limiting factor.

Task 2. Develop feasible packing models based on images and molecular structure

The next phase of the algorithm would use the molecular structure as one input and the spatial coordinates of the packing structures (unit cell vectors) from STM images as a second input. These would be used together to generate feasible models for the molecular packing. Quantum mechanical calculations are being developed by several research groups, but they are computationally expensive and time consuming.[11] We find that students with molecular models can make good guesses about packing structure by hand based on their intuition about electrostatic and other interactions that would drive assembly. We envision an algorithm that could also generate structures in this empirical manner. A machine learning aspect of this would be to have the computer build up a data set of known structures with characteristics of their molecular properties to inform future structure estimations. This would be an extremely efficient

computer algorithm for making initial predictions about structure. The best of those could be used as input into more expensive computational studies and would also serve to improve the determination of structures in future experiments.

Task 3. Statistical analysis of molecular packing structures in STM data

Another major advance in STM analysis of supramolecular packing at surfaces will be the development of algorithms to measure domain sizes and domain boundary locations. After identification of repeating patterns (Task 1), the algorithm would compare multiple images to look for recurrence of the same patterns. Real STM images have tip artifacts that alter the contrast of the molecules, as well as noise and thermal drift. The algorithm would use a statistical analysis to build confidence estimations for the measurements. Comparing many images would allow the algorithm to identify recurring patterns, in spite of the unpredictable variations from image to image due to experimental artifacts and noise. Compiling statistics on the packing structures from multiple images will improve confidence in the measurements, but also provide valuable population data analysis, especially for systems with multiple polymorphs on the surface where the relative population of those is important to the thermodynamic analysis of the system.

STM images are two-dimensional matrices of pixel (height) data, typically 512 x 512 pixels. Each image is 1-3 MB in size. We typically record 20-50 images on a sample at various scan sizes from 20 nm x 20 nm to 500 nm x 500 nm. At these high resolution images, it is often possible to image molecular structures with sub-Angstrom resolution, allowing for submolecular resolution and identification of molecular orientation on the surface. As noted above, there can be significant variability due to changes in tip quality in these high resolution scans.

Based on a previous work on a computerized method to study 3-D crystal structures from X-ray scattering,[12] we estimate that algorithms involved in feature selection, shape matching, object recognition, spatial statistics, and machine learning of QSPR models are required to analyze STM images, identify key features, and generate models. Additional algorithms involving Bayesian analysis, uncertainty analysis, and data reduction techniques are required to generate probabilities of different packing models and analyze large sets of STM data of varying quality.

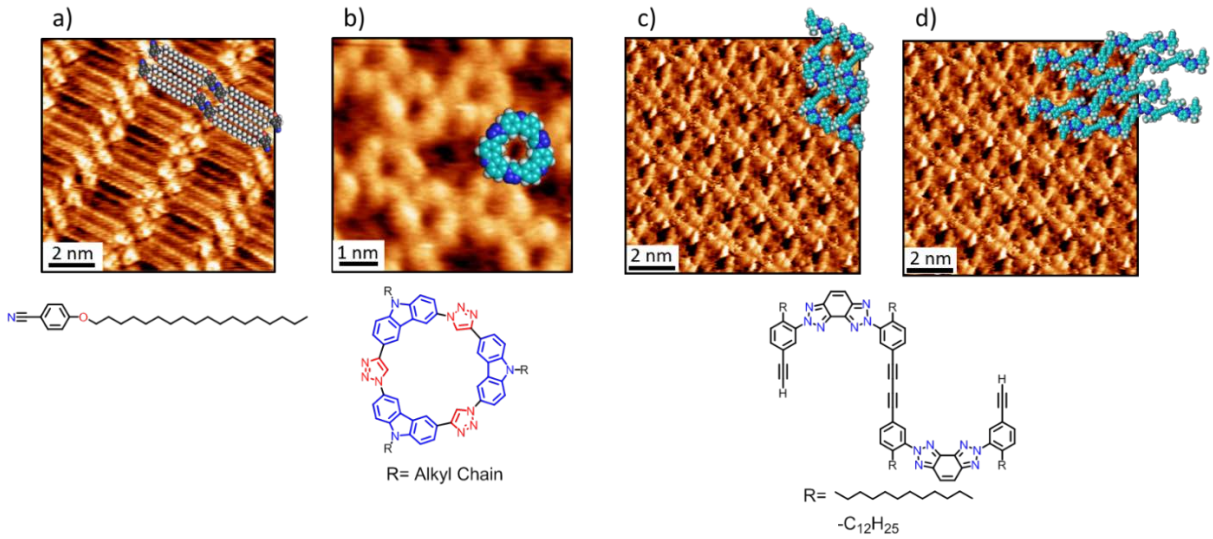

Figure 1. Examples of molecular resolution STM images with overlays of molecular packing models of three molecular systems at the solution-graphite interface: (a) alkoxybenzonitriles, (b) tricarbazolo triazolophane macrocycles,[9] (c-d) two possible models of heteroaryleneethynylenes.

References

1. Steed, J. W.; Atwood, J. L., *Supramolecular Chemistry*. 2 ed.; Wiley: New York, 2009.
2. Mali, K. S.; Pearce, N.; De Feyter, S.; Champness, N. R., Frontiers of supramolecular chemistry at solid surfaces. *Chem. Soc. Rev.* **2017,** *46*, 2520-2542.
3. Binnig, G.; Rohrer, H.; Gerber, C.; Weibel, E., Surface Studies by Scanning Tunneling Microscopy. *Phys. Rev. Lett.* **1982,** *49*, 57-61.
4. Somorjai, G. A.; Li, Y., *Introduction to Surface Chemistry and Catalysis*. John Wiley & Sons: 2010.
5. Ray, B.; Alam, M. A., Random vs regularized OPV: Limits of performance gain of organic bulk heterojunction solar cells by morphology engineering. *Sol. Energy Mater. Sol. Cells* **2012,** *99*, 204-212.
6. Simpson, C. D.; Wu, J.; Watson, M. D.; Mullen, K., From graphite molecules to columnar superstructures – an exercise in nanoscience. *J. Mater. Chem.* **2004,** *14*, 494-504.
7. Angelova, P.; Vieker, H.; Weber, N.-E.; Matei, D.; Reimer, O.; Meier, I.; Kurasch, S.; Biskupek, J.; Lorbach, D.; Wunderlich, K.; Chen, L.; Terfort, A.; Klapper, M.; Müllen, K.; Kaiser, U.; Gölzhäuser, A.; Turchanin, A., A Universal Scheme to Convert Aromatic Molecular Monolayers into Functional Carbon Nanomembranes. *ACS Nano* **2013,** *7*, 6489-6497.
8. Hirsch, B. E.; McDonald, K. P.; Qiao, B.; Flood, A. H.; Tait, S. L., Selective anion-induced crystal switching and binding in surface monolayers modulated by electric fields from scanning probes. *ACS Nano* **2014,** *8*, 10858-10869.
9. Lee, S.; Hirsch, B. E.; Liu, Y.; Dobscha, J. R.; Burke, D. W.; Tait, S. L.; Flood, A. H., Multifunctional Tricarbazolo Triazolophane Macrocycles: One-Pot Preparation, Anion Binding, and Hierarchical Self-Organization of Multilayers. *Chem. Eur. J.* **2016,** *22*, 560-569.

10. Sosa-Vargas, L.; Kim, E.; Attias, A.-J., Beyond “decorative” 2D supramolecular self-assembly: strategies towards functional surfaces for nanotechnology. *Mater. Horiz.* **2017,** *4*, 570-583.
11. Teobaldi, G.; Hofer, W. A.; Bikondoa, O.; Pang, C. L.; Cabailh, G.; Thornton, G., Modelling STM images of TiO2(110) from first-principles: Defects, water adsorption and dissociation products. *Chem. Phys. Lett.* **2007,** *437*, 73-78.
12. Phillips, C. L.; Voth, G. A., Discovering crystals using shape matching and machine learning. *Soft Matter* **2013,** *9*, 8552-8568.