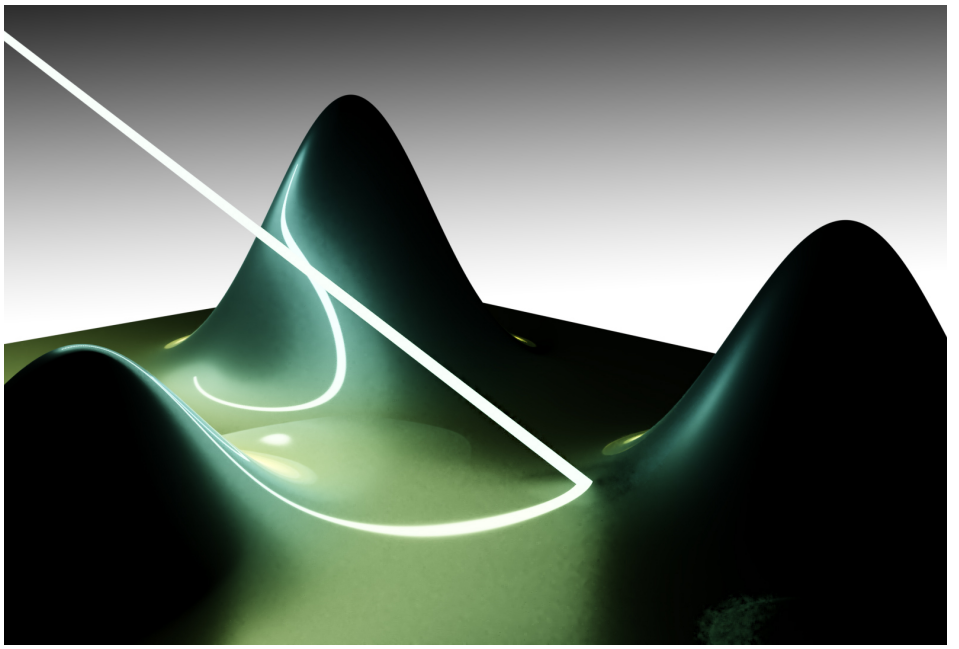
Autonomous Data Acquisition for Scientific Discovery



An artistic illustration of Gaussian functions and a light or particle beam passing through. The image alludes to the inner workings of the algorithm inside gpCAM, a software tool developed by researchers at Berkeley Lab to facilitate autonomous scientific discovery. (Credit: Marcus Noack/Berkeley Lab)

Scientific Achievement

Researchers at large scientific facilities such as the Advanced Light Source (ALS) have applied a quick, robust machine-learning technique to automatically optimize data gathering for a variety of experimental techniques.

Significance and Impact

The work promises to enable experiments with large, complex datasets to be run more quickly, efficiently, and with minimal human intervention.

Keeping up with the data

Experimental facilities around the globe are facing a challenge: their instruments are becoming increasingly powerful, leading to a steady increase in the volume and complexity of the scientific data they collect. To make full use of modern instruments and facilities, researchers need new ways to decrease the amount of data required for scientific discovery and address data acquisition rates humans can no longer keep pace with.

A promising route lies in an emerging field known as autonomous discovery, where algorithms learn from a comparatively small amount of input data and decide themselves on the next steps to take, allowing multidimensional parameter spaces to be explored more quickly, efficiently, and with minimal human intervention.

A Gaussian process reduces uncertainty

Over the last few years, Gaussian process regression (GPR) has emerged as the method of choice for steering many classes of experiments. GPR makes decisions about where to sample next based on minimizing uncertainties. First, a Gaussian-shaped probability density function is applied to a "prior" model—an initial best guess. As real data points are acquired, the probability density function yields a "posterior" uncertainty (mean and variance) at each point. This information is then used to make decisions about future measurements that minimize the uncertainty. This process, which can be based on relatively small datasets, lies at the heart of gpCAM, a software tool developed by the Center for Advanced Mathematics for Energy Research Applications (CAMERA) at Berkeley Lab.

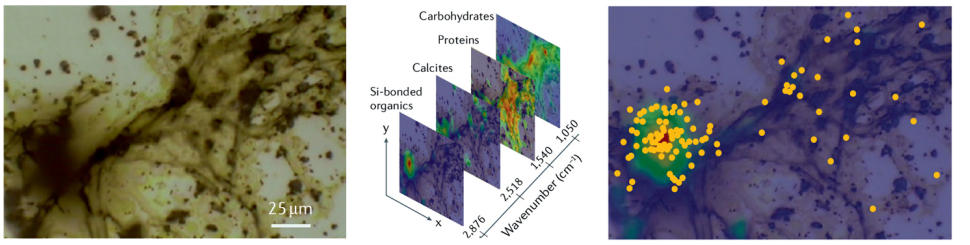
Two case studies at the ALS

At ALS Beamline 1.4, researchers customized gpCAM to steer data collection in an infrared spectromicroscopy study of a well-characterized sample of microbialite, a rock-like sedimentary deposit formed by

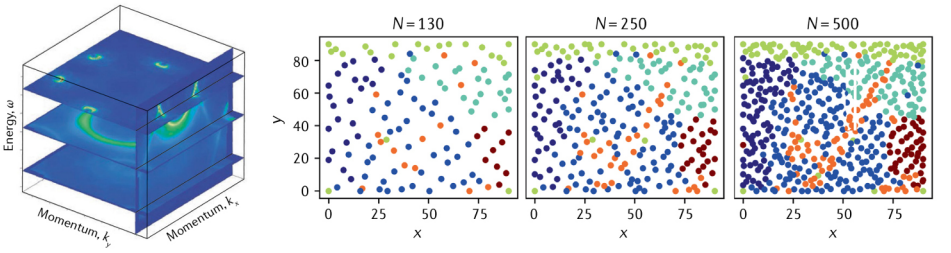
marine microbes. Such studies produce "high-dimensional" infrared images in which each pixel contains rich physical and chemical information. Using gpCAM, regions of interest were identified and sufficiently well resolved after 200 measurements, compared to the approximately 10,000 measurements needed for a full-grid scan. Experiments that once took over eight hours can now be completed within 30 minutes, potentially broadening the scope of biological spectromicroscopy to transient biogeochemical processes.

At Beamline 7.0.2, researchers applied gpCAM to angle-resolved photoemission spectroscopy (ARPES) data from a "twisted" stack of two graphene layers. The lower layer consisted of homogeneous graphene, while the upper layer consisted of distinct regions with different rotations. For certain "magic" rotation angles, the material becomes superconducting. To autonomously find domains near the magic angle, gpCAM was trained to identify areas of the sample with similar spectral characteristics (a method known as *K*-means clustering). For this and other datasets studied, gpCAM with *K*-means clustering greatly outperformed grid-based searches. Often, less than 10% of the dataset was adequate for an accurate description of the number of phases present, together with their boundaries.

While CAMERA's initial case studies have focused primarily on synchrotron beamline experiments, a growing number of scientists in other areas—including math, drug discovery, computer science, and electrical engineering—are now seeing the advantages of incorporating autonomous discovery techniques into their workflows. By automating time- and energy-consuming data collection, these developments liberate researchers to focus on what they're best at: scientific meaning and discovery.



Left: Micrograph of the microbialite used in this study. Center: Multidimensional chemical data showing Si-bonded organics, calcites, proteins, and carbohydrates extracted from a dataset with 1,738 spectral dimensions. Right: Feature-finding sampling locations (orange dots) rendered by gpCAM, using a known reference spectrum associated with Si-bonded organics.



Left: Typical ARPES data for graphene, illustrating the measurement's four-dimensional (x, y, energy, momentum) parameter space. Right: Evolution of the clustered data points. At 100 points, patterns have begun to appear; at 250 data points, the regions are visible; by 500 data points, borders between phases have successfully been defined. It should be noted that this diagram contains just one-sixteenth of the entire dataset.

Contact: Marcus Noack (MarcusNoack@lbl.gov)

Publications: M.M. Noack, P.H. Zwart, D.M. Ushizima, M.Fukuto, K.G. Yager, K.C. Elbert, C.B. Murray, A. Stein, G.S. Doerk, E.H.R. Tsai, R. Li, G. Freychet, M. Zhernenkov, H.-Y.N. Holman, S. Lee, L. Chen, E. Rotenberg, T. Weber, Y. Le Goc, M. Boehm, P. Steffens, P. Mutti, and J.A. Sethian, "Gaussian processes for autonomous data acquisition at large-scale synchrotron and neutron facilities," *Nat. Rev. Phys.* **3**, 685 (2021), doi:10.1038/s42254-021-00345-y.

Researchers: M.M. Noack, P.H. Zwart, H.-Y.N. Holman, and L. Chen (Berkeley Lab); D.M. Ushizima (Berkeley Lab and UC San Francisco); M.Fukuto, K.G. Yager, A. Stein, G.S. Doerk, E.H.R. Tsai, R. Li, G. Freychet, and M. Zhernenkov (Brookhaven National Laboratory); K.C. Elbert and C.B. Murray (Univ. of Pennsylvania); S. Lee and J.A. Sethian (Berkeley Lab and UC Berkeley); E. Rotenberg (ALS); and T. Weber, Y. Le Goc, M. Boehm, P. Steffens, and P. Mutti (Institut Laue-Langevin, France).

Funding: Advanced Scientific Computing Research (ASCR), Biological and Environmental Science (BER), and Basic Energy Sciences (BES) programs of the U.S. Department of Energy (DOE), Office of Science; Laboratory Directed Research and Development program, Berkeley Lab; Office of Naval Research; and National Science Foundation. Operation of the ALS is supported by DOE BES.

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

Published by the
ADVANCED LIGHT SOURCE
COMMUNICATIONS GROUP