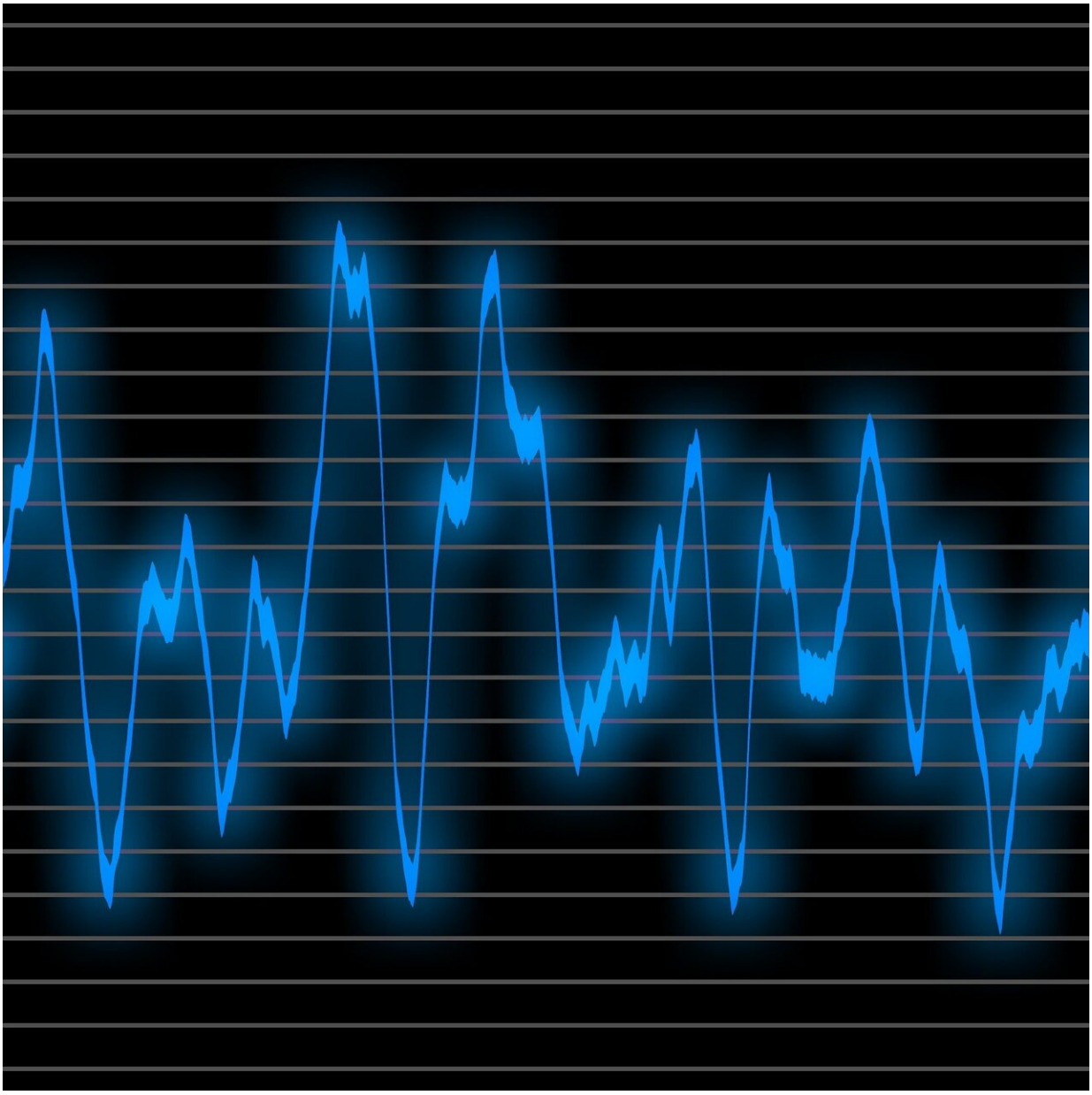
Reimagining the shape of noise leads to improved molecular models

October 21 2020, by Greta Lorge



Credit: CC0 Public Domain

Tenacity comes naturally to a guy who hails from the "mule capital of the world." That trait has stood Columbia, Tennessee, native Elliot Perryman in good stead as an intern at Lawrence Berkeley National Laboratory (Berkeley Lab). Last fall, he began working with staff scientist Peter Zwart in the Center for Advanced Mathematics for Energy Research Applications (CAMERA) through the Berkeley Lab Undergraduate Research program.

CAMERA aims to identify areas in experimental science that can be aided by new applied mathematical insights. These interdisciplinary researchers develop the necessary algorithmic tools and deliver them as user-friendly software. Zwart put Perryman, a computer science and physics major at the University of Tennessee, on a project he likened to "going around in a dark room trying to find a cat."

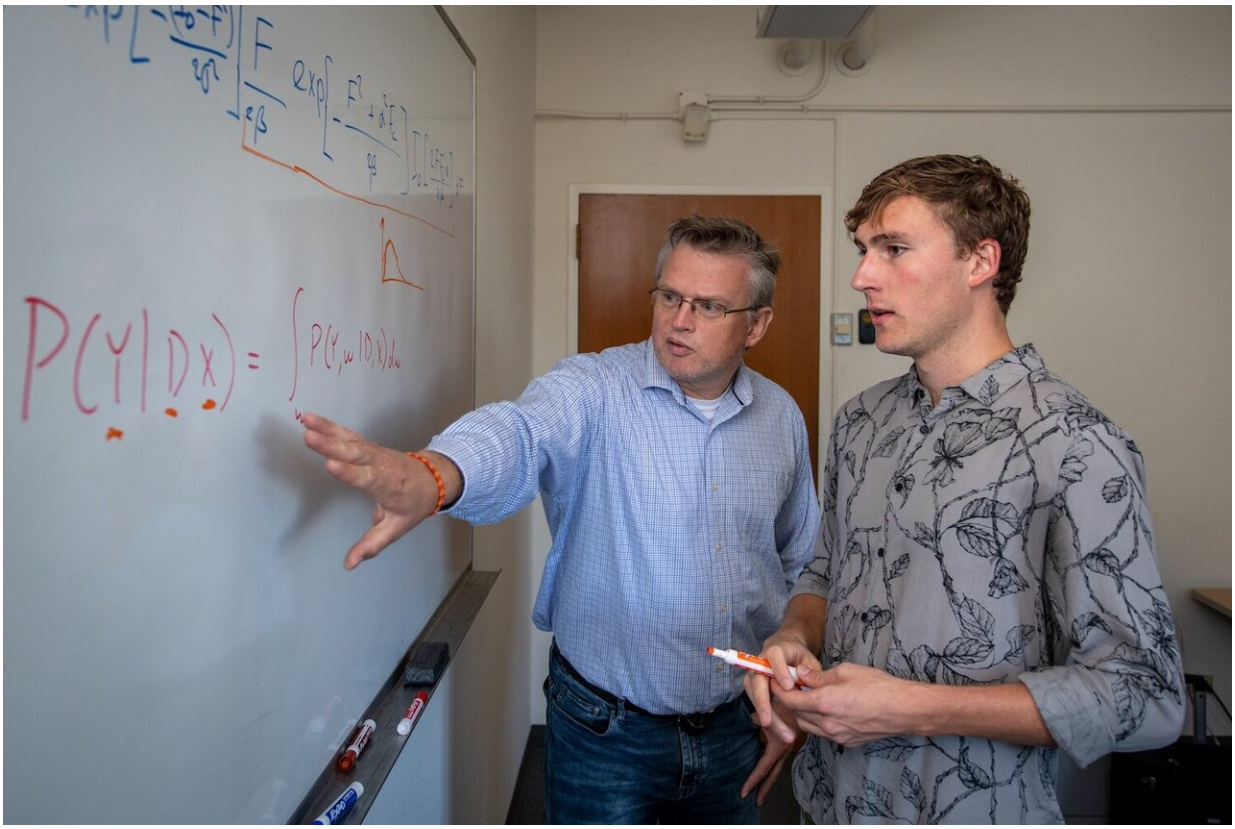
The elusive feline in this case was a mathematical problem that has bedeviled the experimental crystallography community for some time: how to model the presence of noise in data in a more realistic way.

Crystallography is an indispensable tool for determining the atomic structures of molecules—which in turn give researchers insights into their behavior and function. When a focused beam of light is aimed at a purified, crystalline sample, the light diffracts off of the atoms and a detector records the diffracted light. As the sample is rotated, two-dimensional images of the diffraction patterns are captured in various orientations. Algorithms are then applied to the diffraction data to reconstruct a three-dimensional map of the arrangement of atoms in the sample.

When you determine, or solve, a structure from diffraction data, you need to relate the model to your observations, explained Zwart, who is part of Berkeley Lab's Molecular Biophysics and Integrating Bioimaging Division. The target functions that are used to do this are called maximum likelihood functions. They work really well if your data are good, he notes, but when the amount of noise in the data increases—which becomes the case at higher resolutions—the current methods are not able to provide the best possible answer.

The reason target functions fall short in such cases is that there is one step in the calculation, an integration, that can't be done analytically—that is to say, with pencil-and-paper math that gives you an expression you can turn into code. Previous attempts to deal with this problem have either simply ignored the integration step, or come up with approximations that only work in experiment- or technique-specific scenarios. So Zwart and Perryman went back to basics, trying a multitude of different machine learning approaches to numerically derive as exact an approximation as possible in the most efficient way.

Three-quarters of the way through Perryman's 16-week internship, the two arrived at the conclusion that most of the paths that had seemed promising at the outset were actually blind alleys. "I would try things and it took a while just to figure out whether something is a success or a failure because, with a totally new problem, you just don't know," said Perryman. Things finally clicked when they realized that a common assumption people have been making for 30 years could be improved upon.



Univ. of Tennessee undergrad Elliot Perryman (on right) worked with Biosciences staff scientist Peter Zwart during his fall 2019 Berkeley Lab Undergraduate Research (BLUR) internship. Credit: Thor Swift/Berkeley Lab

The assumption has to do with the shape of the noise in the data. The widely accepted view has been that experimental errors fall into a classic normal distribution, like the Gaussian bell curve, where nearly 100 percent of observations fall within 3.5 standard deviations. But a more realistic curve has thicker "tails" owing to rare but predictable events. "Including these slightly more realistic error models in crystallographic target functions allows us to model the presence of what normally might be called outliers in a more realistic way," Zwart said.

Their method, which they published in the journal *Acta*

Crystallographica Section D: Structural Biology, is broadly applicable across the experimental crystallography field and will enable researchers to make better use of marginal or low-quality diffraction data. This research was supported by National Institutes of Health and CAMERA is funded by the U.S. Department of Energy's Office of Science.

A postdoctoral researcher in Zwart's lab is now working to turn the mathematical concept framework into an application that can eventually be implemented in the Phenix software suite. MBIB Director Paul Adams leads the development of Phenix, a collection of tools for automated structure solution that is widely used by the crystallography community.

"Elliot spent a lot of time and energy on approaches that ultimately did not pan out, but were crucial to the total effort because he was able to learn a lot himself and educate me at the same time," Zwart added. And the experience Perryman gained helped him land a follow-up internship working with Tess Smidt, a postdoc in the Computational Research Division, and ultimately a student assistant position working with CAMERA postdoc Marcus Noack on machine-assisted decision-making for experimental sciences.

The project Perryman and Noack have been working on aims to turn traditional methods of automated image sampling on their head. They propose using a random approach that is orders of magnitude more efficient and will give a prediction of how the image could look at some location, as well as an indication of the uncertainty of that prediction. Perryman has been working on a distributed optimisation approach, named HGDL (Hybrid Global Deflated Local), to improve a critical optimization function.

There are a lot of challenging computational problems in the biosciences that can be addressed with approaches that have already been developed

by applied mathematicians, Zwart noted. "Certain ideas just take a longer time to percolate into other areas," he said. "That's why working within CAMERA is so great: mathematicians have a different view on the world, a different set of skills, and read different papers. But they don't know the experimental fields like structural biologists do. It's important to bring these people together so that we can identify problems within the biosciences and find solutions within math and computing."

"That's been one of the big benefits of this internship," said Perryman. "I started out in nuclear physics, so I was just familiar with the types of problems in that field. But after working with Peter, or working with Tess this past spring, or Marcus, I realize there are so many analogous problems. Like, if you have the same problem, Marcus would frame it in terms of some sort of geophysics thing, and Tess would say that it's a geometry problem, but it's probably also a biology problem."

In the end, Perryman has not been deterred by any of these stubborn challenges: "There're so many interesting projects, it's hard not to get excited about them."

More information: Petrus H. Zwart et al. Evaluating crystallographic likelihood functions using numerical quadratures, *Acta Crystallographica Section D Structural Biology* (2020). DOI: 10.1107/S2059798320008372

Provided by Lawrence Berkeley National Laboratory