A Statistical Search for Genomic Truths

The computer scientist Barbara Engelhardt develops machine-learning models and methods to scour human genomes for the elusive causes and mechanisms of disease.

By Jordana Cepelewicz

Sarah Blesener for Quanta Magazine

Barbara Engelhardt, a Princeton University computer scientist, wants to strengthen the foundation of biological knowledge in machine-learning approaches to genomic analysis.

"We don't have much ground truth in biology." According to Barbara Engelhardt, a computer scientist at Princeton University, that's just one of the many challenges that researchers face when trying to prime traditional machine-learning methods to analyze genomic data. Techniques in artificial intelligence and machine learning are dramatically altering the landscape of biological research, but Engelhardt doesn't think those "black box" approaches are enough to provide the insights necessary for understanding, diagnosing and treating disease. Instead, she's been developing new statistical tools that search for expected biological patterns to map out the genome's real but elusive "ground truth."

Engelhardt likens the effort to detective work, as it involves combing through constellations of genetic variation, and even discarded data, for hidden gems. In research published last October, for

example, she used one of her models to determine how mutations relate to the regulation of genes on other chromosomes (referred to as distal genes) in 44 human tissues. Among other findings, the results pointed to a potential genetic target for thyroid cancer therapies. Her work has similarly linked mutations and gene expression to specific features found in pathology images.

The applications of Engelhardt's research extend beyond genomic studies. She built a different kind of machine-learning model, for instance, that makes recommendations to doctors about when to remove their patients from a ventilator and allow them to breathe on their own.

She hopes her statistical approaches will help clinicians catch certain conditions early, unpack their underlying mechanisms, and treat their causes rather than their symptoms. "We're talking about solving diseases," she said.

To this end, she works as a principal investigator with the Genotype-Tissue Expression (GTEx) Consortium, an international research collaboration studying how gene regulation, expression and variation contribute to both healthy phenotypes and disease. Right now, she's particularly interested in working on neuropsychiatric and neurodegenerative diseases, which are difficult to diagnose and treat.

Quanta Magazine recently spoke with Engelhardt about the shortcomings of black-box machine learning when applied to biological data, the methods she's developed to address those shortcomings, and the need to sift through "noise" in the data to uncover interesting information. The interview has been condensed and edited for clarity.

What motivated you to focus your machine-learning work on questions in biology?

I've always been excited about statistics and machine learning. In graduate school, my adviser, Michael Jordan [at the University of California, Berkeley], said something to the effect of: "You can't just develop these methods in a vacuum. You need to think about some motivating applications." I very quickly turned to biology, and ever since, most of the questions that drive my research are not statistical, but rather biological: understanding the genetics and underlying mechanisms of disease, hopefully leading to better diagnostics and therapeutics. But when I think about the field I am in — what papers I read, conferences I attend, classes I teach and students I mentor — my academic focus is on machine learning and applied statistics.

We've been finding many associations between genomic markers and disease risk, but except in a few cases, those associations are not predictive and have not allowed us to understand how to diagnose, target and treat diseases. A genetic marker associated with disease risk is often not the true causal marker of the disease — one disease can have many possible genetic causes, and a complex disease might be caused by many, many genetic markers possibly interacting with the environment. These are all challenges that someone with a background in statistical genetics and machine learning, working together with wet-lab scientists and medical doctors, can begin to address and solve. Which would mean we could actually treat genetic diseases — their causes, not just their symptoms.

You've spoken before about how traditional statistical approaches won't suffice for applications in genomics and health care. Why not?

First, because of a lack of interpretability. In machine learning, we often use "black-box" methods — [classification algorithms called] random forests, or deeper learning approaches. But those don't really allow us to "open" the box, to understand which genes are differentially regulated in

particular cell types or which mutations lead to a higher risk of a disease. I’m interested in understanding what’s going on biologically. I can’t just have something that gives an answer without explaining why.

Sarah Blesener for Quanta Magazine

“The goal of these methods is often prediction,” Engelhardt said, but added, “Prediction is not sufficient for the questions I’m asking,”

The goal of these methods is often prediction, but given a person’s genotype, it is not particularly useful to estimate the probability that they’ll get Type 2 diabetes. I want to know how they’re going to get Type 2 diabetes: which mutation causes the dysregulation of which gene to lead to the development of the condition. Prediction is not sufficient for the questions I’m asking.

A second reason has to do with sample size. Most of the driving applications of statistics assume that you’re working with a large and growing number of data samples — say, the number of Netflix users or emails coming into your inbox — with a limited number of features or observations that have interesting structure. But when it comes to biomedical data, we don’t have that at all. Instead, we have a limited number of patients in the hospital, a limited number of genotypes we can sequence — but a gigantic set of features or observations for any one person, including all the mutations in their genome. Consequently, many theoretical and applied approaches from statistics can’t be used for genomic data.

What makes the genomic data so challenging to analyze?

The most important signals in biomedical data are often incredibly small and completely swamped by technical noise. It’s not just about how you model the real, biological signal — the questions you’re trying to ask about the data — but also how you model that in the presence of this incredibly heavy-handed noise that’s driven by things you don’t care about, like which population the individuals came from or which technician ran the samples in the lab. You have to get rid of that noise carefully. And we often have a lot of questions that we would like to answer using the data, and we need to run an incredibly large number of statistical tests — literally trillions — to figure out the answers. For example, to identify an association between a mutation in a genome and some trait of interest, where that trait might be the expression levels of a specific gene in a tissue. So how can we develop rigorous, robust testing mechanisms where the signals are really, really small and sometimes very hard to distinguish from noise? How do we correct for all this structure and noise that we know is going to exist?

So what approach do we need to take instead?

My group relies heavily on what we call sparse latent factor models, which can sound quite mathematically complicated. The fundamental idea is that these models partition all the variation we observed in the samples, with respect to only a very small number of features. One of these partitions might include 10 genes, for example, or 20 mutations. And then as a scientist, I can look at those 10 genes and figure out what they have in common, determine what this given partition represents in terms of a biological signal that affects sample variance.

So I think of it as a two-step process: First, build a model that separates all the sources of variation as carefully as possible. Then go in as a scientist to understand what all those partitions represent in terms of a biological signal. After this, we can validate those conclusions in other data sets and think about what else we know about these samples (for instance, whether everyone of the same age is included in one of these partitions).

When you say “go in as a scientist,” what do you mean?

I’m trying to find particular biological patterns, so I build these models with a lot of structure and include a lot about what kinds of signals I’m expecting. I establish a scaffold, a set of parameters that will tell me what the data say, and what patterns may or may not be there. The model itself has only a certain amount of expressivity, so I’ll only be able to find certain types of patterns. From what I’ve seen, existing general models don’t do a great job of finding signals we can interpret biologically: They often just determine the biggest influencers of variance in the data, as opposed to the most biologically impactful sources of variance. The scaffold I build instead represents a very structured, very complex family of possible patterns to describe the data. The data then fill in that scaffold to tell me which parts of that structure are represented and which are not.

So instead of using general models, my group and I carefully look at the data, try to understand what’s going on from the biological perspective, and tailor our models based on what types of patterns we see.

How does the latent factor model work in practice?

We applied one of these latent factor models to pathology images [pictures of tissue slices under a microscope], which are often used to diagnose cancer. For every image, we also had data about the set of genes expressed in those tissues. We wanted to see how the images and the corresponding gene expression levels were coordinated.

We developed a set of features describing each of the images, using a deep-learning method to identify not just pixel-level values but also patterns in the image. We pulled out over a thousand features from each image, give or take, and then applied a latent factor model and found some pretty exciting things.

For example, we found sets of genes and features in one of these partitions that described the presence of immune cells in the brain. You don’t necessarily see these cells on the pathology images, but when we looked at our model, we saw a component there that represented only genes and features associated with immune cells, not brain cells. As far as I know, no one’s seen this kind of signal before. But it becomes incredibly clear when we look at these latent factor components.

You’ve worked with dozens of human tissue types to unpack how specific genetic variations help shape complex traits. What insights have your methods provided?

We had 44 tissues, donated from 449 human cadavers, and their genotypes (sequences of their whole genomes). We wanted to understand more about the differences in how those genotypes expressed their genes in all those tissues, so we did more than 3 trillion tests, one by one, comparing every mutation in the genome with every gene expressed in each tissue. (Running that many tests on the computing clusters we’re using now takes about two weeks; when we move this iteration of GTEx to the cloud as planned, we expect it to take around two hours.) We were trying to figure out whether the [mutant] genotype was driving distal gene expression. In other words, we were looking for mutations that weren’t located on the same chromosome as the genes they were regulating. We didn’t find very much: a little over 600 of these distal associations. Their signals were very low.

But one of the signals was strong: an exciting thyroid association, in which a mutation appeared to distally regulate two different genes. We asked ourselves: How is this mutation affecting expression levels in a completely different part of the genome? In collaboration with Alexis Battle’s lab at Johns

Hopkins University, we looked near the mutation on the genome and found a gene called *FOXE1*, for a transcription factor that regulates the transcription of genes all over the genome. The *FOXE1* gene is only expressed in thyroid tissues, which was interesting. But we saw no association between the mutant genotype and the expression levels of *FOXE1*. So we had to look at the components of the original signal we'd removed before — everything that had appeared to be a technical artifact — to see if we could detect the effects of the FOXE1 protein broadly on the genome.

We found a huge impact of FOXE1 in the technical artifacts we'd removed. FOXE1, it seems, regulates a large number of genes only in the thyroid. Its variation is driven by the mutant genotype we found. And that genotype is also associated with thyroid cancer risk. We went back to the thyroid cancer samples — we had about 500 from the Cancer Genome Atlas — and replicated the distal association signal. These things tell a compelling story, but we wouldn't have learned it unless we had tried to understand the signal that we'd removed.

What are the implications of such an association?

Now we have a particular mechanism for the development of thyroid cancer and the dysregulation of thyroid cells. If FOXE1 is a druggable target — if we can go back and think about designing drugs to enhance or suppress the expression of *FOXE1* — then we can hope to prevent people at high thyroid cancer risk from getting it, or to treat people with thyroid cancer more effectively.

The signal from broad-effect transcription factors like FOXE1 actually looks a lot like the effects we typically remove as part of the noise: population structure, or the batches the samples were run in, or the effects of age or sex. A lot of those technical influences are going to affect approximately similar numbers of genes — around 10 percent — in a similar way. That's why we usually remove signals that have that pattern. In this case, though, we had to understand the domain we were working in. As scientists, we looked through all the signals we'd gotten rid of, and this allowed us to find the effects of *FOXE1* showing up so strongly in there. It involved manual labor and insights from a biological background, but we're thinking about how to develop methods to do it in a more automated way.

So with traditional modeling techniques, we're missing a lot of real biological effects because they look too similar to noise?

Yes. There are a ton of cases in which the interesting pattern and the noise look similar. Take these distal effects: Pretty much all of them, if they are broad effects, are going to look like the noise signal we systematically get rid of. It's methodologically challenging. We have to think carefully about how to characterize when a signal is biologically relevant or just noise, and how to distinguish the two. My group is working fairly aggressively on figuring that out.

Why are those relationships so difficult to map, and why look for them?

There are so many tests we have to do; the threshold for the statistical significance of a discovery has to be really, really high. That creates problems for finding these signals, which are often incredibly small; if our threshold is that high, we're going to miss a lot of them. And biologically, it's not clear that there are many of these really broad-effect distal signals. You can imagine that natural selection would eliminate the kinds of mutations that affect 10 percent of genes — that we wouldn't want that kind of variability in the population for so many genes.

But I think there's no doubt that these distal associations play an enormous role in disease, and that they may be considered as druggable targets. Understanding their role broadly is incredibly important for human health.