FOR IMMEDIATE RELEASE

New algorithm may fuel vaccine development

Computational biologists harness machine learning to make sense of immune system data

LA JOLLA, CA—Immune system researchers have designed a computational tool to boost pandemic preparedness. Scientists can use this new algorithm to compare data from vastly different experiments and better predict how individuals may respond to disease.

"We're trying to understand how individuals fight off different viruses, but the beauty of our method is you can apply it generally in other biological settings, such as comparisons of different drugs or different cancer cell lines," says Tal Einav, Ph.D., Assistant Professor at La Jolla Institute for Immunology (LJI) and co-leader of the new study in *Cell Reports Methods*.

This work addresses a major challenge in medical research. Laboratories that study infectious disease—even laboratories focused on the same viruses—collect wildly different kinds of data. "Each dataset becomes its own independent island," says Einav.

Some researchers might study animal models, others might study human patients. Some labs focus on children, others collect samples from immunocompromised senior citizens. Location matters too. Cells collected from patients in Australia might react differently to a virus compared with cells collected from a patient group in Germany, just based on past viral exposures in those regions.

"There's an added level of complexity in biology. Viruses are always evolving, and that changes the data too," says Einav. "And even if two labs looked at the same patients in the same year, they might have run slightly different tests."

Working closely with Rong Ma, Ph.D., a postdoctoral scholar at Stanford University, Einav set out to develop an algorithm to help compare large datasets. His inspiration came from his background in physics, a discipline where—no matter how innovative an experiment is—scientists can be confident that the data will fit within the known laws of physics. E will always equal mc^2.

"What I like to do as a physicist is collect everything together and figure out the unifying principles," says Einav.

The new computational method doesn't need to know precisely where or how each dataset was acquired. Instead, Einav and Ma harnessed machine learning to determine which datasets follow the same underlying patterns.

"You don't have to tell me that some data came from children or adults or teenagers. We just ask the machine 'how similar are the data to each other,' and then we combine the similar datasets into a superset that trains even better algorithms," says Einav. Over time, these comparisons could reveal consistent principles in immune responses—patterns that are hard to detect across the many scattered datasets that abound in immunology.

For example, researchers could design better vaccines by figuring out exactly how human antibodies target viral proteins. This is where biology gets really complicated again. The problem is that humans can make around one quintillion unique antibodies. Meanwhile a single viral protein can have more variations than there are atoms in the universe.

"That's why people are collecting bigger and bigger data sets to try and explore biology’s nearly infinite playground," says Einav.

But scientists don't have infinite time, so they need ways to predict the vast reaches of data they can't realistically collect. Already, Einav and Ma have shown that their new computational method can help scientists fill in these gaps. They demonstrate that their method to compare large datasets can reveal myriad new rules of immunology, and these rules can then be applied to other datasets to predict what missing data should look like.

The new method is also thorough enough to provide scientists with confidence behind their predictions. In statistics, a "confidence interval" is a way to quantify how certain a scientist is of a prediction.

“These predictions work a bit like the Netflix algorithm that predicts which movies you might like to watch,” says Einav. The Netflix algorithm looks for patterns in movies you've selected in the past. The more movies (or data) you add to these prediction tools, the more accurate those predictions will get.

"We can never gather all the data, but we can do a lot with just a few measurements," says Einav. “And not only do we estimate the confidence in predictions, but we can also tell you what further experiments would maximally increase this confidence. For me, true victory has always been to gain a deep understanding of a biological system, and this framework aims to do precisely that.”

Einav recently joined the LJI faculty after completing his postdoctoral training in the laboratory of Jesse Bloom, Ph.D., at the Fred Hutch Cancer Center. As he continues his work at LJI, he plans to focus on the use of computational tools to learn more about human immune responses to many viruses, beginning with influenza. He's looking forward to collaborating with leading immunologists and data scientists at LJI, including Professor Bjoern Peters, Ph.D., also a trained physicist.

"You get beautiful synergy when you have people coming from these different backgrounds," says Einav. "With the right team, solving these big, open problems finally becomes possible."

The study, "Using Interpretable Machine Learning to Extend Heterogeneous Antibody-Virus Datasets," was supported by the Damon Runyon Cancer Research Foundation (grant DRQ 01-20) and by Professor David Donoho at Stanford University.

DOI: https://doi.org/10.1016/j.crmeth.2023.100540

###