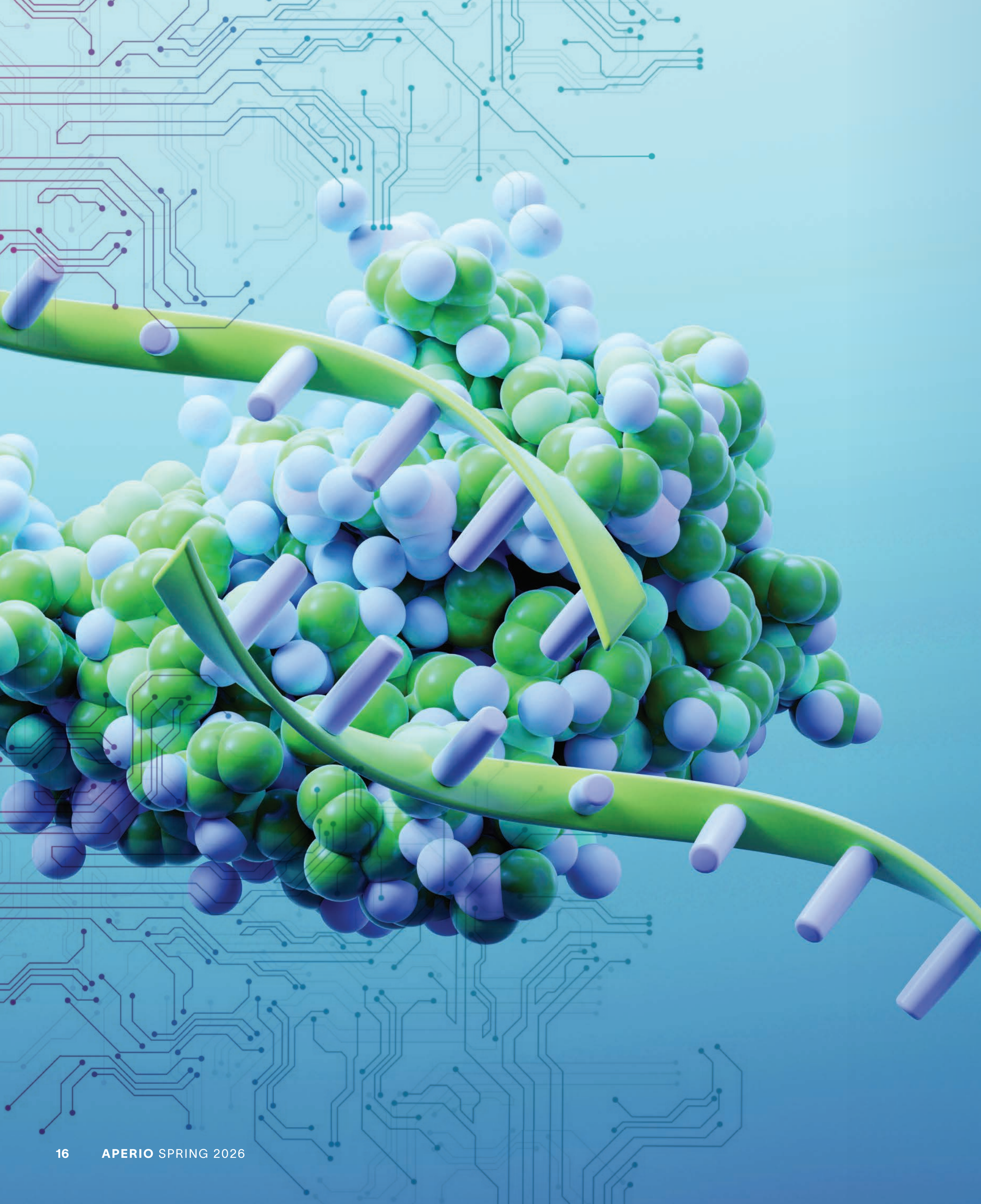


THE AI REVOLUTION IN THE LAB

WHILE CHATGPT GRABS HEADLINES, MACHINE LEARNING ALGORITHMS ARE QUIETLY TRANSFORMING SCIENTIFIC DISCOVERY — FROM DESIGNING CUSTOM PROTEINS TO UNRAVELLING THE GENETIC MYSTERIES OF AUTISM

By Simon Lewsen
Photographs by Naama Stern

When it comes to artificial intelligence, the buzz is all about large language models (LLMs), those massive, energy-hungry algorithms that, today, can write a middling undergraduate essay and, someday, will develop full sentience — at least according to what tech CEOs keep telling us.

But are they overhyped? Every day we encounter new reports about the disruptive — and unproductive — roles that LLMs are playing in our lives. We read about the AI lawyer that invents case law, or the AI book reviewer that recommends nonexistent books. We don't yet know what LLMs are reliably good for, but we do know one thing: they're hogging the media attention.

Meanwhile, the revolution may be quietly happening elsewhere. In scientific disciplines ranging from materials discovery to neurology to genetics, machine-learning technologies — usually algorithms that manage large, complicated datasets — have become essential partners in labs around the world. (Machine-learning technologies are a subfield of artificial intelligence.)

The superpower of machine learning algorithms is their ability to spot correlations that human brains cannot. "It's difficult for people to pick out complicated non-linear relationships," says Ariel Tennenhouse, an Azrieli Graduate Studies Fellow and PhD candidate in molecular biology

Azrieli Graduate Studies Fellow Ariel Tennenhouse is helping develop an antibody library — a database of high-quality proteins that degrade slowly, can withstand fluctuations in temperature and are suitable for human use.

For decades, scientists have dreamed of custom designing much-needed proteins that our immune system won't spontaneously create — antibodies against immune-resistant pathogens, autoimmune disorders or cancers

at the Weizmann Institute of Science, in Rehovot, Israel. "If I'm looking at a graph, there has to be an obvious visual trend there or I'll probably miss it. And I can only handle graphs with two — or, at most three — axes. Machine learning, however, can understand things on a multidimensional plane."

Few machine-learning or AI systems have the novelty factor that has made ChatGPT a massive pop-culture phenomenon, but many are promising research tools. Their value proposition lies in what they do now, rather than what they might do eventually. Not only are they enabling unprecedented breakthroughs, Tennenhouse argues, but they're changing the practice of science itself, from experimental design to data analysis. He has seen the changes up close.

Tennenhouse specializes in proteins, those intricate particles that swim in our blood, line our muscles and fortify our bones. We barely understand them. It wasn't until 1902 that German scientists Franz Hofmeister and Hermann Emil Fischer figured out what proteins are: macromolecules, that is chains of hundreds of smaller molecules, called amino acids, all folded in on themselves in impossibly dense configurations.

Proteins, we now know, are essential to our survival. Antibodies, the foot soldiers of the immune system, are a type of protein. When a hostile virus or bacterium shows up inside us, our bodies go to work, producing millions of novel antibodies in the hopes that one will be perfectly suited to the challenge at hand. The winning candidate will have just the right shape to bind to the enemy pathogen, fusing like pieces in a jigsaw puzzle. When our body lands on such a candidate, it ramps up production, launching a microscopic army to fight the infection.

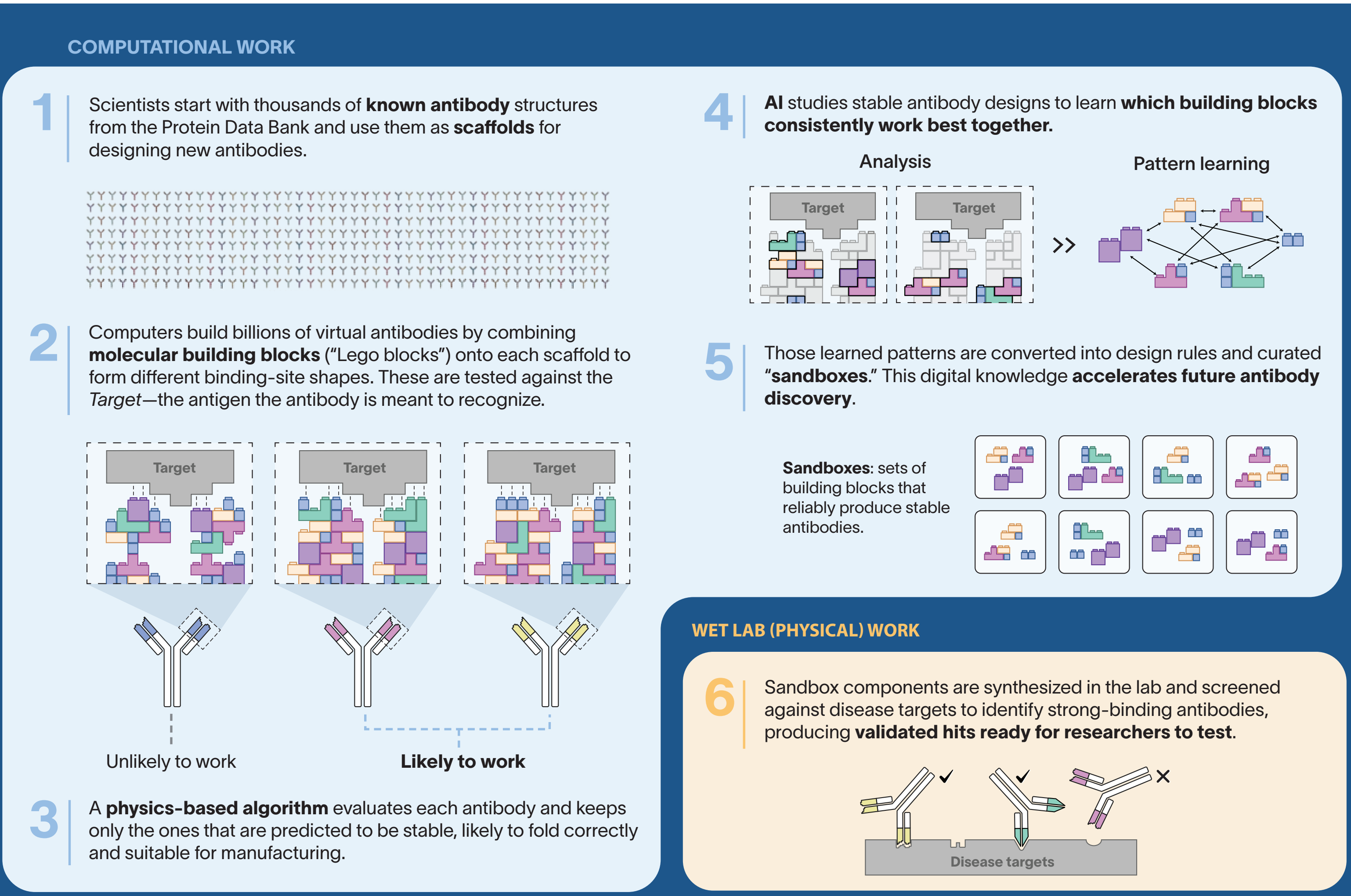
The sheer complexity of proteins is the best and worst thing about them. There are a lot of diseases out there, and pathogens are mutating all the time, so it helps that antibodies can be made in near-endless permutations. "You never know what's going to show up and attack you," says Tennenhouse. "So you want to produce antibodies against pretty much anything."

For decades, scientists have dreamed of custom designing much-needed proteins that our immune system won't spontaneously create — antibodies, that is, against immune-resistant pathogens, autoimmune disorders or cancers.

Over the past four decades, researchers have made strides toward that goal. Biochemist David Baker created algorithms to design proteins never seen before in nature on the computer, and molecular biologist Gregory Winter showed that antibodies generated in mice against a desired target can be adapted to be safe for human use. Winter later created libraries of billions of human-like antibodies that can be screened in a test tube. These are groundbreaking techniques (both Baker and Winter are Nobel laureates) but they have not resulted in a surfeit of promising antibodies. "Today," says Tennenhouse, "there are only 140 or so antibodies that are approved (by the U.S. Food and Drug Administration) for humans." Antibodies are some of the most complex proteins we know of, and the requirements for an antibody to be a therapeutic are extremely harsh.

This is why complexity is a problem. A protein contains hundreds of amino acid building blocks. The number of possible proteins —

The central challenge of building an antibody library is that little is known about what makes an effective antibody. Tennenhouse and his colleagues overcome this by using an advanced tool box that includes computational models based on physics, digital modelling and machine learning algorithms.



COMPUTATIONAL WORK

1 Scientists start with thousands of **known antibody** structures from the Protein Data Bank and use them as **scaffolds** for designing new antibodies.

2 Computers build billions of virtual antibodies by combining **molecular building blocks** ("Lego blocks") onto each scaffold to form different binding-site shapes. These are tested against the *Target*—the antigen the antibody is meant to recognize.

3 A **physics-based algorithm** evaluates each antibody and keeps only the ones that are predicted to be stable, likely to fold correctly and suitable for manufacturing.

4 **AI** studies stable antibody designs to learn **which building blocks consistently work best together.**

5 Those learned patterns are converted into design rules and curated "**sandboxes**." This digital knowledge **accelerates future antibody discovery**.

Sandboxes: sets of building blocks that reliably produce stable antibodies.

WET LAB (PHYSICAL) WORK

6 Sandbox components are synthesized in the lab and screened against disease targets to identify strong-binding antibodies, producing **validated hits ready for researchers to test**.

DIAGRAM BY ARIADNA VILLALBI

Tennenhouse holds a three-dimensional model of an antibody that neutralizes diverse flu variants. In time, he says, machine learning will enable better antibodies to reach patients more quickly and generate new insights from massive datasets. “We’re using AI as a hypothesis generator to suggest things we might investigate further.”

each with a unique folding pattern — is, for all intents and purposes, infinite. In theory, that’s good news: Somewhere within that near-infinite set, there must be a protein that’s compatible with human bodies, that’s heat resistant, that can be safely refrigerated and shipped around the world, and that’s a top-notch treatment for, say, pancreatic cancer. The trouble is finding it.

That’s where AI and machine learning come in. As a researcher at the Weizmann Institute’s Fleishman Lab, which is devoted to protein design, Tennenhouse is helping develop an antibody library — a database of high-quality proteins that degrade slowly, can withstand fluctuations in temperature and are suitable for human use. You can think of this library as the molecular equivalent of Central Casting in Hollywood. Film directors will call up the agency when they need, say, a gangster or a stunt double trained in martial arts. Central Casting has actors for all of these roles. And it offers a quality guarantee: its members have been vetted for professionalism and talent.

In Tennenhouse’s protein research, one of the central challenges is that we simply don’t know what makes an effective antibody. Out of the more than 10 quadrillion possible human antibodies, we have conclusive data on hundreds. This poses a serious issue for data-hungry machine learning algorithms. Tennenhouse overcomes this by first using computational models based on physics to check the quality of billions of possible antibodies. First, his team downloads thousands of antibody structures, which are digital renderings of the three-dimensional shapes that antibodies fold into. These structures have been painstakingly determined by experimentalists all over the world and are made available to researchers on a portal called the Protein Data Bank. Then, using digital modelling, the team tries to build antibodies that match these scaffolds. As its virtual building materials, the team uses the various molecular components — what Tennenhouse calls the “Lego blocks” — that exist within the human immune system. Eventually, they generate a digital dataset of billions of possible antibodies. A physics-based algorithm then goes through and scores each of these antibodies, analyzing their molecular components to predict how robust and stable each is likely to be.

Finally, the AI — in particular, an algorithm called a multilayer perceptron — gets to work, looking at all the possible combinations of Lego blocks and finding groups of them that combine well with every other member of the group. “You can think of these groups as sandboxes, where everyone plays nicely with everyone else,” says Tennenhouse.

Ultimately, the Fleishman Lab seeks to create a resource for fellow scientists. If you’re a researcher seeking an antibody for a given ailment, you’d be better off hunting in one of the lab’s curated sandboxes than trying to find a match within the galaxy of possible proteins.

You can even reproduce one of the sandboxes in a petri dish, using recent advancements in DNA synthesis to actually create the protein components (or Lego blocks). The components in your dish will combine easily with each other, creating a multitude of options, most of them strong, durable and appropriate for human use. You can then introduce the pathogen or cancer cell you’re interested in, to see if any proteins bind to it. Tennenhouse has used this strategy to discover antibodies against four unique targets, and the Fleishman Lab will make this technology available to any academic for free.

“You would come to us with a target disease,” says Tennenhouse. “And we would give you many antibodies, some of which will likely be reasonable therapeutic candidates.” Thanks to machine learning, an unmanageable dataset has been made manageable. Most importantly, this technology may allow better antibodies to get to patients faster.

Machine learning turbocharges the scientific process by performing analytical feats that would confound even the most gifted research teams

That goal — finding coherence within a sprawling, noisy field of data — is also essential to the work of Elad Dvir, an Azrieli Graduate Studies Fellow and a PhD candidate in genomics at The Hebrew University of Jerusalem. Like his peers in protein research, Dvir works in an area where our ability to generate information has vastly outstripped our capacity to manage it, at least until now.

Dvir’s main research interest is autism, a label researchers find maddeningly unspecific. Autism is incredibly heterogeneous. Some people on the autism spectrum are non-verbal and incapable of living independently; others live wildly successful lives. When you look at the brains of autistic people, you see a similar degree of variability. Some are macrocephalic, which means they are noticeably larger than the average human

In an effort to understand genetic pathways leading to autism, Elad Dvir uses gene-editing technology and machine learning approaches to study 39 genes with a known connection to autism. Algorithms have already uncovered intriguing leads.

brain; others are microcephalic, which means they're noticeably smaller. Among autistic people, those with microcephaly are more likely to have severe intellectual disabilities and comorbid intellectual delay. Does it make sense to lump all of these phenotypes into the same diagnostic category? "The diagnosis of autism is a psychological assessment," says Dvir. "But it's not clear that even the psychological aspects of autism are identical for everyone. Simply put, autism is not one thing."

We do know that autism is somehow related to genetics. Family studies confirm this fact. If your monozygotic twin has autism, for instance, you're much more likely than the average person to have the condition yourself. DNA-sequencing studies of people with autism show that they tend to have mutations in specific genes, suggesting that these genes play a role in the condition. Researchers have identified dozens to hundreds of genes that are likely connected with autism.

If we could learn more about the genetic pathways that lead to autism — or to the set of conditions we currently call "autism" — we could start making sense of things. That's what Dvir wants to do. As with Tennenhouse's work, Dvir's research requires him to generate reams of data.

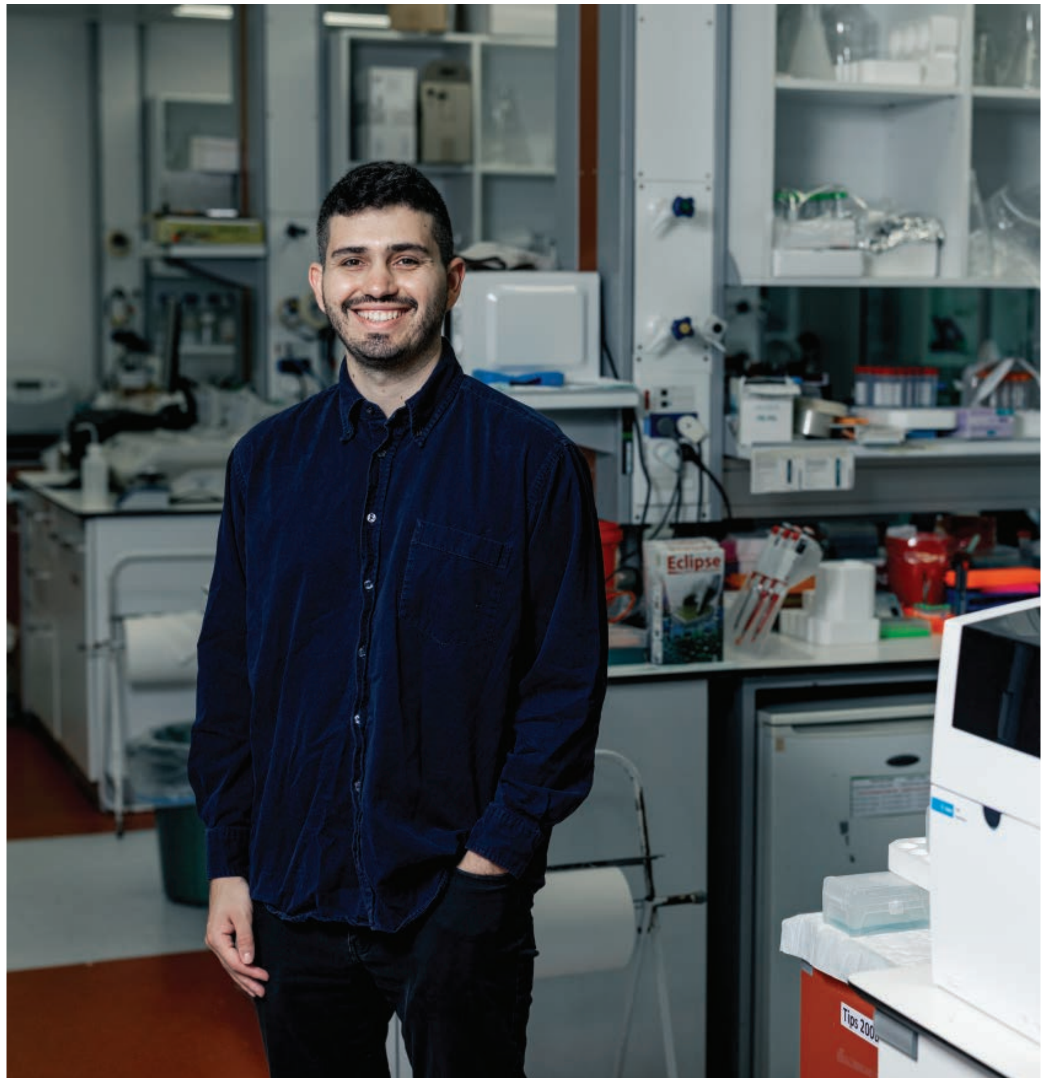
He has been studying 39 different genes with a known connection to autism. Using gene-editing technology, he generates genetic mutations in human embryonic stem cells. As the cells differentiate — that is, as they mature from all-purpose cells to specific cell types — he observes what kinds of cells they become.

Then he uses different statistical and machine-learning approaches to analyze the data. One example is a clustering algorithm, a classical machine-learning model that sorts data points into groups. In the chaos of data, the algorithm has identified a salient feature: Certain genes, when they mutate, make the cells more likely to differentiate into neurons, the building blocks of the brain. Others, when they mutate, make the cells more likely to differentiate into glia, helper cells that provide nutrition and insulation to neurons.

This correlation is striking. Dvir has noticed that the genes in the former group (the ones that, when mutated, differentiate more frequently into neurons) are known to be associated with macrocephalic forms of autism. The genes in the latter group are associated with microcephalic autism. Now, he has a workable theory: Perhaps the different types of autism are partially connected to genetic mutations that affect cell differentiation at the earliest stages of life.

By zeroing in on the underlying genetics of autism, Dvir hopes to better understand the condition and differentiate between autism subtypes. In time, we may find that the single diagnostic category of "autism" makes little sense at all. Perhaps "autism" is a bunch of conditions, all with similar symptoms but different genetic forerunners. "The computer can guide me to the most important features in my data," says Dvir. "From there, I might learn about the underlying biology of autism, which, until now, has been pretty mysterious."

AI and machine learning are exciting technologies not only for the possible discoveries they might yield but also for the way they are revolutionizing the scientific process itself.



Is it problematic for scientists to enlist AI not merely as a tool but as a partner? Will there be a time when scientists outsource their judgment to an alien intelligence?

For both Tennenhouse and Dvir, AI-based methods are disrupting the practice of science. Instead of formulating and testing hypotheses, modern researchers increasingly are in the business of generating data and letting AI draw the links.

"For centuries, science has been hypothesis driven," says Tennenhouse. First, you generate a single, falsifiable hypothesis, based on your intuitions about how things work. Then, you conduct an experiment to either affirm or refute it. If you get an affirmation, you expand your hypothesis; if you get a refutation, you refine it. After that, you conduct another experiment. And another. And another. Years become decades, decades become centuries, and gradually our understanding of the world increases.

Machine learning can turbocharge this process by performing analytical feats that would confound even the most gifted scientists. As a result, the practice of science itself is changing. Instead of formulating and testing hypotheses, Tennenhouse argues, scientists today are, increasingly, in the business of generating data. They don't pretend to know what the data means or how the individual data points might relate to one another. Tennenhouse, for instance, would never dream of guessing which building blocks in his vast library will combine well with one another, and Dvir would never speculate, off the cuff, about which cell-differentiation patterns are most germane to the study of autism. That work is for the algorithms. "We're using AI as a hypothesis generator to suggest things we might investigate further," says Tennenhouse.

Tennenhouse worries about the implications of this novel approach, whereby we are enlisting machines not just as tools but as partners in our scientific investigations. If an algorithm is generating your hypotheses for you, he points out, you may be unable to intuit the underlying logic: Ultimately, you are outsourcing your judgment to an alien intelligence. Then again, there are many crucial questions that we may never answer any other way. "For half a century," he says, "scientists have dreamed about designing proteins as therapeutics." It's high time we got some better results. ▲●■