Follow the Data

THE SUMMER BEFORE HER JUNIOR YEAR at Princeton, Alice Zhang ’10 had an experience that would change the course of her studies, career, and life. As an intern in a National Institutes of Health lab in Washington, D.C., her hometown, Zhang worked on what she describes as a classic molecular biology project: pipetting and doing Western blots — a lab method used to detect proteins — to study brain cancer. “I wasn’t very good at Western blots,” she says, laughing. One day, a lecture by an Austrian physicist turned Zhang’s classic experience on its head. “He showed a picture of a network of hundreds of different genes. It was then I realized diseases are not caused by a single gene in isolation, but by groups of genes in complex biological relationships. I couldn’t stop thinking about it,” says Zhang.

The next day, she approached the physicist, Stefan Wuchty, now an associate professor of computer science at the University of Miami, and asked how she could get involved with that kind of work. “He told me, ‘You need to know [the coding language] Python.’ I said, ‘What’s Python?’ He told me to Google it.” Zhang did, and then began learning to code. She used Python to build a map of how genes interact in brain cancer. Today, Zhang continues to chase the promise of networks of biological information, this time at

Verge Genomics, the drug-discovery startup she founded in 2015, just five years after graduating from Princeton.

Speaking on Zoom as COVID cases ticked up at the end of November, dressed in a baggy Verge Genomics fleece top and clutching a large mug of coffee, Zhang seems more like a college student than a biotech CEO. “I’m not a morning person,” she says, sheepishly, at 11 a.m. But her relaxed and warm demeanor belies a sharp sophistication. She speaks about her work in a melodic voice and precisely chosen words, explaining how a new approach to drug discovery could lead to desperately needed treatments for neurological diseases such as Parkinson’s, Alzheimer’s, and amyotrophic lateral sclerosis, or ALS (Lou Gehrig’s disease), a fatal disease that causes loss of muscle control.

ZHANG, A MOLECULAR BIOLOGY MAJOR at Princeton, kept returning to networks of genes. For her senior thesis, she used computational tools to study gene networks to try to understand cancer-drug resistance. After graduation, during an M.D.-Ph.D. program at the University of California, Los Angeles (UCLA), she coded algorithms that analyzed networks of genes to predict new drugs for nerve regeneration. The algorithms looked at gene networks that distinguished nerves that could recover after injury from nerves that couldn’t, “with the idea that if we knew the networks, we could reprogram the nerves,” says Zhang. This

approach veered from the standard approach taken by scientists at the time, who would either test a series of hypotheses about what causes disease or try a variety of drugs on a cell model of disease to see if something worked. Instead, Zhang's group first tried to understand the root cause of nerve injury and regeneration, looking directly at the biological pieces involved. When used on mice, the drug predicted by the UCLA group helped them recover nerve function four times faster than the leading therapy at the time.

Zhang saw firsthand the power of a computational, data-driven approach to discovering drugs. "I thought, I could publish this in a paper, but is that the best way to get new drugs to patients?" she says. She felt this kind of work would have the greatest impact on patients if she started a company. And she believed she should not wait to do it. She left the M.D.-Ph.D. program without finishing.

Zhang and another M.D.-Ph.D. student, Jason Chen, now a neurosurgery resident in Boston, started Verge Genomics during their fifth year in the program. The two turned to the tech-startup incubator YCombinator, known as YC, to get the company off the ground. YC provides funding and support for three months to fledgling startups, with a chance for companies to present in front of investors. (Other companies that got their start at YC include Airbnb, Dropbox,

and Instacart.) Zhang says she learned about YC through two Google searches: The first was, “What is an incubator?” And the second was, “What is the best incubator in the world?”

Zhang and Chen told YC they saw an opportunity to “parallel process” in drug discovery — to look at hundreds of genes instead of one at a time, and see the genes change together. “Our algorithm was similar to Google’s PageRank, at the time,” says Zhang, offering a tech lens through which to view drug discovery. Before PageRank, webpages were ranked by how often individuals clicked on them. Taking a different approach, PageRank looked at how webpages linked to other pages, and how relevant those linkages were. “We wanted to do the same thing for genes and their relevance in disease,” says Zhang. She and Chen named their company Verge, suggesting the concept of networks, with their nodes and edges. “Verge is a synonym for edge,” says Zhang, “but there’s a double meaning.” With a platform based on networks, the company was on the verge of a watershed moment in drug discovery.

THE PROCESS OF DRUG DISCOVERY AND development is an arduous one, often stretching over a decade. The industry trade association Pharmaceutical Research and Manufacturers of America (PhRMA) estimates the average cost to develop a successful drug to be $2.6 billion.

Scientists start by searching for a molecular structure, typically a protein or set of proteins, that influences disease. Such a structure will become the intended target of a drug that is either built from scratch or selected from a library of existing drugs. (On rare occasions, as with the drug lithium, used to treat bipolar disorder, the actual target and how it works remain unknown to scientists at the time of discovery. At other times, the final target is unintended, as in the case of Viagra, originally formulated to treat blood pressure.) The process of selecting a target and finding a drug that appears to work in lab tests can take up to six years. From there, scientists develop the drug by testing it in additional lab settings, in animals, and finally in people through a series of clinical trials that can take another six to seven years, if not longer. Most drugs fail because at the end of a clinical trial, researchers find out they do not work. In fact, more than 90 percent of drugs fail in clinical trials. “I don’t think people appreciate by and large what really goes into this, and what it means when we say that 90 percent of drug programs fail,” says Ron Cohen ’77, president and CEO of Acorda Therapeutics, a biotech company focused on neurological disorders such as multiple sclerosis.

Drugs don’t work for many reasons. Sometimes the target isn’t right in the first place. Scientists have had to rely on incomplete information to find drug targets and might lack a full picture of a disease. It’s similar to what happens in the

parable of the blind men and the elephant: One blind man touches the elephant's trunk and declares he's touching a snake. Another blind man touches the elephant's leg and says no, it's a tree trunk. "Thanks to technological breakthroughs in the last decade, we can now just take a picture of the elephant," says Zhang.

Now, there are vast new sources of information. Scientists can study an organism's entire biology through DNA sequencing (genomics), data on how genes turn on and off (transcriptomics), and protein catalogs (proteomics), to name a few. These bodies of information and the computational tools needed to process and interpret them can give scientists a more complete picture than ever before of what's going on biologically in a disease, without resorting to a guess or a hypothesis.

This shift, from hypothesis-driven, reductionist science to discovery-based, data-driven science is one of the legacies of genomics, says former Princeton President Shirley Tilghman, professor of molecular biology and public affairs, emerita. Another hallmark of the genomic era is a deep integration of computational and biological expertise, says Tilghman, including the introduction of computational tools in undergraduate research. "It's now impossible to get a degree in any serious molecular biology department without being able to use computers in a sophisticated way," she

says.

Zhang is banking on the idea that improvement in the initial stages of drug discovery, via better target identification, will produce drug molecules more quickly, with greater success down the line in clinical trials. However, even the best targets may not yield drugs that sail through clinical trials, as Cohen knows personally through his experience in developing Acorda's MS drug, Ampyra. While the drug functioned beautifully in the lab, like a liquid bandage on parts of damaged neurons, it failed to show clinically meaningful improvements in walking speed in patients with MS. Cohen analyzed data from a trial by hand and identified the patients whose walking speed increased consistently in at least three of four visits during the study period. These "responders" showed a statistically significant improvement over people who received a placebo, but in a previous analysis, that result had been buried because of the variability in how MS affects a patient's function on any given day or time of day. Later trials replicated Cohen's post-hoc analysis and led to the drug's approval. "It's a constant, bedeviling issue, to come up with correct clinical trial endpoints and know how to measure them, especially in neurological disease," says Cohen.

Neurological disease, in fact, has been the thorniest frontier in drug discovery. With an aging population, rates of

diseases like Alzheimer's, Parkinson's, and ALS are on the rise. In these diseases, parts of the nervous system — the brain, spinal cord, nerve roots, and peripheral nerves — die. The results include memory loss (Alzheimer's), muscle wasting (ALS), and an inability to control movement (Parkinson's). The cause of these diseases is unknown, but likely lies in a murky soup of genes and environmental factors that have thus far been impossible to tease out. There are no cures. ALS is especially cruel, afflicting both the young and old, and leaving patients with a life expectancy of just two to five years.

"Neurological diseases are complex. For many, we don't know the etiology. There's tissue damage and progressive degeneration. It's happening in the nervous system, the most complicated system in the body," says Cohen. Animal models of disease are typically a starting point for drug discovery, but often lose their value in neurological diseases. For example, the core biology behind Alzheimer's may not be found in mice. "Animals are not humans," says Cohen, "and the differences can turn out to be important."

In addition, unlike a disease such as diabetes, defined by blood-sugar levels and thus treated by targeting those levels, neurological diseases present a fuzzy link between a biological feature of the disease, like sticky plaques in the brains of Alzheimer's patients, and disease manifestation.

While Biogen's Alzheimer's drug Aduhelm targets these plaques and recently won accelerated FDA approval last June (despite controversy over its effectiveness and safety), it remains to be seen whether removing plaques will improve the lives of Alzheimer's patients. Prior to Aduhelm, neurological disease drug failures scarred the pharmaceutical landscape, with more than 150 Alzheimer's drugs failing in clinical development since 1998. Big players have now shifted their investments to other diseases. Twenty-five years ago, just about every major drug company had a neurological drug program; by 2020, only a handful of companies did. The exit of larger drug companies coupled with advances in genomics and computation make today's environment a "moneyball" moment for neuroscience drug discovery, says Zhang, referring to Michael Lewis '82's book of that name, about data's disruptive effects on baseball.

ON THE NINTH FLOOR OF 2 TOWER PLACE IN South San Francisco, near the pharmaceutical giants Genentech and Amgen, Verge's 30 employees are taking what they call an "all-in-human" approach to drug discovery, focusing on the elusive goal of drug treatments for neurological disease. Rather than starting with animal models of disease, Verge begins with human data, and lots of it. Zhang led Verge to secure over a dozen partnerships with hospitals, academic centers, and biobanks — repositories of post-mortem human tissue from diseased individuals. From those partnerships,

Verge began building up its own proprietary databases from human tissue, mainly brains and spinal cords, as these would be the most illuminating for understanding the full picture of neurological disease. To find biological gems in the mountains of data, the scientists cannot analyze the data by hand — there's too much of it. Instead, they must use an artificial-intelligence approach — machine learning. Algorithms use "training" datasets containing known linkages between DNA sequences, gene expression patterns, and disease to learn how to find new, powerful biological associations in the datasets of interest. Machine learning requires data of the highest quality so that the algorithms can separate a biological signal from noise. "You can have the most sophisticated algorithms in the world, but you need high-quality data to feed in and train them on," says Zhang.

More than 6,800 human tissue samples make up Verge's human datasets, one of the largest collections of any drug-discovery effort in neurodegenerative disease. Data on gene expression, genotyping (specific DNA sequences), and patient characteristics like age, demographics, and disease progression are fed into machine-learning algorithms that look for associations between networks of genes and disease. Rather than going in with a hypothesis about what's causing disease, "we let the patient data tell us the target," says Victor Hanson-Smith, head of computational biology at

Verge.

When put to the test on human tissue, Verge's algorithms identified a network of 200 genes that were consistently suppressed in patients with ALS, compared to people without the disease. The genes were involved in a biological process called the lysosomal pathway that clears the cell of the protein aggregates seen in patients with neurodegenerative disease. "Most people have been focusing on the toxic protein aggregates, but we're saying, let's fix the root cause — the lysosomal pathway," says Zhang. Verge scientists came up with a handful of drug candidates that could hit the lysosomal pathway. One of them slowed ALS progression in mice, and in a petri dish, the drug rescued dying human neuron cells. This drug candidate is expected to head to clinical trials this year, three years faster than the six years typical in traditional drug discovery.

Verge's all-in-human approach has multiple benefits, Zhang says. Testing drug candidates in human cells in a dish ensures baseline efficacy before jumping into clinical trials. And once the drugs are tested in human cells, the data flow back into the machine-learning algorithms to further improve their predictive power in target identification.

Last July, Verge began a three-year collaboration with Eli Lilly to develop additional ALS targets. And on the heels of Verge's ALS candidate are drug candidates for Parkinson's.

ALTHOUGH OPERATION WARP SPEED ACCELERATED COVID-19 vaccine development from decades to mere months, it isn't clear what impact this will have on other drug programs. "How many marathons can you run simultaneously, and who chooses what to accelerate?" asks Kenneth Moch '76, senior adviser to the chairman at the Center for Global Health Innovation and the Global Health Crisis Coordination Center, both nonprofits. He also has led several biotech companies.

Now, Zhang says, a second wave of companies is aiming to use a data-driven approach, like that used by Verge, to decipher the biology of disease. She hopes these technologies will improve drug targets and dramatically reduce the time and expense involved in developing new medications. With the cost of full genome sequencing falling from more than $100 million per genome 20 years ago to just $1,000 per genome today, garnering vast amounts of data is more afforadable than ever. "With the rapidly advancing technologies of cell therapy, gene therapy, and machine learning," she says, "we are going to see big changes in the next 10 to 20 years."

Susan Reslewic Keatley '99 is a freelance science writer and aspiring medical-thriller author.