Directed evolution has clearly shown great success in engineering antibodies with new functions, some of which are revolutionizing how we manage disease. But the use of directed evolution is not limited to engineering antibodies. In fact, virtually any protein scaffold can be used. Like natural evolution, variation followed by selection is the name of the game. With twenty natural amino acids to choose from at any given position within a protein, the possibilities for engineering new proteins are endless. Consider a tiny protein made up of five amino acids. Since each of the five positions can be any of the twenty amino acids, the number of possible proteins is $20 \times 20 \times 20 \times 20 \times 20$, or 20^5. This means that there are 3.2 million different combinations of amino acid sequences that can be used to make a protein only five amino acids long. Of course, most proteins are far longer than five amino acids. Even a small protein like insulin contains 51 amino acids, while hemoglobin has 574 amino acids. If we take a protein that is 100 amino acids long—not very long by most protein standards—the number of different combinations of amino acids would amount to 20^{100} or 1×10^{130}. That's 1 with 130 zeros! This is an unimaginably large number. In fact, there's not enough matter in the entire observable universe to make every single amino acid combination.[17] This tells us that even with the vast diversity of proteins in all of the animals, plants, fungi, and bacteria that exist today, and have ever existed, nature has explored only a minuscule proportion of the potential amino acid sequence combinations—only a snowflake at the top of the amino-acid-sequence iceberg. This is good news for protein engineers. It means that there will be many, many new ways to string amino acids together in ways that nature has never seen before.

THE POWER OF ONES AND ZEROS

Some of us are old enough to remember the earliest ways we used to connect to the Internet. A computer was hooked up to a dial-up modem through a telephone line, making strange beeps and boops in order to communicate with the World Wide Web. Even a small, heavily pixelated image took minutes to load on the screen, and sending a short message was a delicate and unreliable affair, especially if someone in your home decided to pick up the phone, interrupting the entire process. But computers became faster. Processor speeds made computers more powerful, and wireless technology made communication lightning fast. With the exponential growth in computational power, scientists began to explore the use of computers in the design of proteins. After all, if the protein design problem could be reduced to a set of calculations, then computers, which are in essence giant calculators, should be able to help solve the problem much more quickly.

Initial work on computational protein design used computers to ask limited questions, mostly modeling how mutations could affect the structure of a protein. But in 1997, Stephen Mayo's group at Caltech reported the first protein completely designed by a computer.[18] The new protein, named FSD1 (short for "full sequence design 1"), was just twenty-eight amino acids long. Mayo was one of the early pioneers of computational protein design. At the time FSD1 was made, he was the only Black faculty member at Caltech.[19] Later, he would become the first Black professor to earn tenure there.[20]

Mayo wanted to answer a big question: could a fully folded protein be made entirely by computational methods? He began with the backbone structure of a protein called a zinc finger. The protein is composed of a single alpha helix and two beta strands surrounding one zinc ion. Mayo hoped to strip the original zinc finger protein of its amino acids and to replace them with new ones without disrupting the overall structure. This is a challenging endeavor considering that there are over two trillion trillion trillion different combinations of amino acids that could be put together to create the twenty-eight amino acid–long protein chain. With the computational power at the time, it was nearly impossible to try every single combination of amino acids and systematically determine which ones to test. So instead, Mayo came up with a clever solution to scan through the possible solutions to this problem. He and his coauthor, Bassil Dahiyat, used a computer algorithm known as dead-end elimination (DEE). Rather than try to find the best answers to the problem, the algorithm searched for amino acids that could not possibly be part of the final solution and eliminated them. For example, if an amino acid was too large to fit at a specific position, then the algorithm would reach a dead end, concluding that that specific amino acid chain would not work. The process was repeated over and over, pruning the wrong answers, gradually simplifying the problem as wrong answers were eliminated, and ultimately leading to a reasonable solution. Using the DEE algorithm, Mayo and Dahiyat came up with a new sequence that was different from the original zinc finger but had an identical fold. The work became a milestone in computational protein engineering and marked the beginning of a now flourishing field.

The first protein to be designed completely by a computer program was FSD1. The small protein, composed of one alpha helix and two beta strands, is based on the structure of a zinc finger fold.

After Mayo published his work, researchers began to improve the computational algorithms and use smaller and faster machines as the technology raced forward with better hardware and software. In 2003, David Baker's group at the University of Washington described the design of a new protein fold, a structure never seen before in nature. The protein was ninety-three amino acids long and combined alpha helix and beta sheets in a new topology that they named Top7.[21] Baker's group continued to show resounding success in the field of computational protein design. In 2008, the group reported the design of a brand-new enzyme using computational methods.[22] This was an accomplishment the entire field had been anticipating, and the publication became an instant hit, launching Baker to the forefront of computational protein design. The enzyme reported by the group was shown to carry out a reaction known as Kemp elimination, whereby the protein breaks a bond between a nitrogen and an oxygen in a small organic molecule. The reaction by itself is slow, but with the newly engineered enzyme, the reaction goes a hundred thousand times faster. The group then used directed evolution to improve the designed protein, increasing its activity by an additional ten-fold. That same year, the group repeated their success by reporting a new set of computationally designed enzymes.[23] This time the enzymes were designed to catalyze a retro-aldol reaction, which is often used in carbohydrate metabolism. The group tested seventy-two different designed enzymes, of which thirty-two were able to carry out the reaction. This represented a massive success rate. The group also demonstrated that their designed enzymes could perform the reaction using four different sets of amino acid sequences. Most importantly, the structures of the newly designed proteins were nearly identical to what the computational algorithms had predicted. Using new enzymes to carry out these reactions provides an eco-friendly way to replace traditional methods, which usually require harsh or dangerous chemicals.

David Baker has continued to demonstrate novel ways in which computational algorithms can be used to design proteins with new structures and functions. In 2018 his group designed its own version of the green fluorescent protein found in jellyfish.[24] Unlike GFP, the designed protein could not produce fluorescence on its own. Instead, its beta-barrel structure was designed to surround a fluorescent dye similar to the fluorescent core of the naturally occurring GFP. This novel protein represented a new level of success for computational protein engineering techniques, offering ways to fine-tune and even improve the fluorescent properties of proteins found in nature. The ability to continually engineer new types of proteins and raise the bar for protein engineering requires a great deal of creativity and a network of dedicated scientists working together as a team.

Protein engineering also requires a massive amount of computational power. Due to the incredibly complex nature of proteins, the computational process requires multiple processors working together, each solving little bits of the immense number of calculations. Even with large computational facilities containing hundreds of processors, the demand for computational time is always high. Because of this, David Baker and his group have enlisted the help of the public in carrying out the intense protein design process.[25] Their application, called Rosetta@home, enables people all over the world to help solve protein-design problems.[26] Participants allow researchers to use their personal computers for protein-structure prediction calculations when the computers are otherwise not in use. With more than 1.3 million users on 250,000 active hosts worldwide working together, Baker's group has been able to crowdsource research on protein engineering. The group also developed Foldit, an online platform that has turned protein folding into a collaborative game accessible to players around the world.[27] Participants, whether experienced scientists or just curious gamers, engage in virtual challenges to predict the most stable and biologically functional structures for various proteins. This crowd-driven approach to protein folding has not only expanded the pool of contributors, advancing the fields of biotechnology and computational biology; it has also fostered interdisciplinary collaboration and public engagement in scientific research. But the Baker lab hasn't focused just on surmounting general protein design obstacles. They've also set their sights on developing novel therapeutics to fight some of the greatest health threats of our time.

In 2020, as the COVID-19 pandemic spread across the world, hospitals were overwhelmed with patients, and pharmaceutical companies raced to develop a vaccine. Many scientists shifted their attention to studying how the virus infects humans and how infections can be prevented or treated. In the Baker lab, members were focused on engineering proteins that could help in the battle against the virus. The key to the COVID-19 virus's ability to infect the human body is a protein on its surface known as a spike protein. This is how the virus latches on and gains entry to the inside of the cells of the respiratory tract, where it proliferates and causes disease. The Baker group wanted to block the spike protein from binding to receptors on the surface of human cells. The group designed a set of small proteins that were able to latch onto the spike protein, specifically the parts of the protein responsible for its invasive abilities.[28] The group used computational algorithms to build the tiny proteins with the right shape-complementarity for the COVID-19 spike protein, to try to mimic how an antibody can block a target protein. The researchers knew that the spike protein interacts with a receptor on our lungs known as ACE2. The algorithms started with a 3D model of the part of ACE2 that the spike protein latches onto, then began to build around it in the hope of making proteins that could disrupt the interactions between the spike protein and ACE2. The newly designed proteins latched onto the spike protein with high affinity and in a specific manner.[29] To test if the designed proteins could protect from a COVID-19 infection, the protein was delivered to mice through their nostrils before the mice were exposed to the virus. The results showed that the engineered proteins were sufficient to protect the mice from infection and were effective against several different strains of COVID-19, offering hope for their use as prophylactics in humans.

Over the past three decades, computational scientists have made great strides in the design of new proteins. Yet one major problem has persisted: how to predict the structure of a protein based on only its amino acid sequence. This is called "the protein folding problem." There is no model available today that can predict how a string of amino acids will fold into its functional form, or if it will fold at all. This may seem like a minor problem; after all, proteins naturally fold into precise 3D structures, and they do this in a fraction of a second. But for scientists creating a protein shape from a string of amino acids, knowing how this happens is essential. Even with a very short protein, there are nearly endless ways in which the different parts of the protein can come together as it wiggles and twists. The incredible flexibility of the bonds connecting atoms in the protein compound the problem. Twisting just one bond in the wrong way can have immense consequences for the way the entire protein folds.

For years, scientists grappled with this seemingly unsolvable problem, until AlphaFold came on the scene. The brainchild of DeepMind, the artificial intelligence powerhouse created by Google, AlphaFold quickly became a revolutionary force poised to transform our understanding of protein folding.[30] The turning point was during the Critical Assessment of Structure Prediction (CASP) competition—an Olympic-like event for biochemists and computational biologists striving to predict protein structures. Every year since 1994, CASP participants have raced to see who can most accurately predict protein structures when given only their amino-acid sequences.[31] The participants receive a score on a hundred-point scale reflecting how well their predicted structures match the experimentally determined protein structure. Throughout the history of the competition, the scores for the more challenging protein structures have hovered between forty and sixty, meaning that they did not match well with the actual structures of the proteins.

Using neural networks and other machine learning algorithms, AlphaFold began to decode the language of proteins and the complex symphony of interactions, forces, and shapes they adopt. The 2020 CASP competition was AlphaFold's moment, and as the competition unfolded, AlphaFold dazzled the scientific community with its results. While much of the world was still under quarantine, an announcement was made to the AlphaFold team over a Zoom meeting. The group had not only won the competition; they had also exceeded all expectations. The group's AI algorithms were able to predict the structure of several dozen proteins, given only their amino-acid sequences, with a median score of eighty-seven, a full twenty-five points above the next highest score. The structures generated by AlphaFold's AI algorithms were nearly identical to those determined by experimental methods. The groundbreaking accomplishment inspired protein scientists and biologists worldwide. It wasn't just a victory in a competition; it was a paradigm shift in our ability to decipher life's most fundamental building blocks.

AlphaFold's success lay in its exquisite ability to learn from the two hundred thousand or so known protein structures, finding patterns and deciphering the rules that ultimately govern how proteins fold. This is similar to how AI tools that generate images work. AlphaFold treats protein structures as any three-dimensional object. It learns the structural patterns from the library of available structures, assessing the propensity of certain sequences for forming alpha helices or beta sheets, then generates shapes from new sequences with unknown structures using what it has learned. But image learning is not the only feature of AlphaFold. Like the AI tools used to generate language, another component of AlphaFold treats the sequences of proteins as words or sentences, deciphering their patterns and assigning them structure the same way a text-generating AI tool assigns meaning to words.

Since the development of AlphaFold, many new tools involving AI have emerged, each with new capabilities for specific applications. Some tools allow scientists to see how two different proteins fit together, for example, how a receptor on a cell surface receives a signal from a hormone. Other tools can be used to quickly scan through millions of drug molecules and see which best fits a protein structure, providing insight into drug design. What's most impressive is that tools like AlphaFold are freely available for anyone to use. There is no need for large computational facilities with multiple parallel processors. Anyone can use a laptop or phone to plug a sequence into the AlphaFold online tool and, within minutes, get an incredibly accurate picture of what the structure probably looks like. This combination of AI and biochemistry puts proteins in the hands of more people, redefining who can participate and create. These new tools are democratizing scientific research by increasing accessibility and removing the traditional barriers to who can become a scientist.

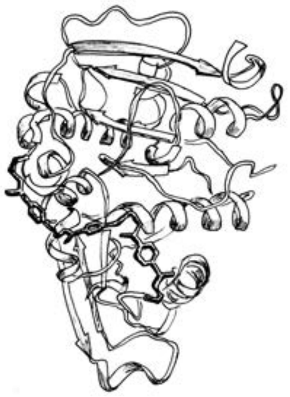
The revolutionary work of David Baker and the team of scientists who developed AlphaFold did not go unnoticed. In 2024, the Nobel Prize for chemistry was awarded to David Baker for his efforts in using computers to design new proteins, and to Demis Hassabis and John Jumper of Google DeepMind for protein structure predictions. The prize recognized the power of AI to help us understand how a sequence of amino acids folds into a functional three-dimensional structure. It also underscored the incredible potential for designing new proteins that nature has never seen before.

ENGINEERING THE FUTURE

Beyond their importance for human health, engineered proteins may become our most powerful weapon against climate change. With the rising temperature of the planet, an increase in greenhouse gas emissions, and the huge amounts of pollutants and plastic waste produced every year, protein engineering may offer hope for a cleaner future. One promising application for protein engineering is the detection of pollutants. Scientists are now turning proteins into "biosensors" that can detect a wide range of molecules, especially those that negatively impact the environment.

Recently, our group at Indiana University engineered a bacterial protein to bind to glyphosate, the active ingredient in RoundUp, the most extensively used herbicide in the world. Each year, 300 million pounds of glyphosate are sprayed on crops, much of which ends up in groundwater, lakes, rivers, and streams. While initially considered harmless, glyphosate has been associated with negative health effects and was classified as a carcinogen by the state of California in 2017. Starting with a bacterial protein that binds to a naturally occurring molecule known as 2-aminoethyl phosphonate, we made mutations in the protein to change its composition, altering its function and turning it into a glyphosate-binding protein.[32] When a fluorescent molecule was attached to the engineered protein, the amount of light emitted by the fluorescent chemical indicated the presence of glyphosate in soil or water.

Protein engineering also offers a new avenue for carbon capture. Each year, coal-burning plants emit nearly nine billion metric tons of CO_2 into the atmosphere, adding to the problem of climate change and contributing to a warmer planet. Current technologies for capturing carbon dioxide as it travels through the chimneys at these plants require large amounts of energy. To tackle this problem, scientists recently sought the help of the enzyme carbonic anhydrase, which naturally combines CO_2 with water to make bicarbonate, known to most of us as baking soda. Carbonic anhydrase is one of the fastest enzymes on the planet, with a rate of about one million reactions per second.[33] Unfortunately, the enzyme is not tolerant of the harsh conditions inside the reaction center of a chimney. The enzyme must be able to withstand temperatures near the boiling point of water and harsh chemical conditions that would denature most proteins. To overcome these challenges, scientists from the biotech firm Codexis used directed evolution to modify carbonic anhydrase with the goal of making a more stable enzyme while preserving its impressively fast activity.[34] They began by making mutations, producing roughly 27,000 different variants of the enzyme. They then meticulously challenged the variants by putting them in alkaline conditions (high pH) until they had identified enzymes that remained active. In the end, the researchers were able to engineer a new carbonic anhydrase that could work at a temperature of 107°C, while enhancing the rate of CO_2 absorption by twenty-five-fold. In addition, the enzyme was able to withstand high pH that would denature most proteins.



A model of an engineered protein that can digest PET, the main component in many plastic products. With nearly 200 million tons of plastics accumulating worldwide, this enzyme is among many new engineered proteins that holds promise as a major weapon against pollution.

More recently, protein engineers have turned their attention to tackling the problem of plastic waste. Of the 380 million tons of plastics produced annually, around 10 percent is dedicated to a material known as polyethylene terephthalate (PET), which is used to make synthetic fabrics and plastic bottles. There is a limit to how many times PET can be recycled without losing its integrity, so each year, tons of PET waste end up in landfills or find their way into the oceans. More concerning is the fact that some PET eventually turns into microplastics, endangering marine ecosystems and posing a potential threat to human health. In 2012, a group of scientists in Japan isolated an enzyme from leaf-branch compost using a large-scale DNA sequencing approach.[35] The enzyme was found to digest a number of fatty acids and was able to break down PET at a modest rate. In 2020, scientists in France took a computational approach to making changes to the PET-eating enzyme. They were ultimately able to engineer a much more stable plastic-digesting enzyme that is 90 percent efficient in breaking down PET.[36] As little as one gram of the engineered enzyme could break down over five hundred grams of PET within ten hours. Based on this new technology, a manufacturing plant for fully bio-recycled PET with the capacity to break down up to fifty thousand tons of PET annually is expected to be completed in 2025. The PET bio-recycling center will reside in France and will be the first of its kind.[37] In the future, engineered enzymes will digest the strong plastic polymer structure into smaller carbon-containing compounds that can then be extracted and used in a wide range of industrial applications.

Protein engineering is a beacon of hope in the battle against environmental pollution and climate change. As our planet faces rising temperatures, escalating greenhouse gas emissions, and a mounting tide of pollutants and plastic waste, engineered proteins offer innovative solutions for a cleaner, more sustainable future. In a few short decades, the immense advancements that have been made toward engineering new proteins stand as a testament to human ingenuity and innovation. One can only imagine what the future holds for protein engineering. It may quite literally save our planet.