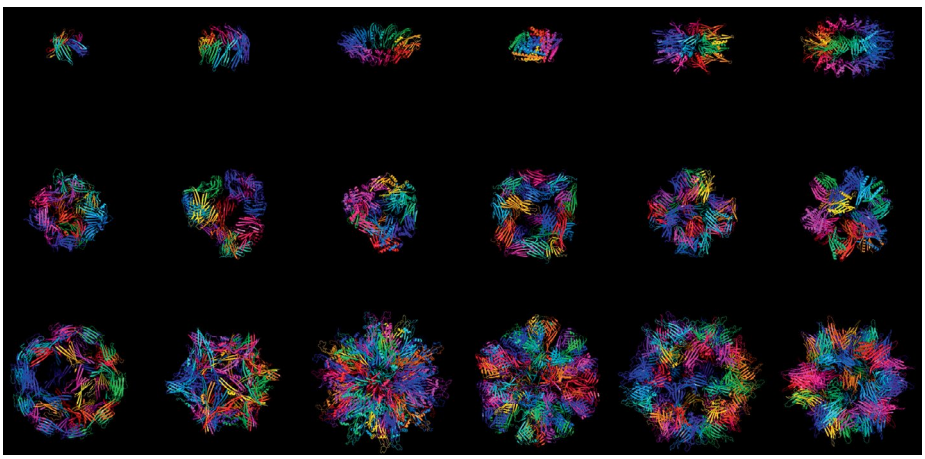
AI-enhanced protein design makes proteins that have never existed

Protein engineers are drawing on rapidly evolving machine learning tools, deep reservoirs of data, and the structure-predicting firepower of AlphaFold2 to pursue more sophisticated de novo protein designs.



AI-based algorithms can guide the design of proteins exhibiting many different kinds of symmetry, from simple spherical forms to complex icosahedral designs.

CREDIT: GENERATE BIOMEDICINES.

By Michael Eisenstein

On 26 January, Profluent came out of stealth mode with $9 million in seed funding to support the company's efforts to apply machine learning (ML) to engineer novel functional proteins. This is just the latest in a steady flurry of investment in this space. Last January, Generate Biomedicines signed a $50 million drug development deal with Amgen that could potentially net the company more than $1.9 billion in total, and a few months later, Arzeda drew $33 million in series B funding to support its ongoing protein design programs. Other startups are also starting to crowd the field, such as computational company Cradle, which exited stealth in November with a $5.5 million seed investment, and Monod Bio, which launched with $25 million in seed funding in August.

ML and other artificial intelligence (AI)-based computational tools have already proven their prowess at predicting real-world protein structures. AlphaFold 2, an algorithm developed by scientists at DeepMind that can confidently predict protein structure purely on the basis of an amino acid sequence, has become a household name since its launch in July 2021. Today, AlphaFold 2 is used routinely by many structural biologists, with over 200 million structures predicted.

This ML toolbox could generate made-to-order proteins too, including those with functions not present in nature. This is an appealing prospect because, despite natural proteins' vast molecular diversity, there are many biomedical and industrial problems that evolution has never been compelled to solve. Scientists are now rapidly moving toward a future in which they can apply careful computational analysis to infer the underlying principles governing the structure and function of real-world proteins and apply them to construct bespoke proteins with functions devised by the user. Lucas Nivon, CEO and cofounder of Cyrus Biotechnology, believes the ultimate impact of such in silico-designed proteins will be massive and compares the field to the fledgling biotech industry of the 1980s. "I think in 30 years 30, 40 or 50% of drugs will be computationally designed proteins," he says.

To date, companies operating in the protein design space have largely focused on retooling existing proteins to perform new tasks or enhance specific properties, rather than true design from scratch. For example, scientists at Generate Biomedicines have drawn on existing knowledge about the SARS-CoV-2 spike protein and its interactions with the receptor protein ACE2 to design a synthetic protein that can consistently block viral entry across diverse variants. "In our internal testing, this molecule is quite resistant to all of the variants that we've seen thus far," says cofounder and CTO Gevorg Grigoryan, adding that Generate aims to file Investigational New Drug paperwork to clear the way for clinical testing in the second quarter of this year. More ambitious programs are on the horizon, although it remains to be seen how soon the leap to de novo design — in which new proteins are built entirely from scratch — will come.

The field of AI-assisted protein design is blossoming, but the roots of the field stretch back more than two decades, with work by academic researchers like David Baker and colleagues at what is now the Institute for Protein Design at the University of Washington. Starting in the late 1990s, Baker — who has co-founded companies in this space including Cyrus, Monod and Arzeda — oversaw the development of Rosetta, a foundational software suite for predicting and manipulating protein structures. Since then, Baker and other researchers have developed many other powerful tools for protein design, powered by rapid progress in ML algorithms — and particularly, by progress in a subset of ML techniques known as deep learning. This past September, for example, Baker's team published their deep learning ProteinMPNN platform, which allows them to input the structure they want and have the algorithm spit out an amino acid sequence likely to produce that de novo backbone structure, achieving a >50% success rate.

Some of the greatest excitement in the deep learning world relates to generative models that can create entirely new proteins, never seen before in nature. These modeling tools belong to the same category of algorithms used to produce eerie and compelling AI-generated artworks in programs like Stable Diffusion or DALL-E 2 and text in programs like

News in brief

Biotech commission accused of conflicts of interest

A new group of influential industry execs and investors have been tasked by government to shape the US Department of Defense's spending in the biotech sector. Many worry that, without a conflict of interest policy, those appointed to the new commission could profit from such federal spending. The US Senate and House Armed Services Committees in December named 12 individuals to the new National Security Commission on Emerging Biotechnology. Among the notable appointees are Jason Kelly, co-founder and CEO of Ginkgo Bioworks, selected as chair; Michelle Rozo, formerly of the National Security Council, as vice chair; Eric Schmidt, formerly Google CEO and executive chairman of Alphabet; and Angela Belcher, head of biological engineering at the Massachusetts Institute of Technology. Commission members are not required to give up their industry roles or divest themselves of relevant personal investments — a particular concern for Schmidt, who holds shares in several biotech companies through the venture capital firm First Spark Ventures, where he serves as a partner and strategic advisor.

After business network CNBC raised questions regarding Schmidt potentially profiting should his portfolio companies be selected by the commission for federal investment, a spokesperson for Schmidt said in January that he would donate all net profits from his investment in First Spark to charity.

That doesn't go far enough for some. "Congress created this commission without adequate safeguards against conflicts of interest," says Walter Shaub, a senior ethics fellow at the nonprofit Project on Government Oversight and previously director of the US Office of Government Ethics. "These are individuals who are going to be helping to shape federal policy on the intersection of biotechnology and national security, and it'll be legal for them to make recommendations that benefit their own personal financial interests." A Senate Armed Services Committee spokesperson said the biotech commission members were selected by bipartisan leaders in the House and Senate and "every member on this commission is required to adhere to all government ethics policies."

ChatGPT. In those cases, the software is trained on vast amounts of annotated image data and then uses those insights to produce new pictures in response to user queries. The same feat can be achieved with protein sequences and structures, where the algorithm draws on a rich repository of real-world biological information to dream up new proteins based on the patterns and principles observed in nature. To do this, however, researchers also need to give the computer guidance on the biochemical and physical constraints that inform protein design, or else the resulting output will offer little more than artistic value.

One effective strategy to understand protein sequence and structure is to approach them as 'text', using language modeling algorithms that follow rules of biological 'grammar' and 'syntax'. "To generate a fluent sentence or a document, the algorithm needs to learn about relationships between different types of words, but it needs to also learn facts about the world to make a document that's cohesive and makes sense," says Ali Madani, a computer scientist formerly at Salesforce Research who recently founded Profluent. In a recent publication, Madani and colleagues describe a language modeling algorithm that can yield novel computer-designed proteins that can be successfully produced in the lab with catalytic activities comparable to those of natural enzymes. Language modeling is also a key part of Arzeda's toolbox, according to co-founder and CEO Alexandre Zanghellini. For one project, the company used multiple rounds of algorithmic design and optimization to engineer an enzyme with improved stability against degradation. "In three rounds of iteration, we were able to go from complete disappearance of the protein after four weeks to retention of effectively 95% activity," he says.

A recent preprint from researchers at Generate describes a new generative modeling-based design algorithm called Chroma, which includes several features that improve its performance and success rate. These include diffusion models, an approach used in many image-generation AI tools that makes it easier to manipulate complex, multidimensional data. Chroma also employs algorithmic techniques to assess long-range interactions between residues that are far apart on the protein's amino acid backbone, but that may be essential for proper folding and function. In a series of initial demonstrations, the Generate team showed that they could obtain sequences that were predicted to fold into a broad array of naturally occurring and arbitrarily chosen structures and subdomains — including the shapes of the letters of the alphabet — although it remains to be seen how many will form these folds in the lab.

In addition to the new algorithms' power, the tremendous amount of structural data captured by biologists has also allowed the protein design field to take off. The Protein Data Bank, a critical resource for protein designers, now contains more than 200,000 experimentally solved structures. The AlphaFold 2 algorithm is also proving to be a game changer here in terms of providing training material and guidance for design algorithms. "They are models, so you have to take them with a grain of salt, but now you have this extraordinarily large amount of predicted structures that you can build upon," says Zanghellini, who says this tool is a core component of Arzeda's computational design workflow.

For AI-guided design, more training data are always better. But existing gene and protein databases are constrained by a limited range of species and a heavy bias towards humans and commonly used model organisms. Basecamp Research is building an ultra-diverse repository of biological information obtained from samples collected in biomes in 17 countries, ranging from the Antarctic to the rainforest to hydrothermal vents on the ocean floor. Chief Technology Officer Philipp Lorenz says that once the genomic data from these specimens are analyzed and annotated, they can assemble a knowledge-graph that can reveal functional relationships between diverse proteins and pathways that would not be obvious purely on the basis of sequence-based analysis. "It's not just generating a new protein," says Lorenz. "We are finding protein families in prokaryotes that have been thought to exist only in eukaryotes." This means many more starting points for AI-guided protein design efforts, and Lorenz says that his team's own design experiments have achieved an 80% success rate at producing functional proteins.

But proteins do not function in a vacuum. Tess van Stekelenburg, an investor at Hummingbird Ventures, notes that Basecamp — one of the companies funded by the firm — captures all manner of environmental and biochemical context for the proteins it identifies. The resulting 'metadata' accompanying each protein sequence can help guide the engineering of proteins that express and function optimally in particular conditions. "It gives you a lot more ability to constrain for things like pH, temperature or pressure, if that's what you're planning to look at," she says.

Some companies are also looking to augment public structural biology resources with data of their own. Generate is in the process of building a multi-instrument cryo-electron microscopy facility, which will allow them to generate near-atomic-resolution structures at relatively high throughput. Such internally generated structural data are more likely to include relevant metadata about individual proteins than data from publicly available resources.

In-house wet lab facilities are another critical component of the design process because experimental results are, in turn, used to retrain the algorithm to achieve even better outcomes in future rounds. Grigoryan notes that, although Generate likes to spotlight its algorithmic toolbox, the majority of its workforce comprises experimentalists. And Bruno Correia, a computational biologist at the École Polytechnique Fédérale de Lausanne, says that the success of a protein design effort depends on close consultation between algorithm experts and experienced wet-lab practitioners. "This notion of how protein molecules are and how they behave experimentally builds in a lot of constraints," says Correia. "I think it's a mistake to handle biological entities just as a piece of data."

Biological validation is an extremely important consideration for investors in this sector, says van Stekelenburg. "If you are doing de novo, the real gold standard is not which architecture are you using – it's what percentage of your designed proteins had the end desired property," she says. "If you can't show that, then it doesn't make sense." Accordingly, most companies pursuing computational design are still focused on tuning protein function rather than overhauling it, shortening the leap between prediction and performance.

Nivon says that Cyrus typically works with existing drugs and proteins that fall short in a particular parameter. "This could be a drug that needs better efficacy, lower immunogenicity or a better toxicity profile," he says. For Cradle, the primary goal is to improve protein therapeutics by optimizing properties like stability. "We've benchmarked our model against empirical studies so that people can get a sense of how well this might work in an experimental setting," says founder and CEO Stef van Grieken.

Arzeda's focus is on enzyme engineering for industrial applications. They have already succeeded in creating proteins with novel catalytic functions for use in agriculture, materials and food science. These projects often begin with a relatively well-established core reaction that is catalyzed in nature. But to adapt these reactions to work with a different substrate, "you need to remodel the active site dramatically," says Zanghellini. Some of the company's projects include a plant enzyme that can break down a widely used herbicide, as well as enzymes that can convert relatively low-value plant byproducts into useful natural sweeteners.

Generate's first-generation engineering projects have focused on optimization. In one published study, company scientists showed that they could 'resurface' the amino acid-metabolizing enzyme L-asparaginase from *Escherichia coli* bacteria, altering the amino acid composition of its exterior to greatly reduce its immunogenicity. But with the new Chroma algorithm, Grigoryan says that Generate is ready to embark on more ambitious projects, in which the algorithm can start building true de novo designs with user-designated structural and functional features. Of course, Chroma's design proposals must then be validated by experimental testing, although Grigoryan says "we're very encouraged by what we've seen."

Zanghellini believes the field is near an inflection point. "We're starting to see the possibility of really truly creating a complex active site and then building the protein around it," he says. But he adds that many more challenges await. For example, a protein with excellent catalytic properties might be exceedingly difficult to manufacture at scale or exhibit poor properties as a drug. In the future, however, next-generation algorithms should make it possible to generate de novo proteins optimized to tick off many boxes on a scientist's wish list rather than just one.

Michael Eisenstein
Philadelphia, PA, USA

Acknowledgements
Additional reporting by Shafaq Zia.

News in brief

Big pharma craves slice of AI-based RNA drug discovery

Interest in companies with artificial intelligence (AI) platforms for RNA-targeted drug discovery has ramped up in recent months. In January Stanford University startup Atomic AI raised $35 million in a series A round and Anima Biotech entered a collaboration with AbbVie in which the biotech gains $42 million up front and potentially up to $580 million. In November Rgenta Therapeutics announced a $52 million series A round, and in September Arpeggio Biosciences finalized a $17 million series A to develop its platform to study transcriptional dysregulation in disease. At least eight companies have been launched to develop RNA-modulating small molecules, and all have signed collaborations with big players, including Skyhawk Therapeutics, partnered with Bristol Myers Squibb, Merck and Takeda; Remix Therapeutics, with Johnson & Johnson; Ribometrix, with Genentech, and Arrakis Therapeutics, with Roche.

All of these companies are exploring how to drug RNA to treat disease, as opposed to targeting proteins. Atomic AI intends to use its platform – which was developed by Raphael Townshend at Stanford and combines machine learning, wet-lab experimentation and 3D RNA structure modeling – to predict and identify parts of the transcriptome that are targetable with small molecules. Anima's platform combines phenotypic screening of live mRNA biology with AI to elucidate the mechanism of action of small-molecule mRNA drugs. Rgenta, which has a collaboration with Eli Lilly, is focused on finding oral medicines to tackle incurable diseases in oncology and neurology.