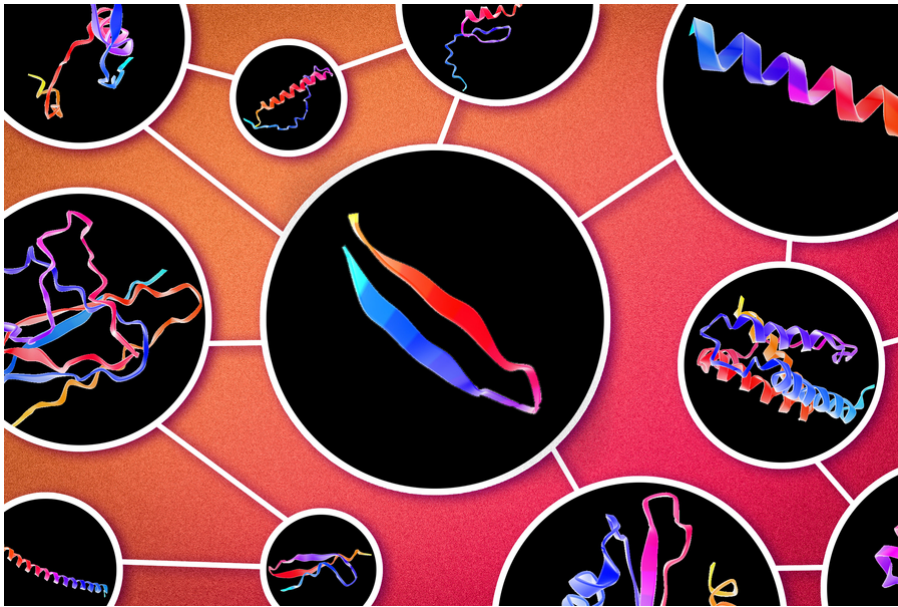


## AI system can generate novel proteins that meet structural design targets

These tunable proteins could be used to create new materials with specific mechanical properties, like toughness or flexibility.

Adam Zewe | MIT News Office

April 20, 2023



A new machine-learning system can generate protein designs with certain structural features, and which do not exist in nature. These proteins could be utilized to make materials that have similar mechanical properties to existing materials, like polymers, but which would have a much smaller carbon footprint.

Image: Jose-Luis Olivares/MIT with figures courtesy of the researchers

MIT researchers are using artificial intelligence to design new proteins that go beyond those found in nature.

They developed machine-learning algorithms that can generate proteins with specific structural features, which could be used to make materials that have certain mechanical

properties, like stiffness or elasticity. Such biologically inspired materials could potentially replace materials made from petroleum or ceramics, but with a much smaller carbon footprint.

The researchers from MIT, the MIT-IBM Watson AI Lab, and Tufts University employed a generative model, which is the same type of machine-learning model architecture used in AI systems like DALL-E 2. But instead of using it to generate realistic images from natural language prompts, like DALL-E 2 does, they adapted the model architecture so it could predict amino acid sequences of proteins that achieve specific structural objectives.

In a [paper](#) published today in *Chem*, the researchers demonstrate how these models can generate realistic, yet novel, proteins. The models, which learn biochemical relationships that control how proteins form, can produce new proteins that could enable unique applications, says senior author Markus Buehler, the Jerry McAfee Professor in Engineering and professor of civil and environmental engineering and of mechanical engineering.

For instance, this tool could be used to develop protein-inspired food coatings, which could keep produce fresh longer while being safe for humans to eat. And the models can generate millions of proteins in a few days, quickly giving scientists a portfolio of new ideas to explore, he adds.

“When you think about designing proteins nature has not discovered yet, it is such a huge design space that you can’t just sort it out with a pencil and paper. You have to figure out the language of life, the way amino acids are encoded by DNA and then come together to form protein structures. Before we had deep learning, we really couldn’t do this,” says Buehler, who is also a member of the MIT-IBM Watson AI Lab.

Joining Buehler on the paper are lead author Bo Ni, a postdoc in Buehler’s Laboratory for Atomistic and Molecular Mechanics; and David Kaplan, the Stern Family Professor of Engineering and professor of bioengineering at Tufts.

### **Adapting new tools for the task**

Proteins are formed by chains of amino acids, folded together in 3D patterns. The sequence of amino acids determines the mechanical properties of the protein. While scientists have identified thousands of proteins created through evolution, they estimate that an enormous number of amino acid sequences remain undiscovered.

To streamline protein discovery, researchers have recently developed deep learning models that can predict the 3D structure of a protein for a set of amino acid sequences. But the inverse problem — predicting a sequence of amino acid structures that meet design targets — has proven even more challenging.

A new advent in machine learning enabled Buehler and his colleagues to tackle this thorny challenge: attention-based diffusion models.

Attention-based models can learn very long-range relationships, which is key to developing proteins because one mutation in a long amino acid sequence can make or break the entire design, Buehler says. A diffusion model learns to generate new data through a process that involves adding noise to training data, then learning to recover the data by removing the noise. They are often more effective than other models at generating high-quality, realistic data that can be conditioned to meet a set of target objectives to meet a design demand.

The researchers used this architecture to build two machine-learning models that can predict a variety of new amino acid sequences which form proteins that meet structural design targets.

“In the biomedical industry, you might not want a protein that is completely unknown because then you don’t know its properties. But in some applications, you might want a brand-new protein that is similar to one found in nature, but does something different. We can generate a spectrum with these models, which we control by tuning certain knobs,” Buehler says.

Common folding patterns of amino acids, known as secondary structures, produce different mechanical properties. For instance, proteins with alpha helix structures yield stretchy materials while those with beta sheet structures yield rigid materials. Combining alpha helices and beta sheets can create materials that are stretchy and strong, like silks.

The researchers developed two models, one that operates on overall structural properties of the protein and one that operates at the amino acid level. Both models work by combining these amino acid structures to generate proteins. For the model that operates on the overall structural properties, a user inputs a desired percentage of different structures (40 percent alpha-helix and 60 percent beta sheet, for instance). Then the

model generates sequences that meet those targets. For the second model, the scientist also specifies the order of amino acid structures, which gives much finer-grained control.

The models are connected to an algorithm that predicts protein folding, which the researchers use to determine the protein's 3D structure. Then they calculate its resulting properties and check those against the design specifications.

### **Realistic yet novel designs**

They tested their models by comparing the new proteins to known proteins that have similar structural properties. Many had some overlap with existing amino acid sequences, about 50 to 60 percent in most cases, but also some entirely new sequences. The level of similarity suggests that many of the generated proteins are synthesizable, Buehler adds.

To ensure the predicted proteins are reasonable, the researchers tried to trick the models by inputting physically impossible design targets. They were impressed to see that, instead of producing improbable proteins, the models generated the closest synthesizable solution.

“The learning algorithm can pick up the hidden relationships in nature. This gives us confidence to say that whatever comes out of our model is very likely to be realistic,” Ni says.

Next, the researchers plan to experimentally validate some of the new protein designs by making them in a lab. They also want to continue augmenting and refining the models so they can develop amino acid sequences that meet more criteria, such as biological functions.


“For the applications we are interested in, like sustainability, medicine, food, health, and materials design, we are going to need to go beyond what nature has done. Here is a new design tool that we can use to create potential solutions that might help us solve some of the really pressing societal issues we are facing,” Buehler says.

“In addition to their natural role in living cells, proteins are increasingly playing a key role in technological applications ranging from biologic drugs to functional materials. In this context, a key challenge is to design protein sequences with desired properties suitable for specific applications. Generative machine-learning approaches, including ones leveraging diffusion models, have recently emerged as powerful tools in this space,” says Tuomas

Knowles, professor of physical chemistry and biophysics at Cambridge University, who was not involved with this research. “Buehler and colleagues demonstrate a crucial advance in this area by providing a design approach which allows the secondary structure of the designed protein to be tailored. This is an exciting advance with implications for many potential areas, including for designing building blocks for functional materials, the properties of which are governed by secondary structure elements.”

“This particular work is fascinating because it is examining the creation of new proteins that mostly do not exist, but then it examines what their characteristics would be from a mechanics-based direction,” adds Philip LeDuc, the William J. Brown Professor of Mechanical Engineering at Carnegie Mellon University, who was also not involved with this work. “I personally have been fascinated by the idea of creating molecules that do not exist that have functionality that we haven’t even imagined yet. This is a tremendous step in that direction.”

This research was supported, in part, by the MIT-IBM Watson AI Lab, the U.S. Department of Agriculture, the U.S. Department of Energy, the Army Research Office, the National Institutes of Health, and the Office of Naval Research.

 [Paper: “Generative design of de novo proteins based on secondary structure constraints using an attention-based diffusion model”](#)