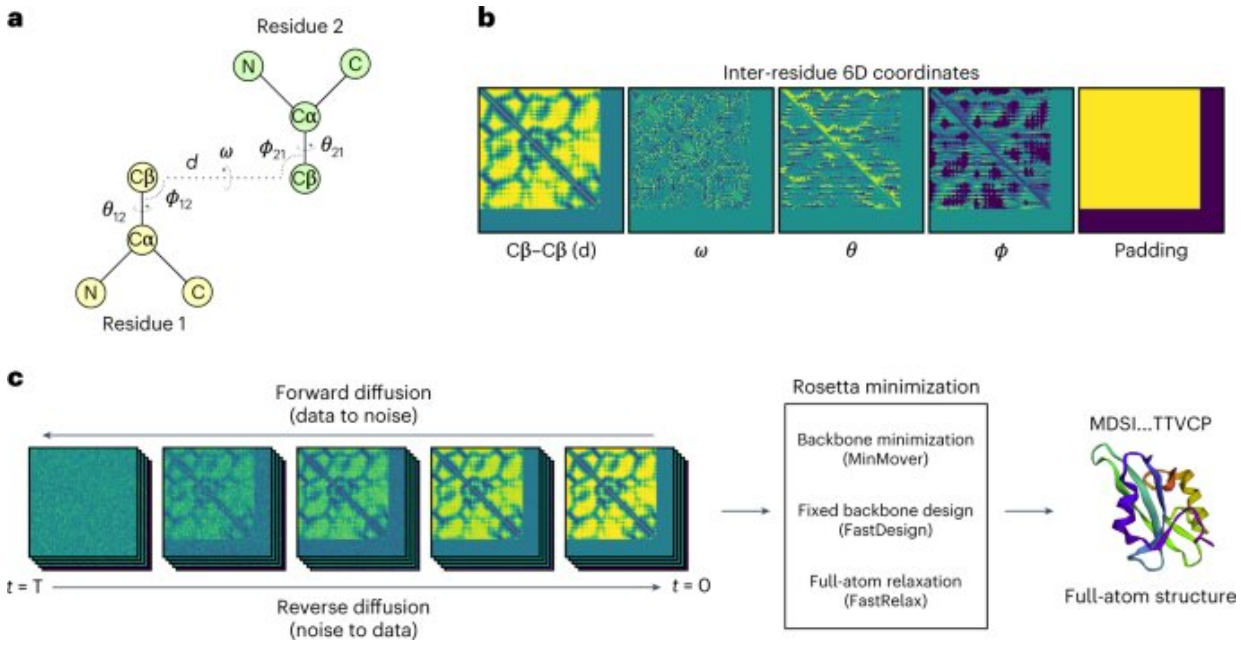
Researchers use generative AI to design novel proteins

May 4 2023, by Jim Oldfield



Model overview. Credit: *Nature Computational Science* (2023). DOI: 10.1038/s43588-023-00440-3

Researchers at the University of Toronto have developed an artificial intelligence system that can create proteins not found in nature using generative diffusion, the same technology behind popular image-creation platforms such as DALL-E and Midjourney.

The system will help advance the field of generative biology, which

promises to speed drug development by making the design and testing of entirely new therapeutic proteins more efficient and flexible.

"Our model learns from image representations to generate fully new proteins, at a very high rate," says Philip M. Kim, a professor in the Donnelly Centre for Cellular and Biomolecular Research at U of T's Temerty Faculty of Medicine. "All our proteins appear to be biophysically real, meaning they fold into configurations that enable them to carry out specific functions within cells."

Today, the journal *Nature Computational Science* published the findings, the first of their kind in a peer-reviewed journal. Kim's lab also published a pre-print on the model last summer through the open-access server *bioRxiv*, ahead of two similar pre-prints from last December, RF Diffusion by the University of Washington and Chroma by Generate Biomedicines.

Proteins are made from chains of amino acids that fold into three-dimensional shapes, which in turn dictate protein function. Those shapes evolved over billions of years and are varied and complex, but also limited in number. With a better understanding of how existing proteins fold, researchers have begun to design folding patterns not produced in nature.

But a major challenge, says Kim, has been to imagine folds that are both possible and functional. "It's been very hard to predict which folds will be real and work in a protein structure," says Kim, who is also a professor in the departments of molecular genetics and computer science at U of T. "By combining biophysics-based representations of protein structure with diffusion methods from the image generation space, we can begin to address this problem."

The new system, which the researchers call ProteinSGM, draws from a

large set of image-like representations of existing proteins that encode their structure accurately. The researchers feed these images into a generative diffusion model, which gradually adds noise until each image becomes all noise. The model tracks how the images become noisier and then runs the process in reverse, learning how to transform random pixels into clear images that correspond to fully novel proteins.

Jin Sub (Michael) Lee, a doctoral student in the Kim lab and first author on the paper, says that optimizing the early stage of this image generation process was one of the biggest challenges in creating ProteinSGM. "A key idea was the proper image-like representation of protein structure, such that the diffusion model can learn how to generate novel proteins accurately," says Lee, who is from Vancouver but did his undergraduate degree in South Korea and master's in Switzerland before choosing U of T for his doctorate.

Also difficult was validation of the proteins produced by ProteinSGM. The system generates many structures, often unlike anything found in nature. Almost all of them look real according to standard metrics, says Lee, but the researchers needed further proof.

To test their new proteins, Lee and his colleagues first turned to OmegaFold, an improved version of DeepMind's software AlphaFold 2. Both platforms use AI to predict the structure of proteins based on amino acid sequences.

With OmegaFold, the team confirmed that almost all their novel sequences fold into the desired and also novel protein structures. They then chose a smaller number to create physically in test tubes, to confirm the structures were proteins and not just stray strings of chemical compounds.

"With matches in OmegaFold and experimental testing in the lab, we

could be confident these were properly folded proteins. It was amazing to see validation of these fully new protein folds that don't exist anywhere in nature," Lee says.

Next steps based on this work include further development of ProteinSGM for antibodies and other proteins with the most therapeutic potential, Kim says. "This will be a very exciting area for research and entrepreneurship," he adds.

Lee says he would like to see generative biology move toward joint design of protein sequences and structures, including protein side-chain conformations. Most research to date has focussed on generation of backbones, the primary chemical structures that hold proteins together.

"Side-chain configurations ultimately determine protein function, and although designing them means an exponential increase in complexity, it may be possible with proper engineering," Lee says. "We hope to find out."

More information: Philip Kim, Score-based generative modeling for de novo protein design, *Nature Computational Science* (2023). DOI: 10.1038/s43588-023-00440-3.
www.nature.com/articles/s43588-023-00440-3

Provided by University of Toronto