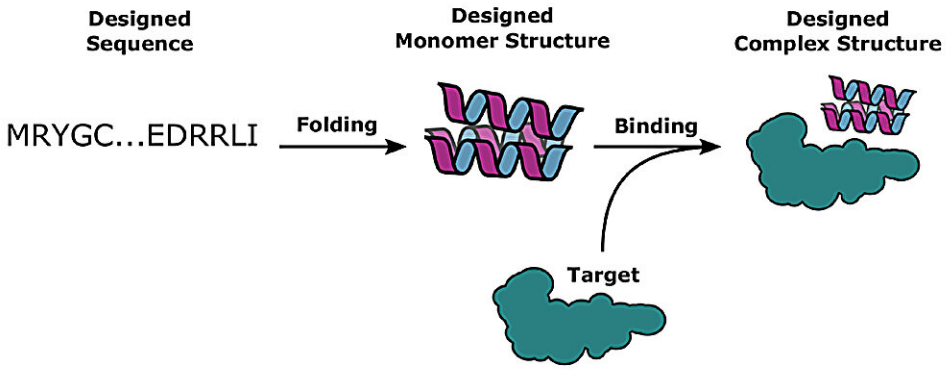
Deep learning for new protein design

August 3 2023, by Jorge Salazar



Deep learning methods have been used to augment existing energy-based physical models in 'do novo' or from-scratch computational protein design, resulting in a 10-fold increase in success rates verified in the lab for binding a designed protein with its target protein. The results will help scientists design better drugs against diseases like cancer and COVID-19. Credit: DOI: 10.1038/s41467-023-38328-5

The key to understanding proteins—such as those that govern cancer, COVID-19, and other diseases—is quite simple: Identify their chemical structure and find which other proteins can bind to them. But there's a catch.

"The search space for proteins is enormous," said Brian Coventry, a research scientist with the Institute for Protein Design, University of Washington and The Howard Hughes Medical Institute.

A protein studied by his lab typically is made of 65 amino acids, and with 20 different amino acid choices at each position, there are 65 to the 20th power binding combinations, a number bigger than the estimated number of atoms there are in the universe.

Coventry is the co-author of a study published May 2023 in the journal *Nature Communications*.

In it, his team used deep learning methods to augment existing energy-based physical models in "de novo" (from scratch) computational protein design, resulting in a 10-fold increase in success rates verified in the lab for binding a designed protein with its target protein.

"We showed that you can have a significantly improved pipeline by incorporating deep learning methods to evaluate the quality of the interfaces where hydrogen bonds form or from hydrophobic interactions ," said study co-author Nathaniel Bennett, a post-doctoral scholar at the Institute for Protein Design, University of Washington.

"This is as opposed to trying to exactly enumerate all of these energies by themselves," he added.

Readers might be familiar with popular examples of deep learning applications such as the language model ChatGPT or the image generator DALL-E.

Deep learning uses computer algorithms to analyze and draw inferences from patterns in data, layering the algorithms to progressively extract higher-level features from the raw input. In the study, deep learning methods were used to learn iterative transformations of representation of the protein sequence and possible structure that very rapidly converge on models that turn out to be very accurate.

The deep learning-augmented de novo protein binder design protocol developed by the authors included the machine learning software tools AlphaFold 2 and also RoseTTA fold, which was developed by the Institute for Protein Design.

The study problem was well-suited for parallelization on Frontera because the protein design trajectories are all independent of one another, meaning that information didn't need to pass between design trajectories as the compute jobs were running.

"We just split up this problem, which has 2 to 6 million designs in it, and run all of those in parallel on the massive computing resources of Frontera. It has a large amount of CPU nodes on it. And we assigned each of these CPUS to do one of these design trajectories, which let us complete an extremely large number of design trajectories in a feasible time," said Bennett.

The authors used the RifDock docking program to generate six million protein "docks," or interactions between potentially bound protein structures, split them into chunks of about 100,000, and assign each chunk to one of Frontera's 8000+ compute nodes using Linux utilities.

Each of those 100,000 docks would be split into 100 jobs of a thousand proteins each. A thousand proteins go into the computational design software Rosetta, where the 1,000 are first screened at the tenth of the second scale, and the ones that survive are screened at the few-minutes scale.

What's more, the authors used the software tool ProteinMPNN developed by the Institute for Protein Design to further increase the computational efficiency of generating protein sequences neural networks to over 200 times faster than the previous best software.

The data used in their modeling is yeast surface display binding data, all publicly available and collected by the Institute for Protein Design. In it, tens of thousands of different strands of DNA were ordered to encode a different protein, which the scientists designed.

The DNA was then combined with yeast such that each yeast cell expresses one of the designed proteins on its surface. The yeast cells were then sorted into cells that bind and cells that don't. In turn, they used tools from the human genome sequencing project to figure out which DNA worked and which DNA didn't work.

Despite the study results that showed a 10-fold increase in the success rate for designed structures to bind on their target protein, there is still a long way to go, according to Coventry.

"We went up an order of magnitude, but we still have three more to go. The future of the research is to increase that success rate even more, and move on to a new class of even harder targets," he said. Viruses and cancer T-cell receptors are prime examples.

The ways to improve the computationally designed proteins are to make the software tools even more optimized, or to sample more.

Said Coventry, "The bigger the computer we can find, the better the proteins we can make. We are building the tools to make the cancer-fighting drugs of tomorrow. Many of the individual binders that we make could go on to become the drugs that save people's lives. We are making the process to make those drugs better."

More information: Nathaniel R. Bennett et al, Improving de novo protein binder design with deep learning, *Nature Communications* (2023). DOI: 10.1038/s41467-023-38328-5

Provided by University of Texas at Austin