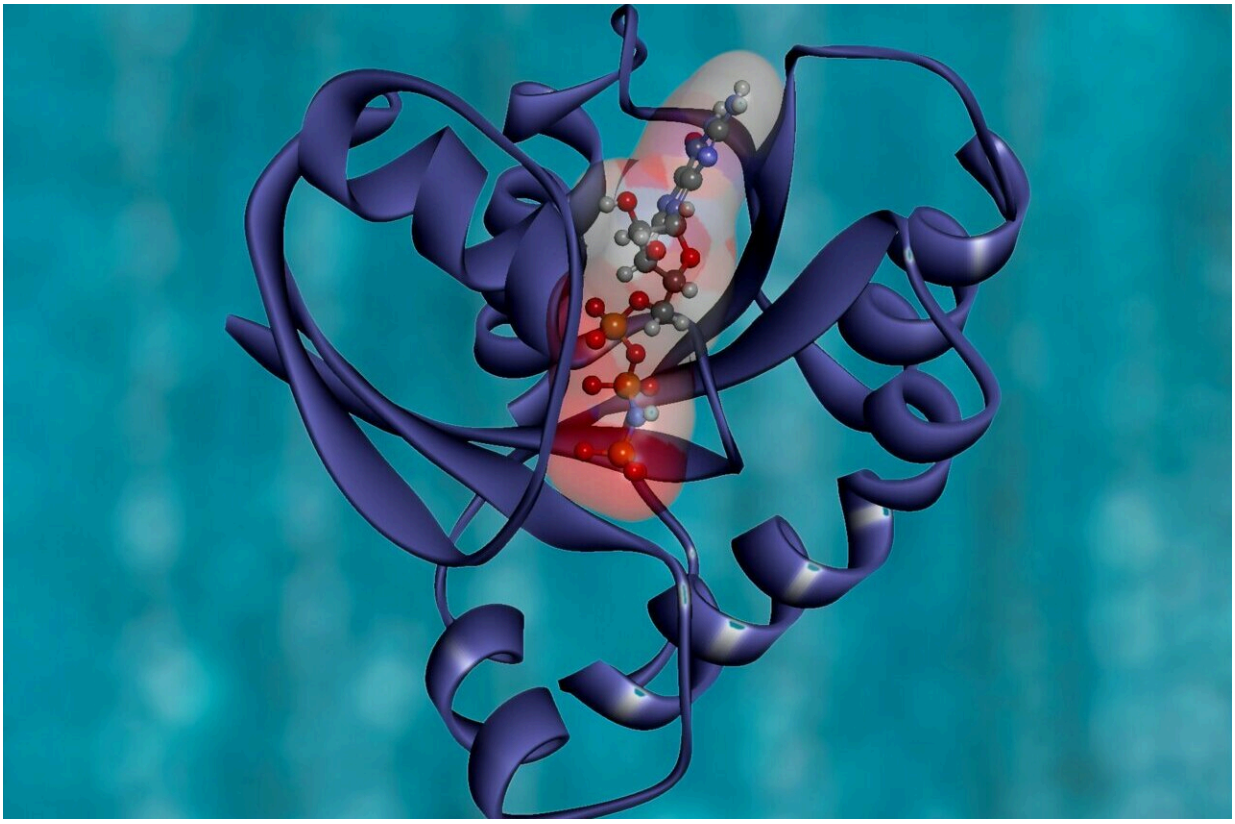
AI technology generates original proteins from scratch

January 26 2023



Credit: Unsplash/CC0 Public Domain

Scientists have created an AI system capable of generating artificial enzymes from scratch. In laboratory tests, some of these enzymes worked as well as those found in nature, even when their artificially

generated amino acid sequences diverged significantly from any known natural protein.

The experiment demonstrates that natural language processing, although it was developed to read and write language text, can learn at least some of the underlying principles of biology. Salesforce Research developed the AI program, called ProGen, which uses next-token prediction to assemble amino acid sequences into artificial proteins.

Scientists said the new technology could become more powerful than directed evolution, the Nobel-prize winning protein design technology, and it will energize the 50-year-old field of protein engineering by speeding the development of new proteins that can be used for almost anything from therapeutics to degrading plastic.

"The artificial designs perform much better than designs that were inspired by the evolutionary process," said James Fraser, Ph.D., professor of bioengineering and therapeutic sciences at the UCSF School of Pharmacy, and an author of the work, which was published Jan. 26, in *Nature Biotechnology*. A previous version of the paper has been available on the preprint server BiorXiv since July of 2021, where it garnered several dozen citations before being published in a peer-reviewed journal.

"The language model is learning aspects of evolution, but it's different than the normal evolutionary process," Fraser said. "We now have the ability to tune the generation of these properties for specific effects. For example, an enzyme that's incredibly thermostable or likes acidic environments or won't interact with other proteins."

To create the model, scientists simply fed the amino acid sequences of 280 million different proteins of all kinds into the machine learning model and let it digest the information for a couple of weeks. Then, they

fine-tuned the model by priming it with 56,000 sequences from five lysozyme families, along with some contextual information about these proteins.

The model quickly generated a million sequences, and the research team selected 100 to test, based on how closely they resembled the sequences of natural proteins, as well how naturalistic the AI proteins' underlying amino acid "grammar" and "semantics" were.

Out of this first batch of a 100 proteins, which were screened in vitro by Tierra Biosciences, the team made five artificial proteins to test in cells and compared their activity to an enzyme found in the whites of chicken eggs, known as hen egg white lysozyme (HEWL). Similar lysozymes are found in human tears, saliva and milk, where they defend against bacteria and fungi.

Two of the artificial enzymes were able to break down the cell walls of bacteria with activity comparable to HEWL, yet their sequences were only about 18% identical to one another. The two sequences were about 90% and 70% identical to any known protein.

Just one mutation in a natural protein can make it stop working, but in a different round of screening, the team found that the AI-generated enzymes showed activity even when as little as 31.4% of their sequence resembled any known natural protein.

The AI was even able to learn how the enzymes should be shaped, simply from studying the raw sequence data. Measured with X-ray crystallography, the atomic structures of the artificial proteins looked just as they should, although the sequences were like nothing seen before.

Salesforce Research developed ProGen in 2020, based on a kind of

natural language programming their researchers originally developed to generate English language text.

They knew from their previous work that the AI system could teach itself grammar and the meaning of words, along with other underlying rules that make writing well-composed.

"When you train sequence-based models with lots of data, they are really powerful in learning structure and rules," said Nikhil Naik, Ph.D., Director of AI Research at Salesforce Research, and the senior author of the paper. "They learn what words can co-occur, and also compositionality."

With proteins, the design choices were almost limitless. Lysozymes are small as proteins go, with up to about 300 amino acids. But with 20 possible amino acids, there are an enormous number (20300) of possible combinations. That's greater than taking all the humans who lived throughout time, multiplied by the number of grains of sand on Earth, multiplied by the number of atoms in the universe.

Given the limitless possibilities, it's remarkable that the model can so easily generate working enzymes.

"The capability to generate functional proteins from scratch out-of-the-box demonstrates we are entering into a new era of protein design," said Ali Madani, Ph.D., founder of Profluent Bio, former research scientist at Salesforce Research, and the paper's first author. "This is a versatile new tool available to protein engineers, and we're looking forward to seeing the therapeutic applications."

A comprehensive codebase for the methods described in the paper is publicly available at github.com/salesforce/progen .

More information: Ali Madani, Large language models generate functional protein sequences across diverse families, *Nature Biotechnology* (2023). DOI: 10.1038/s41587-022-01618-2.
www.nature.com/articles/s41587-022-01618-2

Provided by University of California, San Francisco