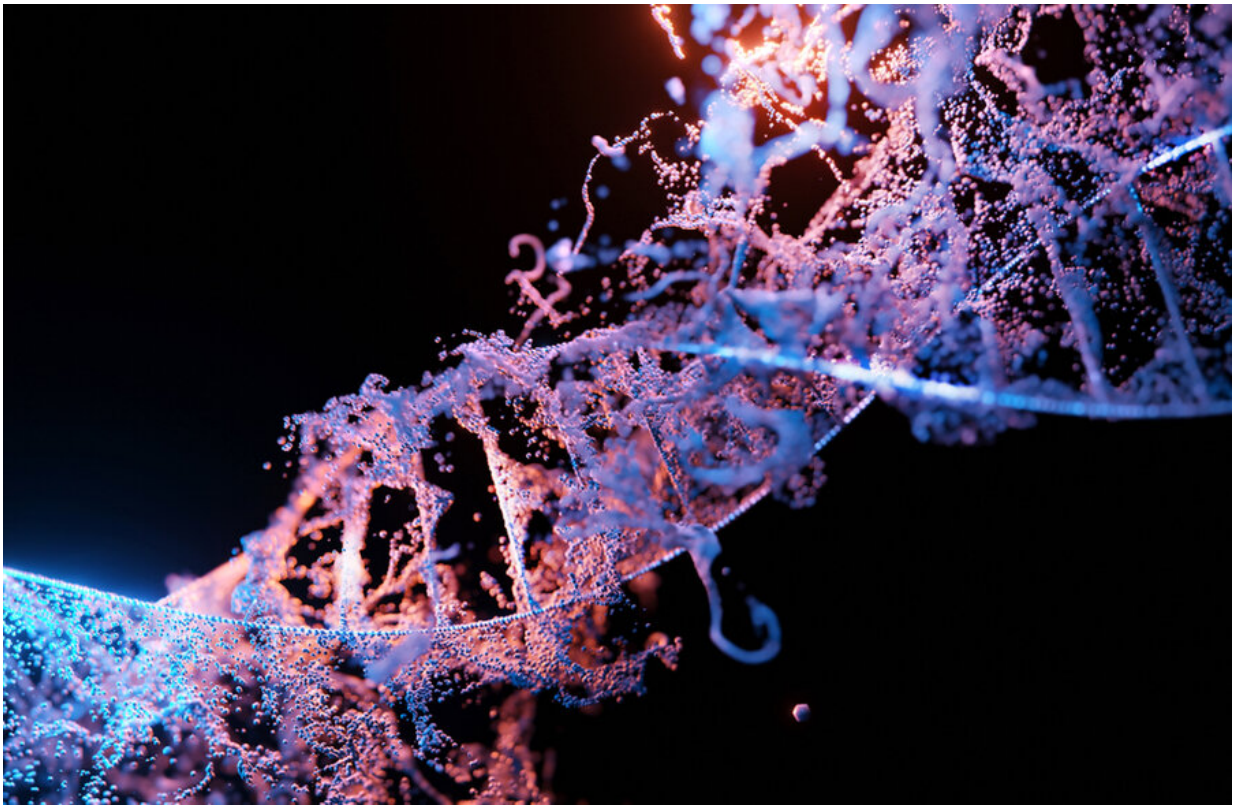
Scientists build a system that can generate AI models for biology research

July 7 2023, by Alex Ouyang



Graphical abstract. Credit: Massachusetts Institute of Technology

Is it possible to build machine-learning models without machine-learning expertise?

Jim Collins, the Termeer Professor of Medical Engineering and Science

in the Department of Biological Engineering at MIT and the life sciences faculty lead at the Abdul Latif Jameel Clinic for Machine Learning in Health (Jameel Clinic), along with a number of colleagues decided to tackle this problem when facing a similar conundrum. An open-access paper on their proposed solution, called BioAutoMATED, was published in *Cell Systems*.

Recruiting machine-learning researchers can be a time-consuming and financially costly process for science and engineering labs. Even with a machine-learning expert, selecting the appropriate model, formatting the dataset for the model, then fine-tuning it can dramatically change how the model performs, and takes a lot of work.

"In your machine-learning project, how much time will you typically spend on data preparation and transformation?" asks a 2022 Google course on the Foundations of Machine Learning (ML). The two choices offered are either "Less than half the project time" or "More than half the project time." If you guessed the latter, you would be correct. Google states that it takes over 80% of project time to format the data, and that's not even taking into account the time needed to frame the problem in machine-learning terms.

"It would take many weeks of effort to figure out the appropriate model for our dataset, and this is a really prohibitive step for a lot of folks that want to use machine learning or biology," says Jacqueline Valeri, a fifth-year Ph.D. student of biological engineering in Collins's lab who is first co-author of the paper.

BioAutoMATED is an automated machine-learning system that can select and build an appropriate model for a given dataset and even take care of the laborious task of data preprocessing, whittling down a months-long process to just a few hours. Automated machine-learning (AutoML) systems are still in a relatively nascent stage of development, with

current usage primarily focused on image and text recognition, but largely unused in subfields of biology, points out first co-author and Jameel Clinic postdoc Luis Soenksen Ph.D.

"The fundamental language of biology is based on sequences," explains Soenksen, who earned his doctorate in the MIT Department of Mechanical Engineering. "Biological sequences such as DNA, RNA, proteins, and glycans have the amazing informational property of being intrinsically standardized, like an alphabet. A lot of AutoML tools are developed for text, so it made sense to extend it to [biological] sequences."

Moreover, most AutoML tools can only explore and build reduced types of models. "But you can't really know from the start of a project which model will be best for your dataset," Valeri says. "By incorporating multiple tools under one umbrella tool, we really allow a much larger search space than any individual AutoML tool could achieve on its own."

BioAutoMATED's repertoire of supervised ML models includes three types: binary classification models (dividing data into two classes), multi-class classification models (dividing data into multiple classes), and regression models (fitting continuous numerical values or measuring the strength of key relationships between variables). BioAutoMATED is even able to help determine how much data is required to appropriately train the chosen model.

"Our tool explores models that are better-suited for smaller, sparser biological datasets as well as more complex neural networks," Valeri says. This is an advantage for research groups with new data that may or may not be suited for a machine learning problem.

"Conducting novel and successful experiments at the intersection of biology and machine learning can cost a lot of money," Soenksen

explains. "Currently, biology-centric labs need to invest in significant digital infrastructure and AI-ML trained human resources before they can even see if their ideas are poised to pan out. We want to lower these barriers for domain experts in biology."

With BioAutoMATED, researchers have the freedom to run initial experiments to assess if it's worthwhile to hire a machine-learning expert to build a different model for further experimentation.

The open-source code is publicly available and, researchers emphasize, it is easy to run. "What we would love to see is for people to take our code, improve it, and collaborate with larger communities to make it a tool for all," Soenksen says. "We want to prime the biological research community and generate awareness related to AutoML techniques, as a seriously useful pathway that could merge rigorous biological practice with fast-paced AI-ML practice better than it is achieved today."

More information: Jacqueline A. Valeri et al, BioAutoMATED: An end-to-end automated machine learning tool for explanation and design of biological sequences, *Cell Systems* (2023). DOI: 10.1016/j.cels.2023.05.007

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology