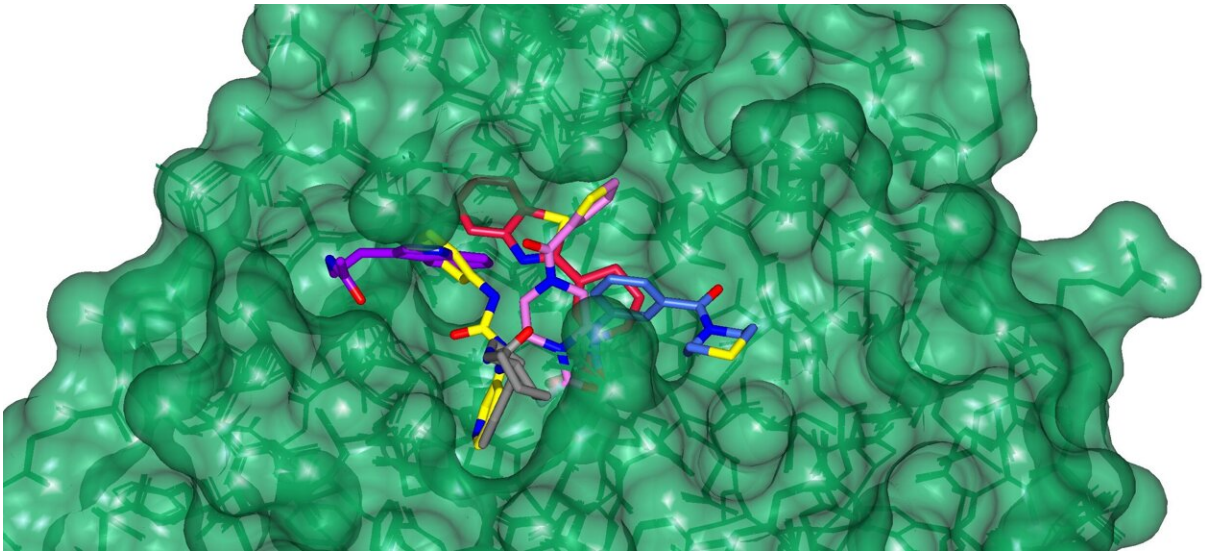
The potential of generative AI to accelerate antiviral development and drug discovery

June 28 2023



Surface representation of SARS CoV-2 Mpro protein with fragment hits from Diamond XChem platform bound in active site (Green). Credit: Diamond Light Source Ltd

In a new study, researchers from IBM, Oxford University and Diamond Light Source show that IBM's AI Model, MoLFormer, can generate antiviral molecules for multiple target virus proteins, including SARS-CoV-2, that can accelerate the drug discovery process and bolster our response to future pandemics.

The results are laid out in a new paper published in *Science Advances*,

and at the time of the paper's submission, the antiviral properties of eleven molecules were successfully validated by Oxford researchers. This breakthrough has the potential to get drugs to people faster in the next crisis and bring treatments for urgent, life-threatening illnesses within reach.

Early in the pandemic, a group of computer scientists at IBM wanted to explore if generative AI could be used to design never-before seen molecules to block SARS-CoV-2, the virus that causes COVID-19. David Stuart, Head of the Division of Structural Biology in the Department of Clinical Medicine at the University of Oxford and Life Sciences Director at Diamond Light Source, the UK's national synchrotron who is an authority on pathogens HIV, SARS, and Ebola, among other viruses explains he was initially skeptical. "The idea that you could take a protein sequence and, with AI, pluck out of thin air chemicals that would bind to a 3D site on the virus seemed very unlikely," he said.

However, he and Martin Walsh also an expert structural biologist and Life Sciences Deputy director at Diamond joined up with the IBM team and over the course of three years, demonstrated that generative AI could, "pluck viable starting points for antivirals out of thin air," in collaboration with Enamine Ltd., a chemical supplier in Ukraine, and other researchers at Oxford.

Because the generative model was also a foundation model, pre-trained on massive amounts of raw data, it was versatile enough to create new inhibitors for multiple protein targets without extra training or any knowledge of its 3D structure.

The Stuart and Walsh groups had commenced working on two essential SARS-CoV-2 proteins, namely the spike protein and the main protease. Using these targets, the team hit on four potential COVID-19 antivirals

in a fraction of the time it would have taken using conventional methods. The work then exploited Diamond's high-throughput macromolecular crystallography beamlines to visualize how a subset of the AI generated compounds bound to the main protease.

Their work is showcased in their new paper in *Science Advances* and IBM has released a web-based interface for interacting with the model and chemical foundation models like it in IBM Cloud.

The team stated that the validated molecules have many more hurdles to clear, including clinical trials, before companies could potentially turn them into drugs. But even if the AI-generated "hits" never materialize into actual drugs, the work provides confirmation that generative AI has an important role to play in the future of drug development, especially in a time of crisis.

"It took time to develop and validate these methods, but now that we have a working pipeline in place, we can generate results much faster," said study co-senior author, Payel Das, a researcher at IBM Research. "When the next virus emerges, generative AI could be pivotal in the search for new treatments."

"Generating initial compounds that bind with high affinity to a drug target of interest accelerates the structure-based drug discovery pipeline and underpins our efforts to be better prepared for future pandemics," said, Martin Walsh, who was co-senior author at Diamond

The researchers built their model, Controlled Generation of Molecules (or CogMol), on a generative AI architecture known as variational autoencoders, or VAEs. VAEs encode raw data into a compressed representation, and then decode, or translate, it back into a statistical variation on the original sample. Their model was trained on a large dataset of molecules represented as strings of text, along with general

information about proteins and their binding properties. But they deliberately left out information about SARS-CoV-2's 3D structure or molecules known to bind to it. Their goal was to give their generative foundation model a broad base of knowledge so that it could be more easily deployed for molecular design tasks it has never seen before.

Their goal was to find drug-like molecules that would bind with two COVID protein targets: the spike, which transmits the virus to the host cell, and the main protease, which helps to spread it. Though the 3D structures of both proteins had been discovered by that time, the IBM researchers chose to use only their amino acid sequences, derived from their DNA. By limiting themselves in this way, they hoped that the model could learn to generate molecules without knowing the shape of their target.

The researchers input only the amino acid sequence for each protein target into CogMol, which generated 875,000 candidate molecules in three days. To narrow the pool, the researchers ran the candidates through a retrosynthesis platform, IBM RXN for Chemistry, to understand what ingredients would be needed to synthesize the compounds. Based on the platform's predicted recipes, they selected 100 molecules for each target. Chemists at Enamine further pared the list to four molecules for each target, selecting those deemed easiest to manufacture.

After synthesizing the eight novel molecules, Enamine shipped them to Oxford for testing their ability to disrupt the functions of the two protein targets in the labs of Prof Chris Schofield and PRof Gavin Screaton. . The intense X-ray beam generated from Diamond which are 10 billion times brighter than the sun were used to visualize how the compounds interacted with proteins to inactivate their function. The novel compounds were further tested in target inhibition and live virus neutralization tests. Two of the validated antivirals target the main

protease; the other two not only targeted the spike protein but proved capable of neutralizing all six major COVID variants. "You get a map that shows exactly where things bind, and bang! you've got a confirmation," said Stuart.

Diamond Light Source, UK's national synchrotron aerial View. Credit: Diamond Light Source Ltd 2021

CogMol is one of several chemical foundation models that IBM has since developed. The largest, MoLFormer-XL, was trained on a database of more than 1.1 billion molecules and is currently being used by Moderna to design mRNA medicines. "We created valid starting points for accelerated development of antivirals using a generative foundation

model that knew relatively little about its protein targets," said the study's co-senior author, Jason Crain, a researcher at IBM Research and professor at Oxford. "I'm hopeful that these methods will allow us to create antivirals and other urgently needed compounds much faster and more inexpensively in the future."

Though the researchers focused on validating antivirals for COVID, they argue that these methods can be extended to existing viruses that continue to mutate, like the flu, or viruses that have yet to surface. "If you want to be prepared for the next pandemic, you want drugs that act on different sites of the protein," concluded Stuart. "It becomes much harder for the virus to escape."

More information: Vijil Chenthamarakshan et al, Accelerating drug target inhibitor discovery with a deep generative foundation model, *Science Advances* (2023). DOI: 10.1126/sciadv.adg7865

Provided by Diamond Light Source