NVIDIA Expands Large Language Models to Biology

Leading pharma companies, biotech startups and pioneering biology researchers are developing AI applications with the NVIDIA BioNeMo LLM service and framework to generate, predict and understand biomolecular data.

Author: Abraham Stern

As scientists probe for new insights about DNA, proteins and other building blocks of life, the NVIDIA BioNeMo framework — announced today at NVIDIA GTC — will accelerate their research.

NVIDIA BioNeMo is a framework for training and deploying large biomolecular language models at supercomputing scale — helping scientists better understand disease and find therapies for patients. The large language model (LLM) framework will support chemistry, protein, DNA and RNA data formats.

It’s part of the NVIDIA Clara Discovery collection of frameworks, applications and AI models for drug discovery.

Just as AI is learning to understand human languages with LLMs, it’s also learning the languages of biology and chemistry. By making it easier to train massive neural networks on biomolecular data, NVIDIA BioNeMo helps researchers discover new patterns and insights in biological sequences — insights that researchers can connect to biological properties or functions, and even human health conditions.

NVIDIA BioNeMo provides a framework for scientists to train large-scale language models using bigger datasets, resulting in better-performing neural networks. The framework will be available in early access on NVIDIA NGC , a hub for GPU-optimized software.

In addition to the language model framework, NVIDIA BioNeMo has a cloud API service that will support a growing list of pretrained AI models .

Scientists using natural language processing models for biological data today often train relatively small neural networks that require custom preprocessing. By adopting BioNeMo, they can scale up to LLMs with billions of parameters that capture information about molecular structure, protein solubility and more.

BioNeMo is an extension of the NVIDIA NeMo Megatron framework for GPU-accelerated training of large-scale, self-supervised language models. It’s domain specific, designed to support molecular data represented in the SMILES notation for chemical structures, and in FASTA sequence strings for amino acids and nucleic acids.

“The framework allows researchers across the healthcare and life sciences industry to take advantage of their rapidly growing biological and chemical datasets,” said Mohammed AlQuraishi, founding member of the OpenFold Consortium and assistant professor at Columbia University’s Department of Systems Biology. “This makes it easier to discover and design therapeutics that precisely target the molecular signature of a disease.”

For developers looking to quickly get started with LLMs for digital biology and chemistry applications, the NVIDIA BioNeMo LLM service will include four pretrained language models. These are optimized for inference and will be available under early access through a cloud API running on NVIDIA DGX Foundry .

ESM-1 : This protein LLM, based on the state-of-the-art ESM-1b model published by Meta AI, processes amino acid sequences to generate representations that can be used to predict a wide variety of protein properties and functions. It also improves scientists’ ability to understand protein structure.

OpenFold : The public-private consortium creating state-of-the-art protein modeling tools will make its open-source AI pipeline accessible through the BioNeMo service.

MegaMolBART : Trained on 1.4 billion molecules, this generative chemistry model can be used for reaction prediction, molecular optimization and de novo molecular generation.

ProtT5 : The model, developed in a collaboration led by the Technical University of Munich's RostLab and including NVIDIA, extends the capabilities of protein LLMs like Meta AI's ESM-1b to sequence generation.

In the future, researchers using the BioNeMo LLM service will be able to customize the LLM models for higher accuracy on their applications in a few hours — with fine-tuning and new techniques such as p-tuning, a training method that requires a dataset with just a few hundred examples instead of millions.

A wave of experts in biotech and pharma are adopting NVIDIA BioNeMo to support drug discovery research.

AstraZeneca and NVIDIA have used the Cambridge-1 supercomputer to develop the MegaMolBART model included in the BioNeMo LLM service. The global biopharmaceuticals company will use the BioNeMo framework to help train some of the world's largest language models on datasets of small molecules, proteins and, soon, DNA.

Researchers at the Broad Institute of MIT and Harvard are working with NVIDIA to develop next-generation DNA language models using the BioNeMo framework. These models will be integrated into Terra, a cloud platform co-developed by the Broad Institute, Microsoft and Verily that enables biomedical researchers to share, access and analyze data securely and at scale. The AI models will also be added to the BioNeMo service's collection.

The OpenFold consortium plans to use the BioNeMo framework to advance its work developing AI models that can predict molecular structures from amino acid sequences with near-experimental accuracy.

Peptone is focused on modeling intrinsically disordered proteins — proteins that lack a stable 3D structure. The company is working with NVIDIA to develop versions of the ESM model using the NeMo framework, which BioNeMo is also based on. The project, which is scheduled to run on NVIDIA's Cambridge-1 supercomputer, will advance Peptone's drug discovery work.

Evozyne , a Chicago-based biotechnology company, combines engineering and deep learning technology to design novel proteins to solve long-standing challenges in therapeutics and sustainability.

"The BioNeMo framework is an enabling technology to efficiently leverage the power of LLMs for data-driven protein design within our design-build-test cycle," said Andrew Ferguson, co-founder and head of computation at Evozyne. "This will have an immediate impact on our design of novel functional proteins, with applications in human health and sustainability."

"As we see the ever-widening adoption of large language models in the protein space, being able to efficiently train LLMs and quickly modulate model architectures is becoming hugely important," said Istvan Redl, machine learning lead at Peptone, a biotech startup in the NVIDIA Inception program. "We believe that these two engineering aspects — scalability and rapid experimentation — are exactly what the BioNeMo framework could provide."

Sign up for early access to the NVIDIA BioNeMo LLM service or BioNeMo framework. For hands on-experience with the MegaMolBART chemistry model in BioNeMo, request a free lab from NVIDIA LaunchPad on training and deploying LLMs.

Discover the latest in AI and healthcare at GTC , running online through Thursday, Sept. 22. Registration is free.

Watch the GTC keynote address by NVIDIA founder and CEO Jensen Huang below:

Main image by Mahendra awale, licensed under CC BY-SA 3.0 via Wikimedia Commons

Original URL:
https://blogs.nvidia.com/blog/2022/09/20/bionemo-large-language-models-drug-discovery/