IDC PERSPECTIVE

GPU-Powered Transformer Models Poised to Accelerate Drug Discovery and Disrupt Drug Development

Nimita Limaye

EXECUTIVE SNAPSHOT

FIGURE 1

Executive Snapshot: GPU-Powered Transformer Models Poised to Accelerate Drug Discovery and Disrupt Drug Development

This IDC Perspective provides analysts insights on the importance of GPU-powered transformer models in accelerating drug discovery and development. It analyzes industry challenges and reviews use cases, including the value of these transformer models in fueling innovation to deal with COVID-19. It highlights the value of domain-optimized, open source frameworks and federated learning models in democratizing high-performance computing. It addresses the use of evolutionary AI to validate these models.

Key Takeaways

- Spiraling costs of drug development, as well as high attrition of drugs across the drug development cycle, have created a bleeding need to accelerate innovation in the drug discovery process.
- SOTA biomedical GPU-powered transformer models are being used to mine real-world data (RWD) to optimize recruitment, better predict clinical outcomes, and develop precision medicine strategies. These models are being utilized to hyper-accelerate the development of innovative therapies to treat COVID-19.
- Federated learning models will drive collaboration and accelerate drug discovery while ensuring data privacy.

Recommended Actions

- Investments in a GPU-based transformer model counterbalance the escalating costs of R&D and significantly accelerate innovation.
- It is important to strategize around the level of support required when outlining a vendor partnership — Would it require only the technology solution? Would it require domain-specific pretrained models? Would it require data scientists and subject matter experts as well?
- Since transformer models are based on deep neural network architecture, evolutionary AI can be used to improve and customize the design of these transformer-based deep networks.

Source: IDC, 2021

May 2021, IDC #US47660321

SITUATION OVERVIEW

Finding the right drug to treat a disease has been an insurmountable task, more challenging than finding a needle in a haystack, costing approximately $2.5 billion per drug as per the Tufts Center for the Study of Drug Development. There is an urgent need to transition from manual, laborious trial-and-error approaches to the use of advanced computational software to accelerate drug discovery. *In silico* medicine is paving the way for innovation. However, the massive size of "multiomics" data sets used for drug discovery, and of real-world data sets to accelerate drug development, requires the industry to evolve hyperscalable infrastructure and develop a very powerful computing data fabric.

Addressing the Challenges

While simulation and modeling have great potential, there still exist multiple challenges that need to be addressed. First, the transition from a purely "genomics" to a "multiomics" world, complemented by deep phenotyping, is resulting in the generation of terabytes of data every day from diverse data sources. Genomic sequencers alone can generate over 10TB of data in one run. Hence, one needs highly scalable infrastructure. Second, humans do not have the capacity to analyze this data. Therefore, one needs to train machine learning algorithms on massive domain-specific labeled data sets, which are not always readily available. Third, one needs to run at least four or five iterations to make these domain-specific machine learning algorithms accurate enough to be predictive. This can be immensely time consuming and would require considerable effort and time to achieve the desired levels of accuracy. Thus there is a clear need for science-at-scale platforms.

Companies such as NVIDIA and AMD have been leaders in the development of powerful GPU-based high-performance computing (HPC) models.

Cloud-native platforms with huge data storage capacity can address the data storage issues. For example, NVIDIA GPU Cloud (NGC) empowers researchers by providing performance-engineered deep learning (DL) software containers containing the NGC Software Stack and a deep learning framework. These frameworks provide highly optimized GPU-enabled code specific to the computations required for training deep neural networks (DNNs). Compute Unified Device Architecture (CUDA) is NVIDIA's parallel computing platform and application programming interface (API), which hyperscales computing performance.

A solution that can accelerate computing, artificial intelligence (AI), and machine learning to speed up the drug discovery and the development process is much needed. Transformer-based neural network architectures, which have become available only in the past few years, operate in a two-stage process. The first step is unsupervised learning, which involves pretraining a large model of algorithms based on unlabeled data. The second step of supervised learning involves fine-tuning those algorithms on smaller amounts of labeled data. This process eliminates the need for very large, labeled data sets, thus accelerating time to market for new lifesaving innovations and reducing the cost of R&D by taking out a lot of the laborious steps.

Use Cases for Transformer Models

NVIDIA's Clara platform is designed to address the aforementioned computational challenges and accelerate drug discovery and healthcare delivery. It provides a collection of over 40 pretrained models, 12+ libraries, AI application frameworks, and reference applications. It leverages the power of GPUs to accelerate discovery and leverage data from diverse sources including genomic, proteomic, microscopy, virtual screening, computational chemistry, visualization, and clinical imaging data.

Translating Chemical Structures to Establish Molecular Interrelationships

NVIDIA has partnered with AstraZeneca to develop a drug discovery model called MegaMolBART. It leverages AstraZeneca's MolBART transformer model and NVIDIA's Megatron framework (which is based on transformer-based neural network architecture), which provide a hyperscalable infrastructure and supercomputing capability. This model is trained on the ZINC chemical compound database to pretrain it to understand chemical structures. ZINC is a free database of commercially available compounds, and ZINC is a recursive acronym that stands for "ZINC is not commercial." It is then trained to translate the chemical structure into a sequence of characters called smile strings (simplified molecular-input line-entry system, a line notation for describing the structure of chemical species using short ASCII strings) and effectively train these models on the language of chemistry in an unsupervised manner. The expectation is that the neural networks trained on molecular structure data will be able to interpret relationships between atoms in real-world molecules.

Fueling Innovation and Hyperscaling Preclinical Candidate Selection

These predictive models will significantly accelerate discovery and reduce lab work and costs. The industry is digitizing data like never before, and GPU-based transformer models will serve as the digital disruptors in bioinnovation. Insilico Medicine has a Pharma.ai drug discovery suite that also includes target discovery and multiomics data analysis (PandaOmics) and clinical trial outcomes predictions (InClinico). Chemistry42 is a core component of this platform and is used for de novo small molecule design and integrates artificial intelligence with computational and medicinal chemistry methods. Insilico Medicine partnered with NVIDIA, leveraging NVIDIA Tensor Core GPUs, to develop a novel preclinical candidate to treat idiopathic pulmonary fibrosis within 18 months of target hypothesis, and the process cost less than $2 million. This is one of the rare examples where an AI-designed molecule for a new disease target has been designated for clinical trials.

3D Protein Structure Prediction

Understanding the 3D structure of proteins is key to assessing the ability of a drug to bind to it, directly impacting the drug's efficacy. Predicting the 3D protein form is very complex, challenging, expensive, and time consuming. DeepMind Technologies, acquired by Google in 2014, developed AlphaFold, a machine learning software that can rapidly and accurately predict the structure of proteins in a matter of minutes, based on its amino acid sequence. This has the potential to significantly accelerate drug development. In 2020, DeepMind's AlphaFold won the Critical Assessment of Protein Structure Prediction (CASP) competition, which compares the code for predicting protein folding, achieving a score of 87 GDT (Global Distance Test) or effectively 87% accuracy, the highest score ever. This score is considered to be as good as results obtained from physical observations of proteins and would be so accurate that the guesstimated positions of the amino acids may be off by distance of just 1.6 Angstrom, equivalent to the width of an atom. AlphaFold identifies the folded protein structure as a "spatial graph" using a neural network system where residues are treated as nodes connected with edges. About 100–200 GPUs of computing power were used to run the system for a few weeks and train it on around 170,000 protein structures from the Protein Data Bank. AlphaFold could determine the structure of a bacterial protein that the Max Planck Institute for Developmental Biology has been working on for years without success using x-ray diffraction data, within half an hour.

Mining Real-World Data to Accelerate Drug Development and Training Language Models Based on Deep Learning Transformer Architecture

The ability to leverage real-world data (RWD) is gaining increasing importance in drug development. Its applications are diverse and include deriving meaningful insights to help drive a precision medicine

strategy, improving the prediction of health outcomes, and accelerating patient recruitment, especially in the rare disease therapeutic area. However, the challenge lies in the fact that this data resides as unstructured data in diverse data sources, such as scientific journal articles, physician notes, and medical imaging reports, making it very difficult to analyze. The University of Florida partnered with NVIDIA to build Gatotron, the world's largest clinical language model. It leveraged NVIDIA's Megatron, a PyTorch-based framework, used to train giant language models based on the deep learning transformer architecture. Over 300 million unstructured notes across 2 million patients and 50 million patient encounters were used to pretrain NVIDIA's state-of-the-art (SOTA) language model for biomedical and clinical NLP, BioMegatron, available on NVIDIA's GPU Cloud.

Modern NLP models involve an initial step of unsupervised pretraining of the model on a large volume of text, followed by supervised fine-tuning on a smaller volume of data. Since pretraining is a highly computationally intensive step, memory constraints can pose challenges as the volume of data on which the model is being trained increases. Therefore, splitting the model parameters across multiple GPUs can help address this problem. The Megatron-LM model achieved 76% scaling efficiency on 512 GPUs as compared with a fast, single-GPU baseline. The second step of fine-tuning was done using NeMo, an open source toolkit for conversational AI.

BioMegatron is the largest biomedical transformer-based language model ever trained. It has been trained on over 6.1 billion words from PubMed as well as on abstracts and full-text biomedical journal articles. While previous NLP models did not perform very well in clinical research, BioMegatron, which had been trained specifically on this data, could perform very well on common biomedical NLP tasks, such as named entity recognition (NER), relation extraction (RE), and question answering (QA).

Gatotron is being used by the University of Florida to extract information from massive untapped data sets including electronic medical records (EMRs). The objective is to map patients to clinical trials, predict life-threatening conditions to enable timely intervention, and use this data to power clinical decision support systems (CDSS). NVIDIA's Megatron training framework has democratized the ability for academic medical centers to build their own clinical language models.

Powering Cancer Research and Leveraging a GPU-Accelerated Deep Learning Data Fabric

Vyasa Analytics is dealing with the challenges associated with mining text-heavy unstructured data, which typically does not have a predefined data model or is not organized in a predefined manner. It is utilizing deep learning algorithms to derive insights at scale from unstructured data as well as from images and data streams. It has created a GPU-accelerated deep learning data fabric, Vyasa Layar, to drive powerful analytics cancer research, clinical trial protocol design, and biomedical data harmonization. It leverages NVIDIA's Clara Discovery's BioMegatron language model and NVIDIA DGX systems, along with its own proprietary deep learning analytics modules, to create a next-generation data integration and analytics architecture.

Fundamentally, to be able to understand the large and diverse language of chemistry, biology, imaging, doctors' notes, and more, one needs superpods, which contain the highest-end GPUs in the world, the A100s, as well as the networking fabric needed for internodal, intercommunication between GPUs, complemented by CPU cores. The combination of these technologies forms the SuperPOD. The SuperPOD architecture includes 140 of the A100s, complemented by the InfiniBand HDR networking fabric, and networking interfaces, enabled by Mellanox, yielding a speed of over 200Gbps. The DGX A100 is a building block that has 5 petaflops of computing power. This is possible since one does not have to spend a lot of time communicating between the nodes. This feature of the SuperPod

architecture is very important because it enables multinode AI training. The NVIDIA DGX SuperPOD provides access to large storage for diverse data types, along with very high-speed bandwidth.

Computing as the New Instrument of Science and Discovery

Deep learning is compute intensive and has led to the adoption of GPUs. The datacenter is the new unit of computing. NVIDIA's full-stack HPC and AI optimization delivers the performance of 1,000 CPU servers in a single DGX A100. DGX SuperPOD, with NVIDIA DGX A100 systems, is a fully integrated, fully network-optimized AI datacenter as a product. MLPerf, an industry benchmark for AI, has emerged and continues to evolve to mission-critical workloads from computer vision to translation to 3D medical imaging and inferencing. NVIDIA GPUs have won all tests of AI inference in datacenter and edge computing systems in the MLPerf's latest benchmarking exercise, with its A100 outperforming CPUs by up to 237 times in datacenter inference, as per MLPerf Inference 0.7 benchmarks.

Leveraging Supercomputing to Mine Data and Drive Pipeline Development

GlaxoSmithKline (GSK) has established a partnership between its United Kingdom–based AI lab and NVIDIA, leveraging NVIDIA's deep expertise in GPU optimization and high-performance computational pipeline development, complemented by GSK's imaging, genomic, and genetic data sets to drive innovation of new therapies. This includes the use of NVIDIA's Clara Discovery platform and NVIDIA DGX A100 systems. GSK will also have access to NVIDIA's Cambridge-1, the United Kingdom's most powerful AI supercomputer, and NVIDIA's data scientists. This partnership will provide GSK with additional computational power and state-of-the-art AI technology.

Using High-Performance Computing to Hyperscale Drug Discovery

NVIDIA and Schrödinger have partnered to develop a joint solution around the NVIDIA DGX SuperPOD to evaluate billions of compounds in minutes. The intent is to complement the physics-based modeling of the Schrödinger platform, with the NVIDIA Ampere architecture and its multi-instance GPU technology and state-of-the-art AI frameworks to optimize the Schrödinger suite. Customers would have the option of deploying the Schrödinger software on a single DGX system or on a cluster of 20+ systems to create a DGX SuperPOD that can power super high-throughput lead generation and drive accelerated drug discovery. As against the hundreds of thousands of hours of GPU time required to evaluate tens of thousands of molecules on high-performance computers, the transformer-enabled supercomputing model can replace these computationally intensive physics-based approaches. This will provide the pharmaceutical industry the opportunity to accelerate drug discovery at supercomputing scale on its own private clouds.

Classifying Cell Responses to Small Molecules to Drive Discovery Using DGX SuperPOD Reference Architecture

Recursion, a clinical-stage biotech company, flips the drug discovery model. Robots carry out a million experiments a week, which involve making healthy human cells sick, taking pictures of these cells, and leveraging machine learning to interpret how the sick cells differ from the healthy ones. This information is used to discover the drugs that can treat the sick cells. Recursion built BioHive-1, a supercomputer with NVIDIA DGX SuperPOD reference architecture. As of January 2021, BioHive-1 is estimated to rank at number 58 on the list of the top 500 of the world's most powerful computer systems. BioHive-1 will enable Recursion to run deep learning projects that once took a week, to now be completed within a day, and automatically classify cell responses to small molecules, extract insights, and establish relationships between the data sets, a process which would be way too

complex for humans. NVIDIA built BioHive-1 for Recursion in just three weeks. This was possible since the NVIDIA DGX SuperPOD architecture is really a datacenter architecture that is already established and is being utilized all over the world.

Studying Epigenetic Changes in Rare Diseases Using GPU-Accelerated Deep Learning Technology

ATAC-seq (Assay for Transposase-Accessible Chromatin with high-throughput sequencing) is a molecular technique used to assess genome-wide chromatin accessibility, using the hyperactive transposase Tn5, which integrates next-generation sequencing (NGS) adapters into open chromatin regions. This allows the genome to be sequenced by NGS and the regions of the genome with open or accessible chromatin to be analyzed using bioinformatics. Usually, the ATAC-seq approach requires tens of thousands of cells so as to be able to eliminate unnecessary noise. The AI-based AtacWorks toolkit, developed by Harvard University's Department of Stem Cell and Regenerative Biology and NVIDIA, has leveraged GPU-accelerated deep learning technology to remove the noise from sequencing. Hence a low cell count, as low as tens of cells, against the tens of thousands required previously can be utilized to obtain fast and clean results. This provides researchers with the capability to study epigenetic changes associated with rare cell development and diseases and identify rare mutations as well.

Accelerating Discovery by Employing AMD EPYC Processors and Radeon Instinct Accelerators

Symmetric Computing, through its Venture Development Center at the University of Massachusetts (UMass), Boston, is working on developing a cure for Alzheimer's disease. It has developed a Virtual Drug Discovery Platform using AutoDock Vina (a molecular modeling simulation software), NAMD (Nanoscale Molecular Dynamics, a molecular dynamics simulation software), and TensorFlow (an open source machine learning platform). AutoDock Vina is used to simulate protein-ligand docking to identify likely drug candidates. Then, NAMD is used to further short-list the best candidates, which will then go on for validation in the lab.

The simulations are performed using a database of over 500 million small molecules with 3D structure, 23,000 human proteins with 3D structure, and data on all known drugs, both deployed and experimental. These must all be tested together to see which compounds block or enhance the function of which proteins. One has to not only evaluate whether the given molecule binds to and inhibits the proteins specifically associated with Alzheimer's, but one must also determine whether the compound binds to other human proteins, resulting in undesirable side effects. Typically, this would be a complex, time-consuming trial-and-error process that would take years if performed using the wet laboratory process. To address the significant challenges associated with the storage of data and the lack of speed, Symmetric Computing employed AMD EPYC processors and Radeon Instinct accelerators (AMD's deep learning-oriented GPUs). The AMD Instinct MI100 accelerator is the world's fastest HPC GPU and the first x86 server GPU to surpass the 10 teraflops (FP64) performance barrier and is the world's fastest accelerator for scientific research.

Leveraging Embedded AI to Develop Smart Sensors and Intelligent Medical Devices

AI in the healthcare space needs two computers, the AI supercomputer to develop and continuously improve algorithms and the AI edge and embedded computer to run the AI applications in real time to help healthcare professionals make real-time decisions. One increasingly sees the use of smart

sensors (wherein AI algorithms are being updated on an ongoing basis, connecting the latest research breakthroughs with day-to-day practice of medicine) in medical instruments.

Clara AGX is an embedded AI platform for medical devices and is being leveraged by companies such as Carestream Health for creating smart x-ray rooms, and for companies like Activ Surgical to deliver real-time surgical guidance, and by researchers to create AI models for ultrasound and endoscopy. The Clara AGX architecture (NVIDIA Jetson AGX Xavier, NVIDIA RTX 6000 GPU, and NVIDIA Mellanox ConnectX-6 SmartNIC) allows developers with a software development kit (SDK) to streamline AI models for the creation of live video feeds to enable smart medical operations.

Accelerating COVID-19 Research Leveraging GPUs

The COVID-19 pandemic has emphasized the urgent need for a more efficient and an effective drug discovery process, and the industry has been collaborating to accelerate discovery of vaccines and therapies for COVID-19. Considering the speed at which SARS-CoV-2 has been mutating and new variants have been developing, there is a heightened need for accelerating innovation. The global COVID R&D Alliance is an example of the global collaborative effort to accelerate discovery of an antiviral therapeutic for SARS-CoV-2. Google Cloud has donated over 16 million hours of NVIDIA GPU time to support this initiative. The NVIDIA Clara Parabricks genome analysis toolkit provides researchers the opportunity to sequence and analyze genomes up to 50 times faster, and this solution has been used across the globe to sequence the viral genome and the DNA of COVID-19 patients.

It is equally important to understand the structure of the viral spike protein, which mediates SARS-CoV-2 entry into host cells and serves as a key target for vaccines, therapeutic antibodies, and diagnostics. Understanding the 3D structure of proteins requires the analysis of thousands of images and complex computing, making it difficult to deliver high-resolution structures. Cryogenic electron microscopy (cryo-EM) is used to analyze protein structures in their native and near-native states and derive the 3D structure of the protein. Thousands of images, amounting to several terabytes of data, can be generated in a single cryo-EM run. GPU processors help shorten the significantly long processing times and the requirement for expert intervention, and prior structural knowledge, which are typically associated with these compute-intensive steps associated with single-particle workflows. The need of the hour is commercial-grade, nonexpert software, leveraging algorithms to automate specialized and time-intensive tasks. Structura Biotechnology, a Toronto-based start-up, developed cryoSPARC, a GPU-accelerated software to address exactly this need. Merck uses structure-based drug design (SBDD) to build its pipeline, by using the cryoSPARC AI-infused platform, powered by NVIDIA's V100 Tensor Core GPUs to automate the high-quality, high-throughput discovery of protein structure. The first 3D atomic-scale map of SARS-CoV-2 was created in February 2020. It was developed in just 12 days by researchers at the National Institutes of Health and the University of Texas at Austin using cryoSPARC.

The "Corona" cluster at Lawrence Livermore National Laboratory (LLNL) focuses on COVID-19 drug discovery and vaccine research. It simulates the development of antibody candidates that have the potential to block SARS-CoV-2 from human receptors and prevent infection. The COVID-19 HPC Consortium includes a dozen member institutions across government, industry, and academia and is spearheaded by the White House Office of Science and Technology Policy, the U.S. Department of Energy, and IBM. It provides free compute cycles to COVID-19 researchers from outside institutions. LLNL uses the National Nuclear Security Administration's (NNSA's) Corona system for predictive biomedical modeling of COVID-19 therapies. This has been powered by nearly 1,000 AMD Radeon

Instinct MI50 GPU accelerators, increasing the throughput of the cluster from about 4.5 petaflops to more than 11 petaflops, at peak performance.

About 121 additional AMD MI50 accelerators enable deep learning, delivering up to 26.5 teraflops of native peak theoretical half precision or up to 13.3 teraflops of single-precision peak theoretical floating-point performance, combined with 32GB of high-bandwidth memory. The objective is to identify antibodies that might bind to the SARS-CoV-2 spike protein and to study the attributes of over a billion chemical compounds to screen against four binding sites in the SARS-CoV-2 proteins by machine learning and molecular docking studies.

Open, Robust, and Trustworthy AI

Open Source, Domain-Specific AI

The Medical Open Network for AI (MONAI) is a domain-optimized, open source framework for AI in healthcare. It is a PyTorch-based framework supporting the development of AI for medical imaging, enabling reproducibility of research experiments for comparisons against state-of-the-art implementations. It offers over 20 pretrained models, including ones recently developed for COVID-19, as well as the latest training optimizations on NVIDIA DGX A100 GPUs that provide up to a sixfold acceleration in training turnaround time. It is providing a production-ready framework, powering the academic community to further innovation to build enterprise-ready features and accelerate the transition from research to production. It is being used as the AI framework for imaging by the German Cancer Research Center (DKFZ), King's College London, Mass General, Stanford University, and Vanderbilt University and has had over 50,000 downloads and has 58 contributors for world-leading healthcare organizations.

Collaborative AI

There are many data characteristics that challenge AI in healthcare and can be addressed with a collaborative and federated learning paradigm. Federated learning allows algorithms to be trained collaboratively without exchanging the data itself, thus addressing the problem of data governance and privacy. By moving the training to the data instead of the data to the training, learning can happen at the edge, behind firewalls. Data privacy can be respected by never sharing data and only contributing to the model weights. Data paucity can be overcome by working together to learn from diverse and much larger data sets than any one entity.

A generalized AI model for COVID-19 patients was developed by 20 hospitals around the world using Clara Federated Learning. Using this federated learning model, researchers could use chest x-ray, patient vitals, and lab data to train a local model and share only a subset of model weights back with the global model. The goal of this model is to predict the likelihood that a person showing up in the emergency room will need supplemental oxygen, thus helping physicians determine the appropriate level of care and ICU needs for patients.

Machine Learning Ledger Orchestration for Drug Discovery (MELLODDY) is a 17-member industry-academia consortium, including 10 global pharma companies, that is leveraging NVIDIA's federated learning model to accelerate drug discovery. A distributed deep learning model that can travel across individual pharma cloud clusters of NVIDIA V100 Tensor Core GPUs will enable training on annotated data for over 10 million chemical compounds. This will virtualize drug discovery and enable pharma to leverage the world's largest collaborative drug compound data set for AI training without sacrificing data privacy while being able to fine-tune the AI model to its specific needs.

The Broad Institute of MIT and Harvard, Verily (an Alphabet company), and Microsoft Corp. have partnered to create a federated ecosystem accelerating biomedical research. Verily and the Broad Institute have developed Terra, a secure, scalable, open source platform that will provide open access to a federated ecosystem of gen(omics) and biomedical data repositories; Microsoft provides the cloud infrastructure, data, and AI technologies, creating a scalable model, enabling collaboration, and powering innovation.

Evolutionary AI

One of the challenges associated with deep learning models is the explainability of these models — it is often difficult to know what the model is actually doing. Evolutionary AI can be leveraged to validate these models and to author a set of human understandable rules to interpret exactly what happened. Cognizant has created the Learning Evolutionary AI Framework (LEAF), a toolkit that uses evolutionary algorithms, deep learning, and distributed computation technology. LEAF can be used to automatically develop deep learning models, develop adaptive Prescriptor models to recommend best alternatives and augment decision making, and assess the trustability of the prediction of other DL models.

In this approach, a Predictor model is trained on context, action, and outcome (CAO) data from the real world. The Predictor model then serves as a surrogate model to train a Prescriptor model, which recommends optimal actions. Once these are implemented in the real world, data that is generated enriches the CAO data set, creating a feedback loop for further training the Predictor model. Since the Predictor model is a "surrogate" model, not based on real-world data, this makes the process faster and more cost effective. The Predictor and Prescriptor Models are usually deep learning, neural network-based models.

Cognizant has developed RIO, a third model, to validate the Prescriptor model (or any external DL model) based on Gaussian Programming techniques. It can provide certainty estimations for individual predictions that a model makes.

The company has used this model to predict how COVID-19 will unfold around the world and optimize COVID-19 non-pharmaceutical interventions (NPIs), such as the creation of regional policies for reopening.

RIO could also be used to improve and customize the design of GPU-based transformer models in enabling drug discovery and reduce false positives for such systems.

ADVICE FOR THE TECHNOLOGY BUYER

Computational drug discovery, driven by GPU-enabled transformer models, isn't the future, it is the present. GPUs have a large number of cores, which enable better computation of multiple parallel processes. This is crucial for supporting various deep learning applications. While the ability to hyperscale and accelerate innovation comes at a cost, one needs to compare this not only with the typical ever-increasing costs of drug development but also with the average time of over a decade that it takes to bring a drug to the market and the extremely high failure rate of over 90%.

A lot of time and compute cost goes into developing pretrained models. It helps to partner with a technology provider that can provide you with models that are pretrained on clinical data as that would not only contribute significantly to the success of the model but also result in significant savings in time and money. Fine-tuning of the model on your own data, specific to your model, is important. One

should also assess whether one has the data science expertise within the organization or one would prefer to lean on the vendor for support. A strong governance model and transparency regarding the inputs required from both parties are key. Evolutionary AI models can be used to improve and customize the design of transformer-based deep networks.

LEARN MORE

Related Research

- *Leveraging Cloud, AI, and Analytics to Derive Insights for Biomedical Research Using Terra, a Next-Generation, Open Source Platform – A Microsoft, Verily, and Broad Partnership* (IDC #IcUS47316621, January 2021)
- *Future of Intelligence in Life Sciences* (forthcoming)
- *Finding a Clear Path Forward – Post COVID-19* (forthcoming)
- *Precision Medicine* (forthcoming)

Synopsis

This IDC Perspective provides deep insight into the crucial role of GPU-powered transformer models in accelerating innovation and designing new therapies. This is the era of a data-driven drug discovery and development. New deep learning models are replacing the conventional strategy of knocking down gene expression using small interfering RNA (siRNA), by binding to the corresponding pieces of messenger RNA (mRNA) of specific genes and blocking their expression. A neural network can now be trained on thousands of mutations (equivalent to massive data sets) to encode phenotypes that can allow drugs to be evaluated for their efficacy and safety. The two-stage model of modern NLP – involving unsupervised pretraining on massive data sets, followed by supervised training on smaller annotated data sets to fine-tune the process – is being widely implemented. However, the massively parallel architecture of GPUs provides the high-compute performance that is key to accelerating the drug discovery and development process. This document discusses the use cases of GPU-powered transformer models and examines their critical value in driving innovation in a world torn by the COVID-19 pandemic. It also highlights the importance of federated learning platforms in democratizing high-performance computing and accelerating drug discovery.

"The pandemic has brought to the forefront the urgent need to accelerate drug discovery and development like never before. Conventional models need to be revisited. The world is undergoing a digital biology revolution as we speak. GPU-enabled deep learning transformer models will fuel innovation and will help deliver lifesaving therapies to patients. Federated learning models and domain-optimized, open source frameworks can drive collaboration and accelerate innovation. Evolutionary AI can play a role in building trust in these models," said Dr. Nimita Limaye, research VP, Life Sciences R&D Strategy and Technology at IDC.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com