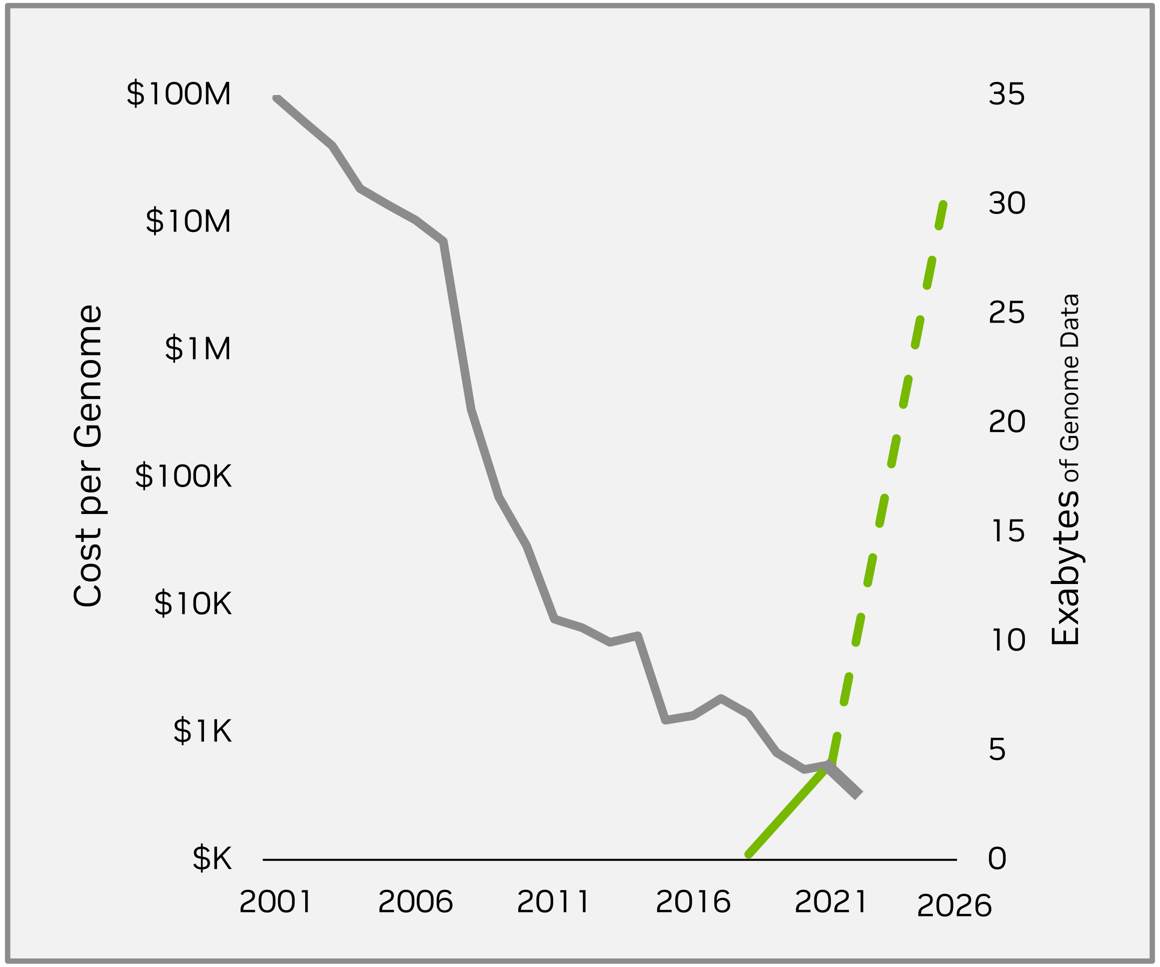
Accelerating Biomedical Research with BioNeMo & Parabricks

陳映嘉 Ying-Ja Chen, Solutions Architect, Healthcare & Life Sciences

Genomics Projects Will Exceed 40 Exabytes in the Next Decade

As sequencing becomes less expensive, the data deluge grows



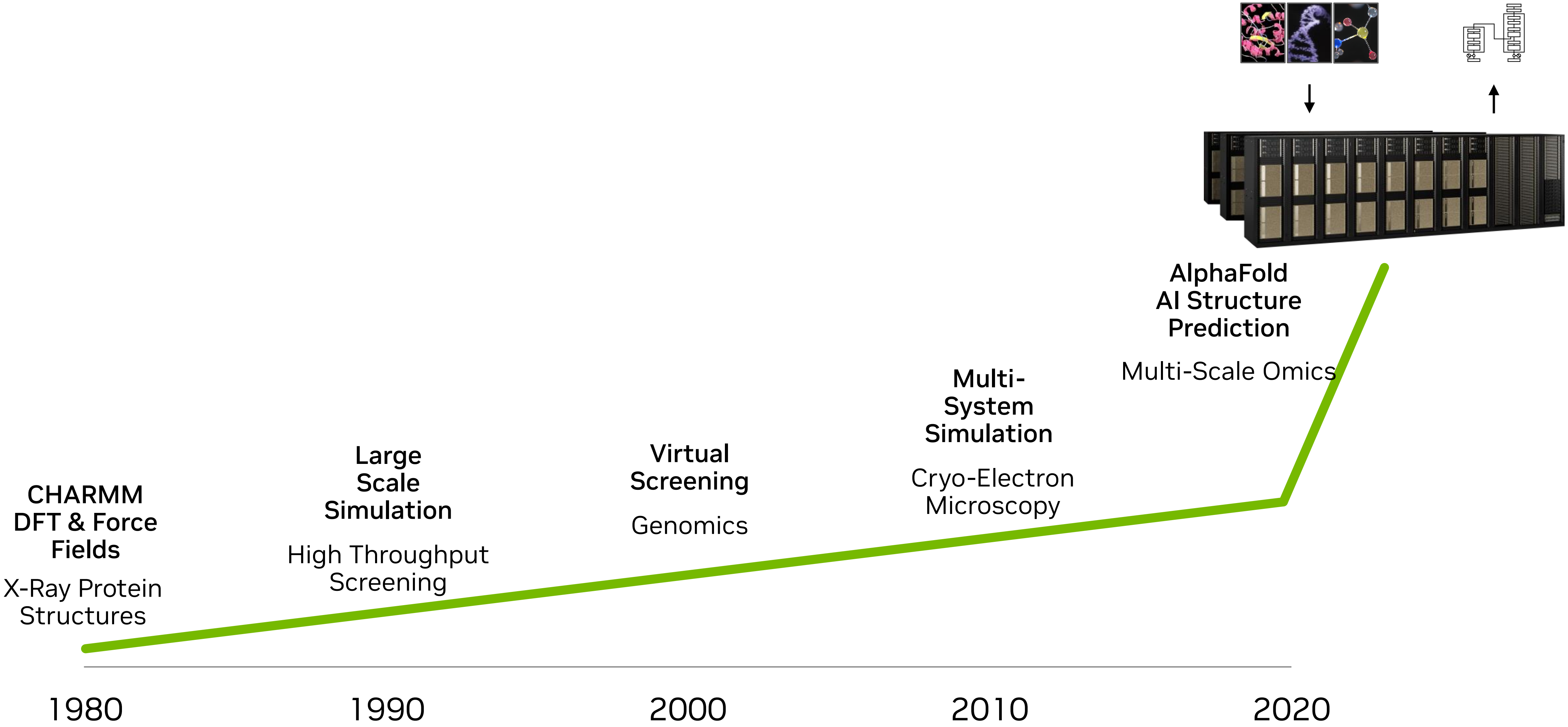
> “Our ability to sequence DNA has far outpaced our ability to decipher the information it contains, so genomic data science will be a vibrant field of research for many years to come.”

National Human Genome Research Institute

Source: https://www.genome.gov/about-genomics/fact-sheets/Genomic-Data-Science

Biomedical Research and Drug Discovery Is at an Inflection Point

Computer Aided Drug Discovery is Expanding Exponentially



CHARMM DFT & Force Fields

X-Ray Protein Structures

Large Scale Simulation

High Throughput Screening

Virtual Screening

Genomics

Multi-System Simulation

Cryo-Electron Microscopy

AlphaFold AI Structure Prediction

Multi-Scale Omics

AI is Transforming the Drug Discovery Process

Deep learning is an essential tool for modern R&D

nature machine intelligence

REVIEW ARTICLE

https://doi.org/10.1038/s42256-022-00463-x

Check for updates

The transformational role of GPU computing and deep learning in drug discovery

Mohit Pandey[1,5], Michael Fernandez[1,5], Francesco Gentile[1], Olexandr Isayev[2], Alexander Tropsha[3], Abraham C. Stern[4]✉ and Artem Cherkasov[1]✉

Deep learning has disrupted nearly every field of research, including those of direct importance to drug discovery, such as medicinal chemistry and pharmacology. This revolution has largely been attributed to the unprecedented advances in highly parallelizable graphics processing units (GPUs) and the development of GPU-enabled algorithms. In this Review, we present a comprehensive overview of historical trends and recent advances in GPU algorithms and discuss their immediate impact on the discovery of new drugs and drug targets. We also cover the state-of-the-art of deep learning architectures that have found practical applications in both early drug discovery and consequent hit-to-lead optimization stages, including the acceleration of molecular docking, the evaluation of off-target effects and the prediction of pharmacological properties. We conclude by discussing the impacts of GPU acceleration and deep learning models on the global democratization of the field of drug discovery that may lead to efficient exploration of the ever-expanding chemical universe to accelerate the discovery of novel medicines.

Originally developed to accelerate three-dimensional graphics, the benefits of GPUs for powerful parallel computing were quickly praised by the scientific community. The earliest attempts to use GPUs for scientific purposes employed the programmable shader language to run calculations. In 2007, NVIDIA released Compute Unified Device Architecture (CUDA) as an extension of the C programming language, together with compilers and debuggers, opening the floodgates for porting computationally intensive workloads into GPU accelerators. Further advances came from the release of common maths libraries such as fast Fourier transforms and basic linear algebra subroutines, which were foundational to scientific computing. In the same year, the first computational chemistry programs were ported to GPUs, enabling efficient parallelization of molecular mechanics and quantum Monte Carlo[1] calculations.

In September 2014, NVIDIA released cuDNN, a GPU-accelerated library of primitives for deep neural networks (DNNs) implementing standard routines such as forward and backward convolution, pooling, normalization and activation layers. The architectural support for training and testing subprocesses enabled by GPUs seemed to be particularly effective for standard deep learning (DL) procedures. As a result, an entire ecosystem of GPU-accelerated DL[2] platforms has emerged. While NVIDIA's CUDA is a more established GPU programming framework, AMD's ROCm[3] represents a universal platform for GPU-accelerated computing. ROCm introduced new numerical formats to support common open-source machine learning libraries such as TensorFlow and PyTorch; it also provides the means for porting NVIDIA CUDA code into AMD hardware[4]. It is important to note that AMD not only is catching up to the ROCm platform in the GPU computing race, but also recently introduced the new flagship GPU architecture AMD Instinct M1200 Series[5] to compete with the latest NVIDIA Ampere A100 GPU architecture[6].

The fields of bioinformatics, cheminformatics and chemogenomics in particular, including computer-aided drug discovery (CADD), have taken advantage of DL methods running on GPUs. Most challenges in CADD have routinely faced combinatorics and optimization problems, and machine learning has been effective at providing solutions for them[7]. Thus, major progress has been made in DL for CADD applications such as virtual screening, de novo drug design, absorption, distribution, metabolism, excretion and toxicity (ADMET) properties prediction and so on (Fig. 1).

Herein, we discuss the effects of GPU-supported parallelization and DL model development and application on the timescale and accuracy of simulations of proteins and protein–ligand complexes. We also provide examples of DL algorithms used for structure determination in cryo-electron microscopy (cryo-EM) and 3D structure prediction of proteins.

GPU computing and DL for molecular simulations

GPU acceleration comes from massive data parallelism, which arises from similar independent operations performed on many elements of the data. In graphics, an example of a common data parallel operation is the use of a rotation matrix across coordinates describing the positions of objects as a view is rotated. In a molecular simulation, data parallelism can be applied to independent calculation of atomic potential energies. Similarly, DL model training involves forward and backward passes that are commonly expressed as matrix transformations that are readily parallelizable (Fig. 2).

Accelerating molecular dynamics simulations on GPUs. The development of GPU-centred molecular dynamics codes in the past decade led hundred-fold reductions in the computational costs of simulations compared with central processing unit (CPU)-based algorithms[8]. Consequently, most molecular dynamics engines (such as AMBER (assisted model building with energy refinement)[9], GROMACS (Groningen machine for chemical simulations)[10] and NAMD (nanoscale molecular dynamics)[11]) now provide GPU-accelerated implementations. GPUs not only are well suited

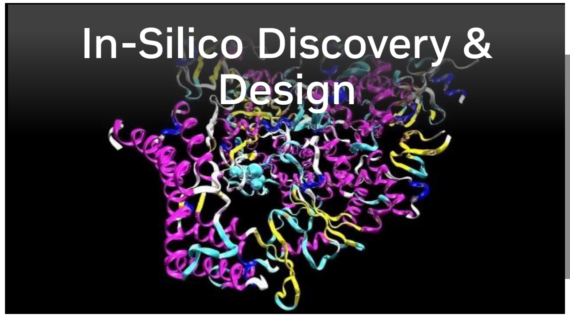
[1] Vancouver Prostate Centre, Department of Urologic Sciences, University of British Columbia, Vancouver, British Columbia, Canada. [2] Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA, USA. [3] Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [4] Nvidia Corporation, Santa Clara, CA, USA. [5] These authors contributed equally: Mohit Pandey, Michael Fernandez. ✉e-mail: astern@nvidia.com; acherkasov@prostatecentre.com

NATURE MACHINE INTELLIGENCE | VOL 4 | MARCH 2022 | 211–221 | www.nature.com/natmachintell 211

Source: https://www.nature.com/articles/s42256-022-00463-x

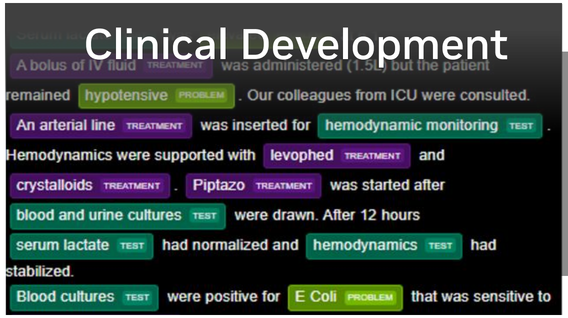
Target ID

- Gene Expression Prediction, scRNA analysis
- Accelerated cryo-em & protein structure prediction
- Knowledge synthesis from scientific literature



In-Silico Discovery & Design

- Active-learning virtual screening
- AI powered molecular property prediction & generation
- Drug-target interaction prediction

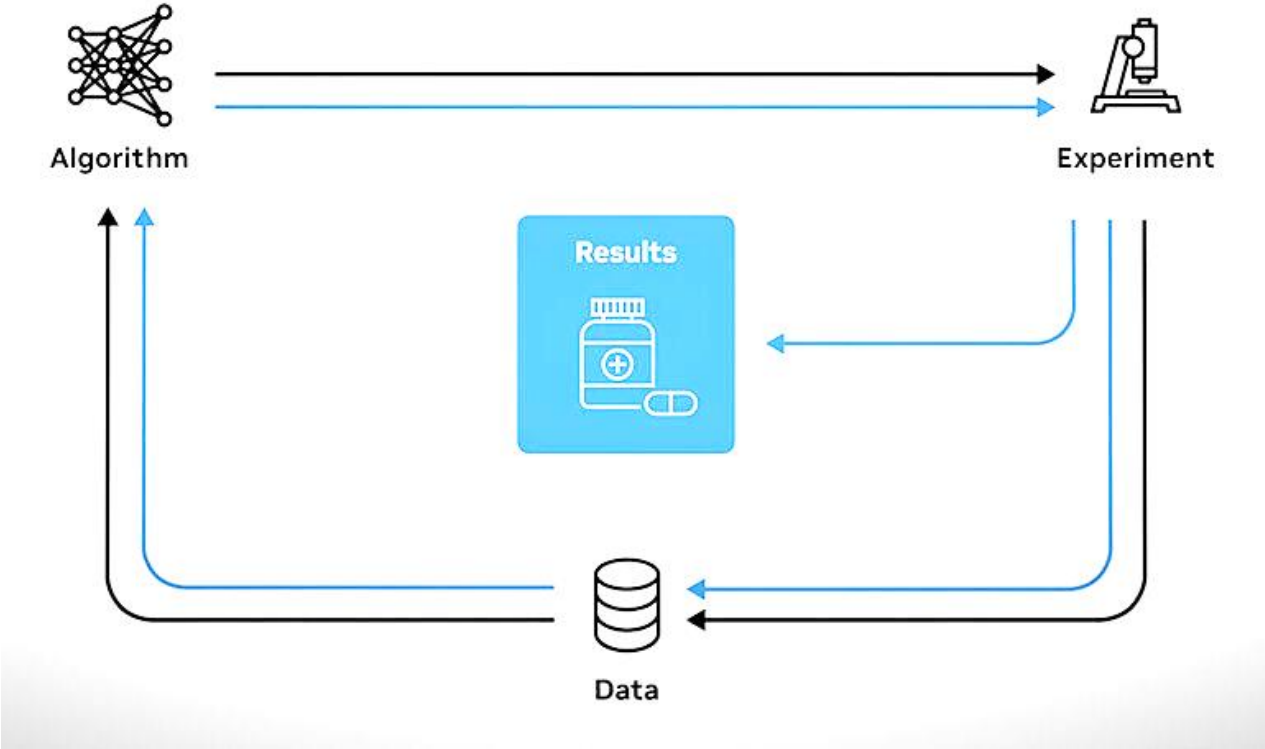


Clinical Development

A bolus of IV fluid TREATMENT was administered (1.5L) but the patient remained hypotensive PROBLEM . Our colleagues from ICU were consulted. An arterial line TREATMENT was inserted for hemodynamic monitoring TEST . Hemodynamics were supported with levophed TREATMENT and crystalloids TREATMENT . Piptazo TREATMENT was started after blood and urine cultures TEST were drawn. After 12 hours serum lactate TEST had normalized and hemodynamics TEST had stabilized.
Blood cultures TEST were positive for E Coli PROBLEM that was sensitive to

- NLP for Clinical trial matching
- Adverse event monitoring
- Histopathology/Radiology/OMICS biomarker ID

Lab in a Loop: AI to Transform Drug Discovery and Development



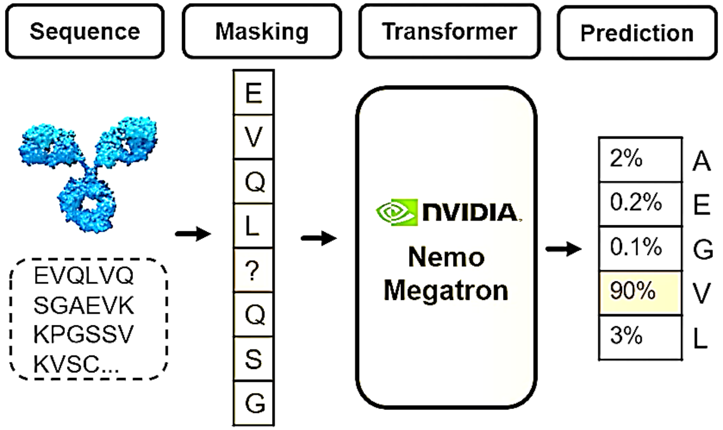
Aviv Regev, Head of Genentech Research and Early Development

Genentech
A Member of the Roche Group

Generative AI Is Used to Design Biologics

Antibody Foundation Models | Post-processing Analysis

Antibody LLM based on ESM-1nv in BioNeMo



Sequence → Masking → Transformer → Prediction

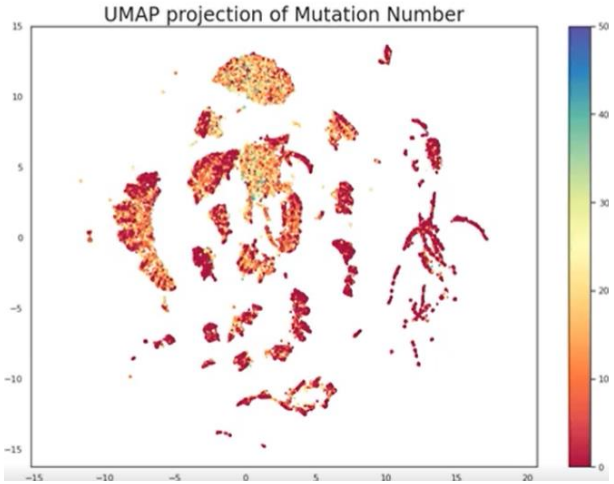
EVQLVQ SGAEVK KPGSSV KVSC...

E V Q L ? Q S G

NVIDIA Nemo Megatron

2%	A
0.2%	E
0.1%	G
90%	V
3%	L

Downstream processing by RAPIDS, *e.g.*, UMAP



20sec/ structure — Faster protein structure prediction

100x — Faster post-training analysis

< 1 month — From onboarding to first pretrained protein LLM

Christopher Langmead
Director, Digital Biologics Discovery
AMGEN

NVIDIA DGX Cloud
AI-training-as-a-service solution

NVIDIA Base Command Platform
for workflow management

NVIDIA AI Enterprise
RAPIDS for data post-processing

NVIDIA BioNeMo
For training and inferencing

NVIDIA Clara for Healthcare and Life Sciences

World's Largest Data Industry | 36% CAGR by 2025

GENOMICS

MEDICAL DEVICES

DRUG DISCOVERY

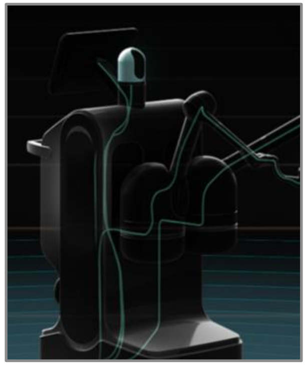
MEDICAL IMAGING

DIGITAL HEALTH

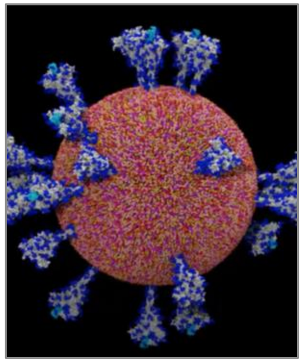
NVIDIA CLARA

PARABRICKS
Genomics

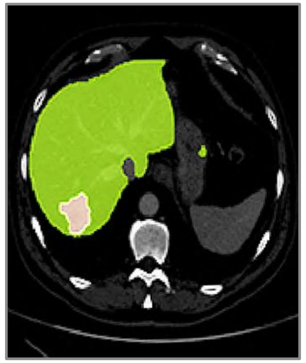
ISAAC
Robotics



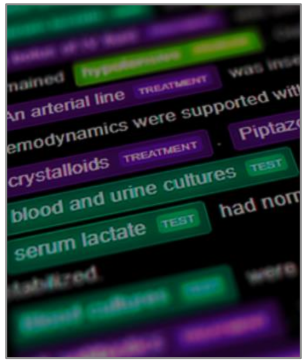
HOLOSCAN
Instruments



BIONEMO
Biomolecules



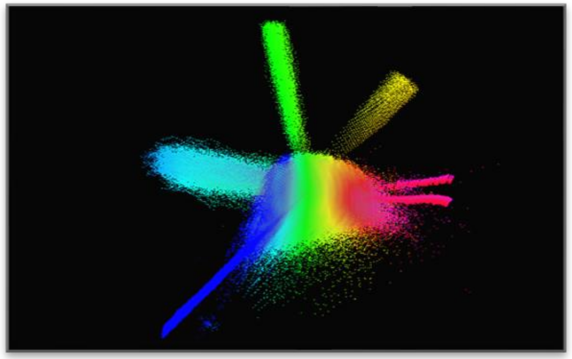
MONAI
Imaging



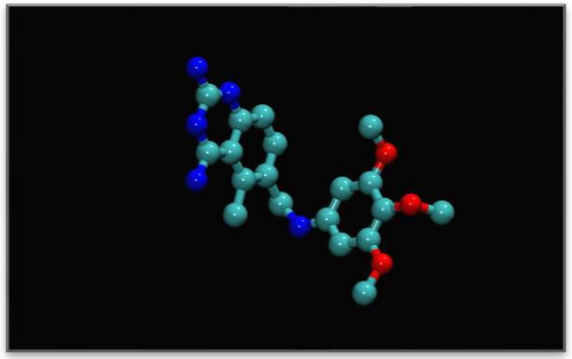
NEMO
Natural Language

BioNeMo Framework Supports Optimized Biomolecular Models

Proteins | Small Molecules | Genomics

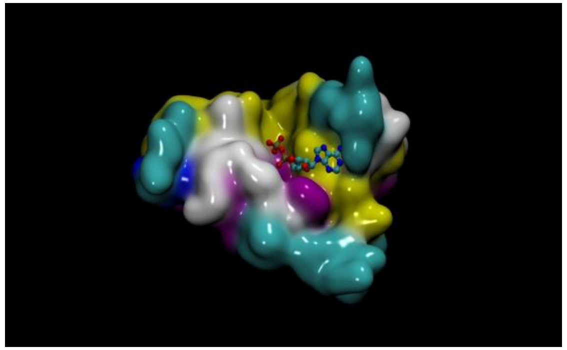


ESM-1 | ESM-2
Protein LLMs



MegaMolBART
Generative Chemistry Model

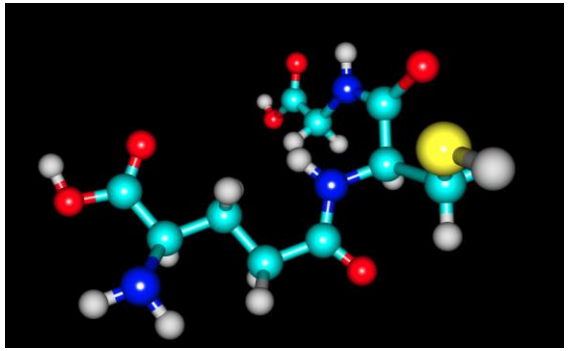
ProtT5
Protein Sequence Generation



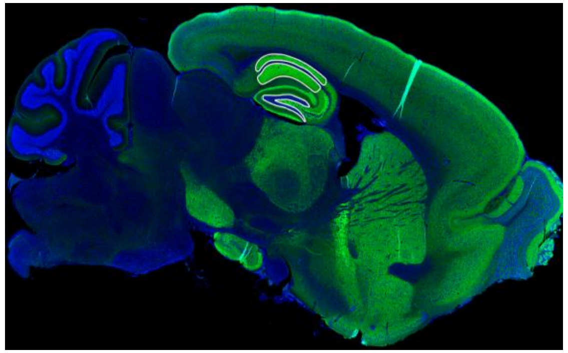
DiffDock | EquiDock
Docking Prediction

NEW: OpenFold
3D Protein Structure Prediction

NEW: DNABERT
DNA Sequence Model



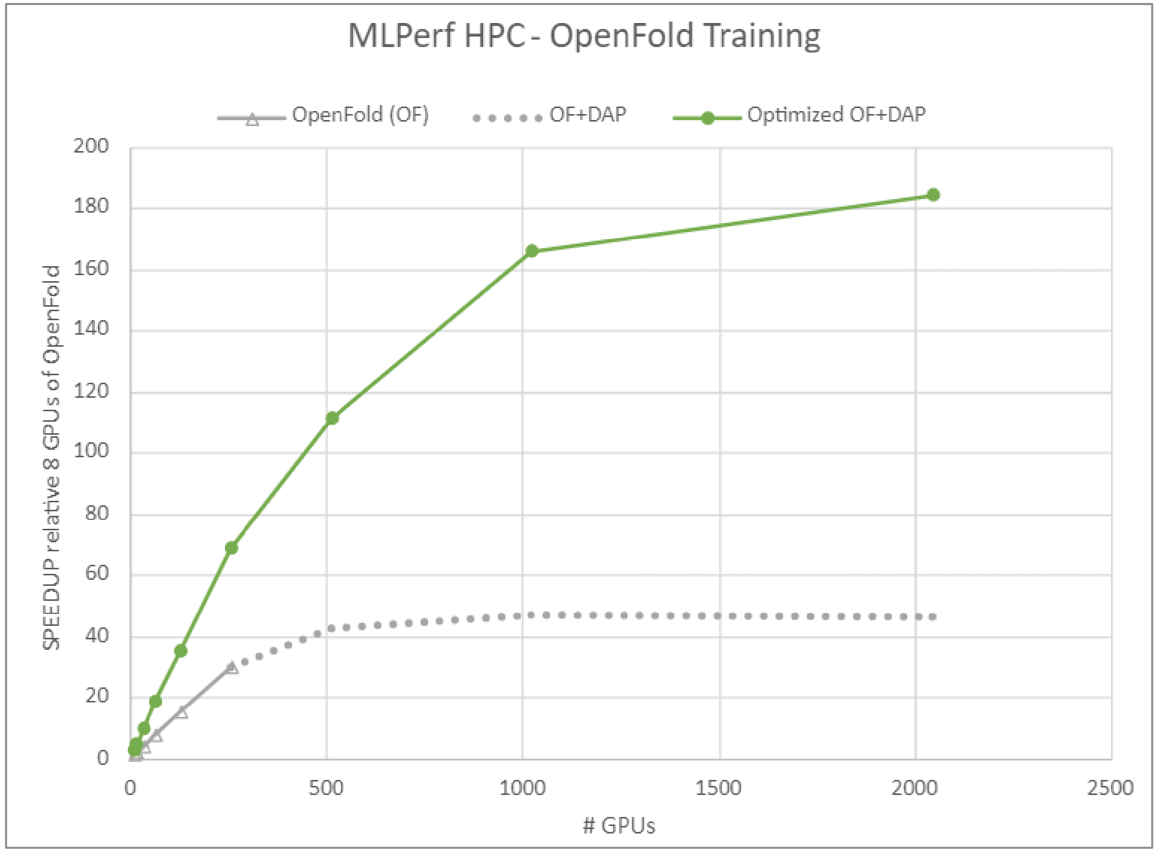
NEW: MolMIM
Molecular Generation



BETA: Geneformer
Single Cell Expression Model

Optimizing OpenFold Training for Drug Discovery

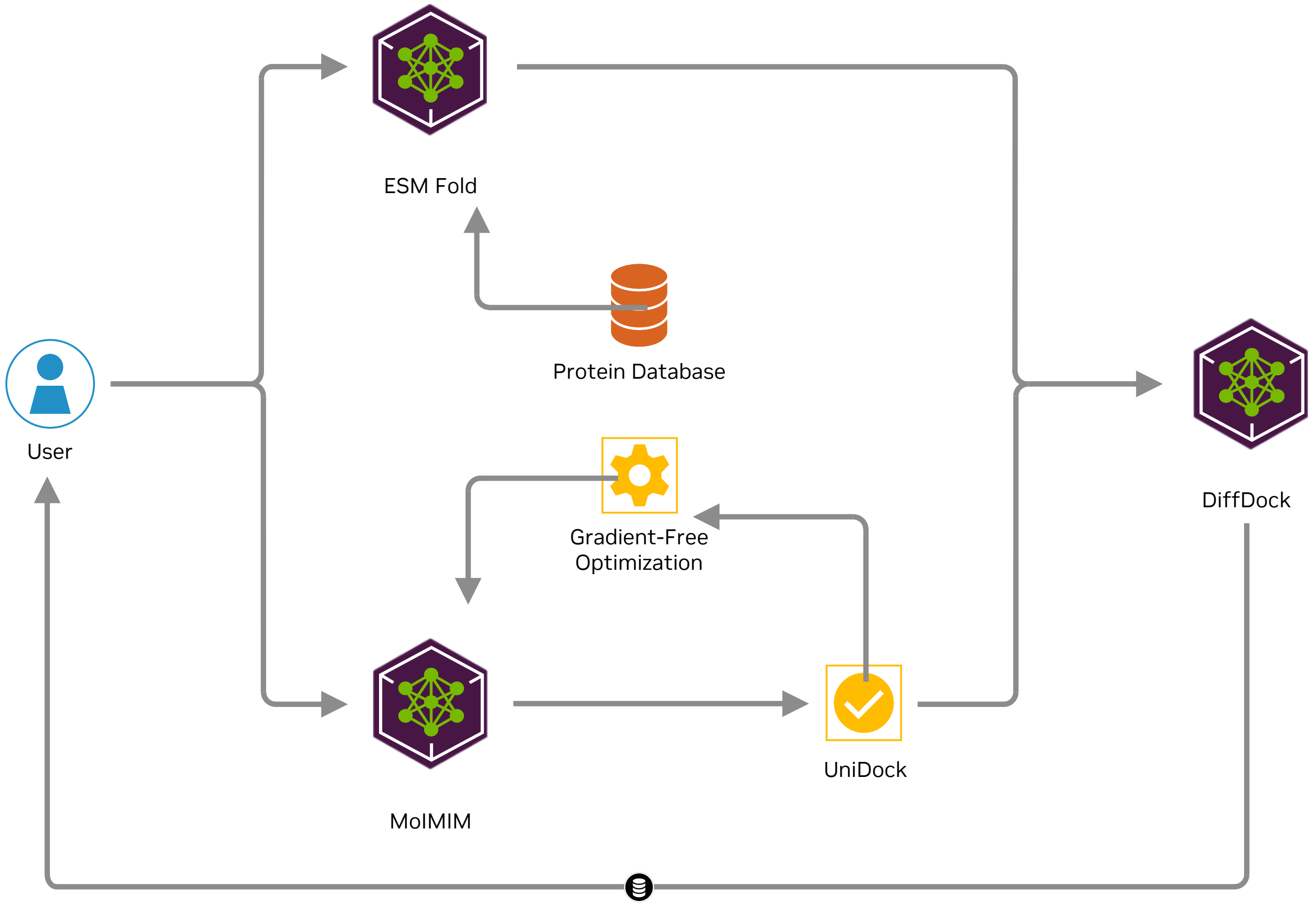
6x performance improvement in MLPerf HPC v3.0 Benchmark over baseline



- Training time to reach 0.9 lDDT-Cα
  - AlphaFold2: 7 days
  - 1056 H100s: 12.4 hrs
  - 2080 H100s: 10 hrs
- MLPerf HPC v3.0 benchmark results
  - OpenFold partial training task finished in 7.51 min, **6x** faster than baseline

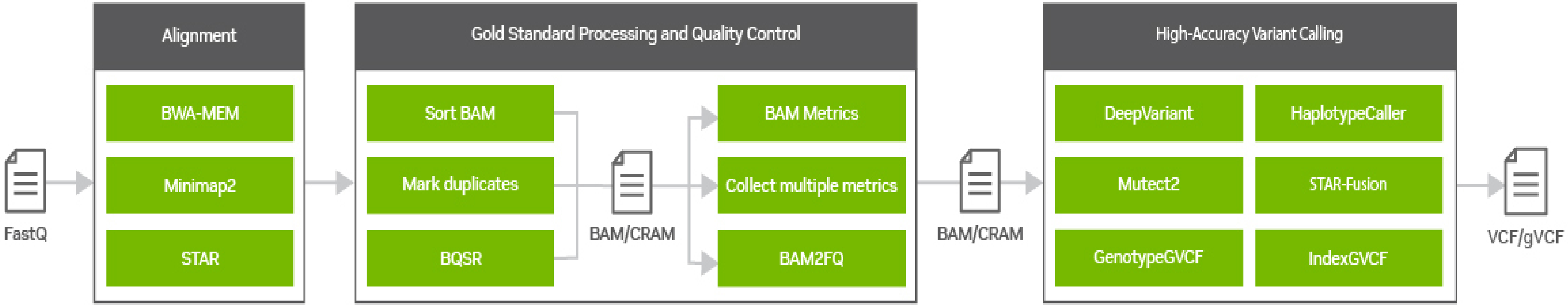
Build Generative AI Virtual Screening Workflows with NVIDIA NIM

Use composable NVIDIA NIMS to build workflows for CADD applications



NVIDIA Parabricks for Alignment & Variant Calling

Speed, Scale, Accuracy



Universal Analysis

Industry-standard tools for all major sequencers, ported to GPU

Up to 100x Acceleration

Up to 100x faster for WGS compared to CPU-only

Up to 50% Lower Cost

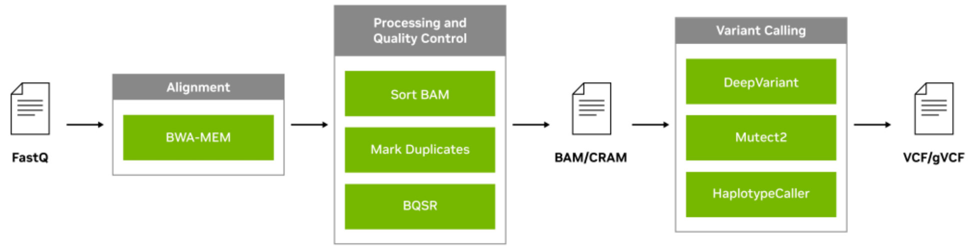
Up to 50% lower compute cost for WGS compared to CPU-only

Higher Accuracy with AI

The power of deep learning for customized high accuracy analysis

A Universal Analysis Solution

Short-Read



Element Biosciences

illumina

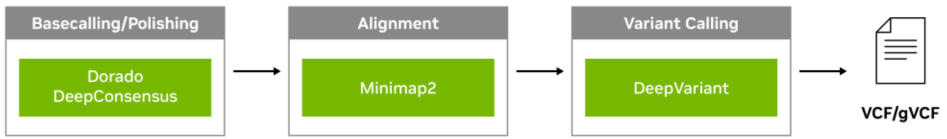
MGI

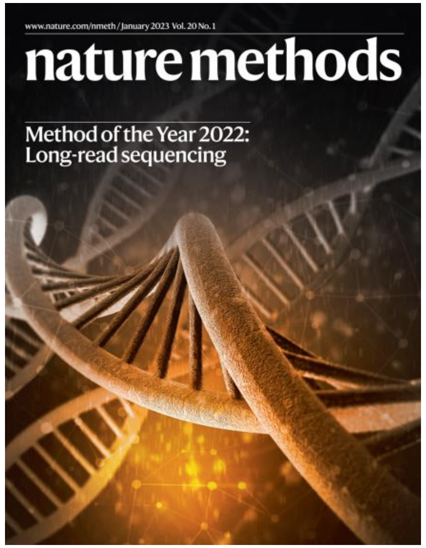
SINGULAR GENOMICS

ThermoFisher SCIENTIFIC

ULTIMA GENOMICS

Long-Read





www.nature.com/nmeth / January 2023 Vol. 20 No. 1

nature methods

Method of the Year 2022: Long-read sequencing

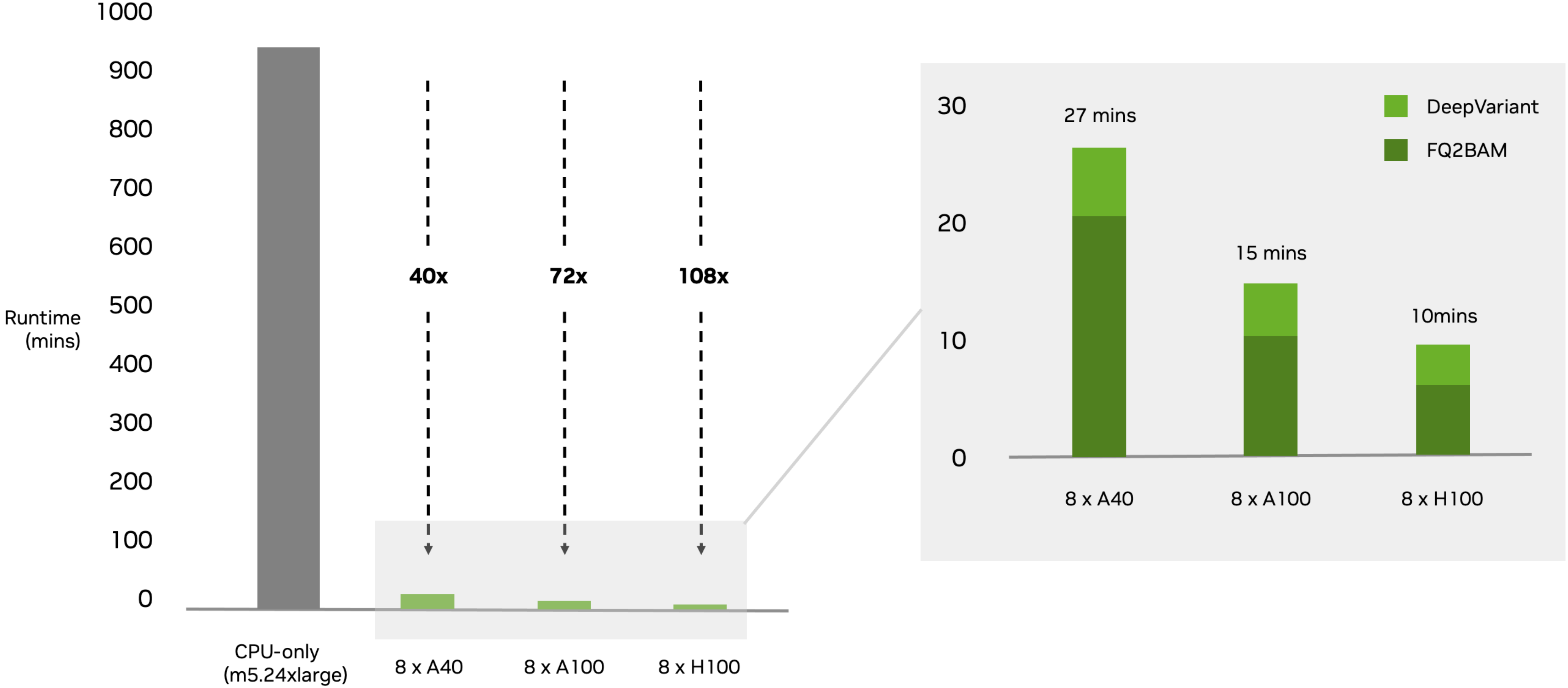
Oxford NANOPORE Technologies

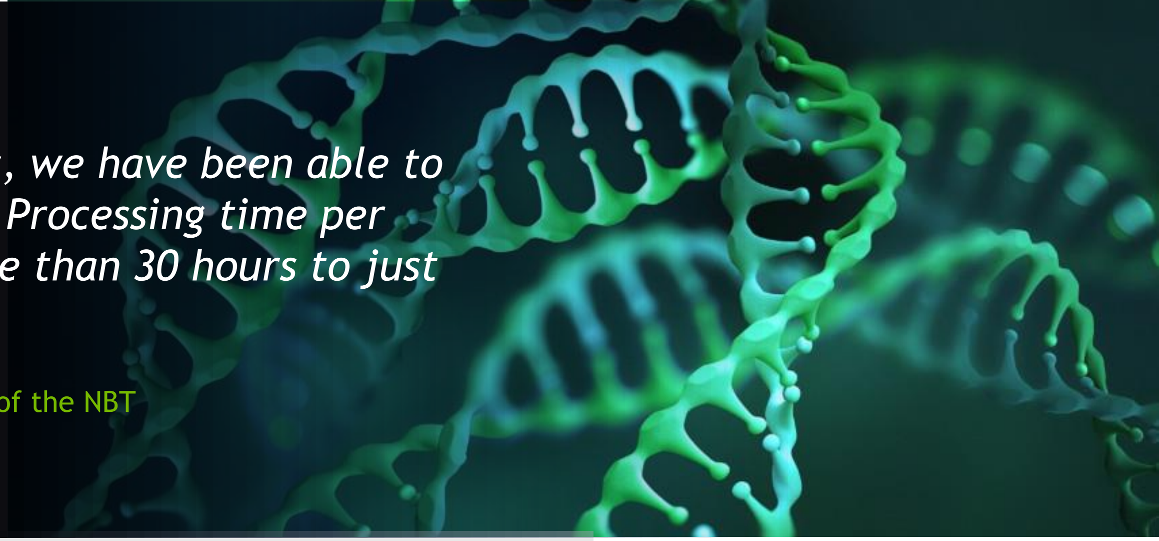
PacBio

Germline Analysis from 18 hours to 10 minutes

108x Acceleration using H100s Dynamic Programming Core

Germline workflow runtime per whole genome
(HG002, 30x Illumina)





> “ *By pairing NVIDIA DGX A100 with NVIDIA Parabricks, we have been able to reduce our WGS data processing by four months. Processing time per individual user has also been shortened from more than 30 hours to just one to two hours.* ”

- Sissades Tongsima, Director of the NBT

NATIONAL BIOBANK OF THAILAND ACCELERATES GENOMIC ANALYSIS BY 30X

Challenge

The National Biobank of Thailand (NBT) is the leading HPC facility and computational science R&D center in the ASEAN region.

Tasked with analyzing massive genome sequencing data from over **50,0000** individuals.

Their goal was to perform **whole genome sequencing (WGS)** to help accurately identify causative mutations and rare variants.

Solution

NBT leveraged **NVIDIA Parabricks** for genomic analysis on NVIDIA DGX A100, processing 5 PB of data in parallel with speed and accuracy.

The solution accelerated genomic analysis from 30 hours per individual to 1-2 hours.

NBT was able to reduce the whole genome sequencing (WGS) by 4 months, leading to faster genomic discoveries. NBT continues to use DGX A100 for their AI related projects.

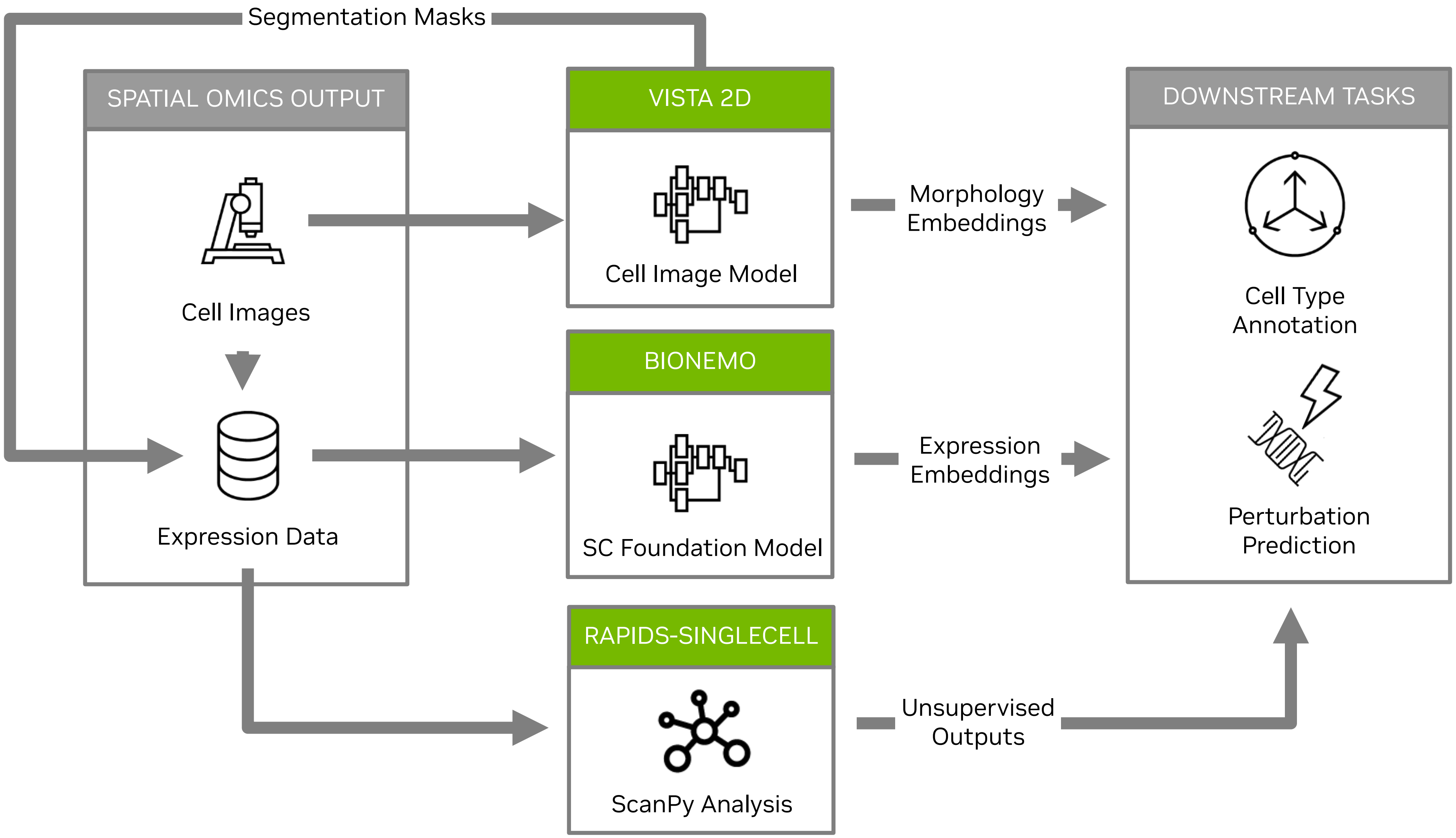
NVIDIA DGX A100
Unprecedented compute performance in the world’s first 5 petaFLOPS AI system

NVIDIA Parabricks
Computational genomic analysis framework supporting DNA & RNA

93%	Reduction in WGS data processing time per individual
30x	Faster genomic analysis vs CPU

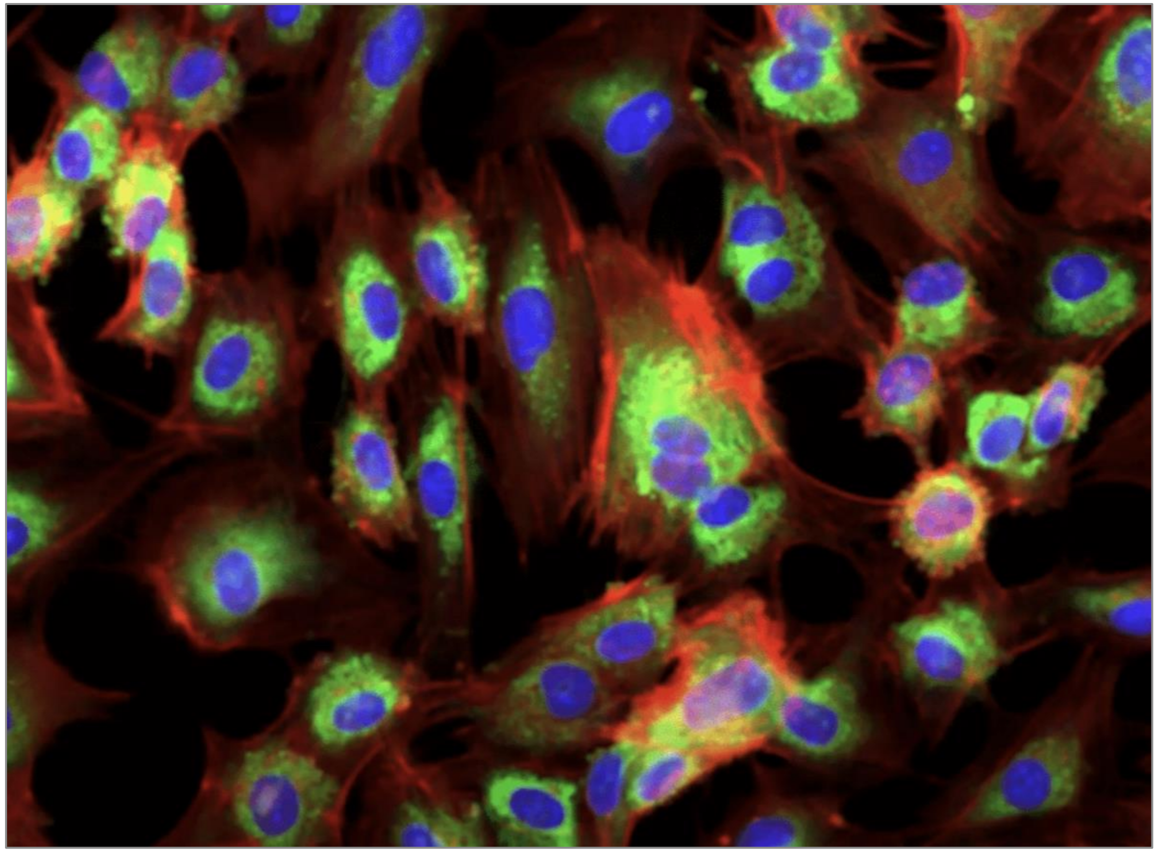
NEW: Single Cell & Spatial Omics Workflow

Build generative foundation models | Segment at high accuracy | Extract morphology embeddings



BioNeMo Microservices Activates Partner Ecosystem

Hosting Partners building models with BioNeMo and contributing as NVIDIA NIMS



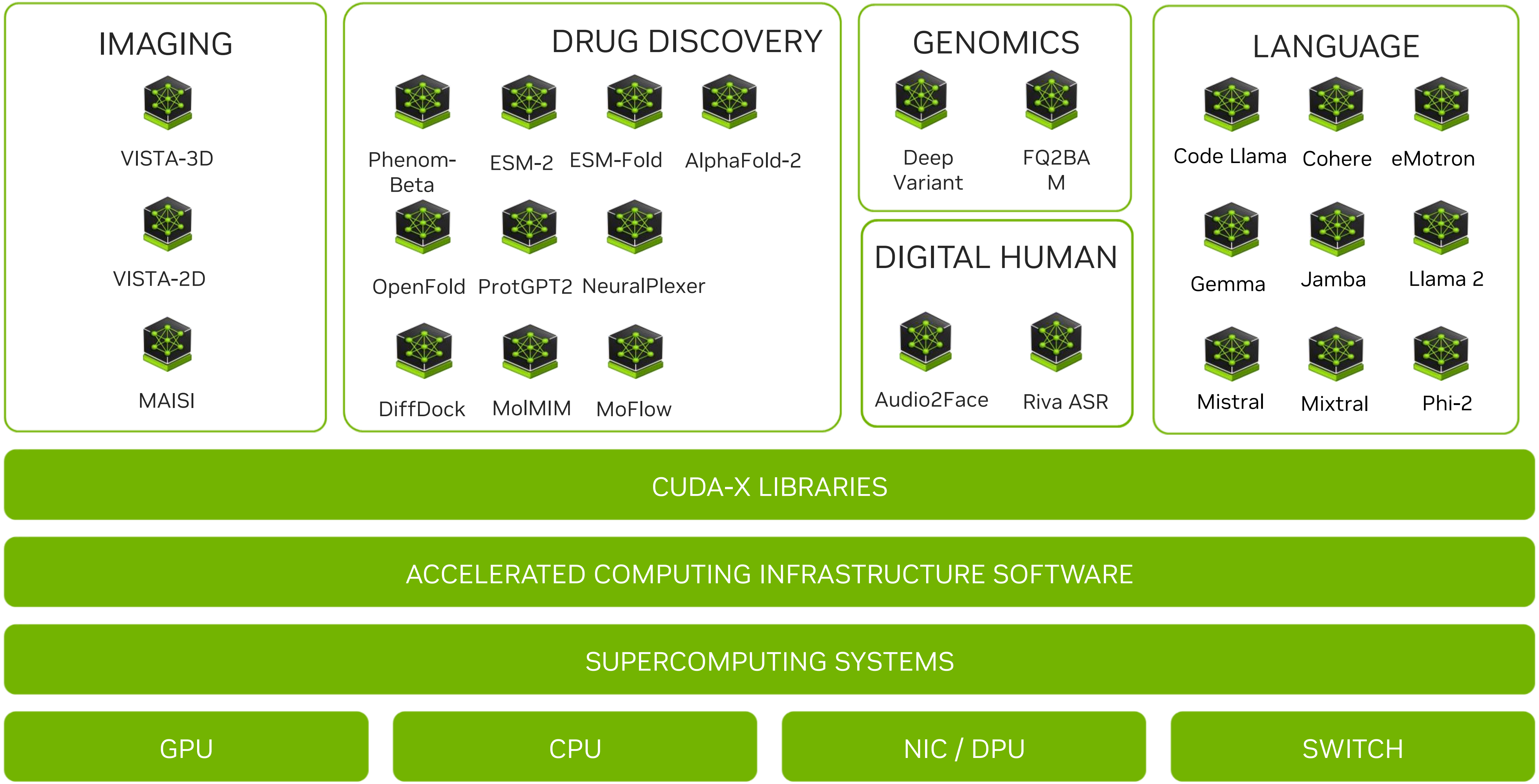
Recursion

Phenom-Beta model for cell morphology

- Phenom-Beta: First vision transformer model targeting cellular data
  - RxRx3 dataset: 17,063 CRISPR-KO genes, 2.2M HUVEC cell images, 1674 compounds, 8 dilutions

- BioHive-1 Supercomputer with NVIDIA DGX SuperPod reference architecture
  - 500 NVIDIA H100 TensorCore GPUs

NVIDIA NIM – A New Layer on the NVIDIA Clara Stack



IMAGING

- VISTA-3D
- VISTA-2D
- MAISI

DRUG DISCOVERY

- Phenom-Beta
- ESM-2
- ESM-Fold
- AlphaFold-2
- OpenFold
- ProtGPT2
- NeuralPlexer
- DiffDock
- MolMIM
- MoFlow

GENOMICS

- Deep Variant
- FQ2BAM

DIGITAL HUMAN

- Audio2Face
- Riva ASR

LANGUAGE

- Code Llama
- Cohere
- eMotron
- Gemma
- Jamba
- Llama 2
- Mistral
- Mixtral
- Phi-2

CUDA-X LIBRARIES

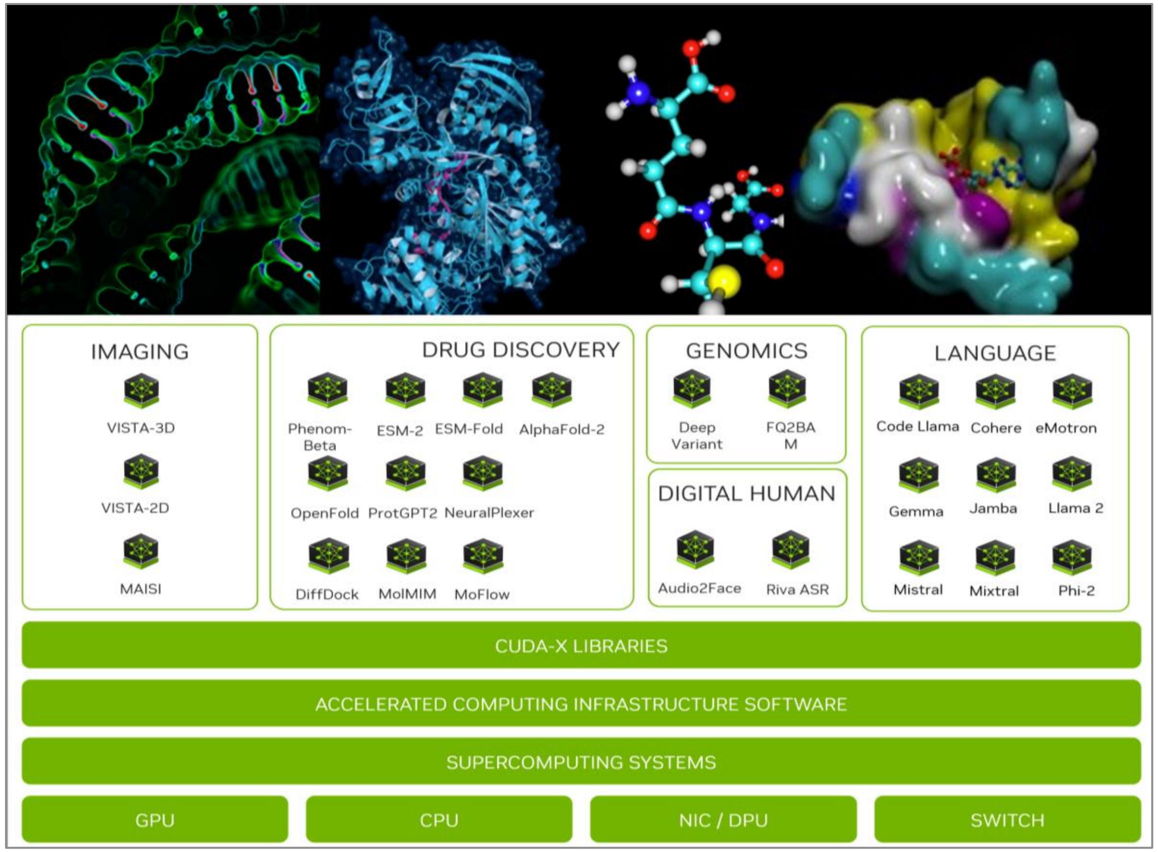
ACCELERATED COMPUTING INFRASTRUCTURE SOFTWARE

SUPERCOMPUTING SYSTEMS

GPU | CPU | NIC / DPU | SWITCH

Summary

Start accelerating your biomedical research with BioNeMo & Parabricks



IMAGING

VISTA-3D

VISTA-2D

MAISI

DRUG DISCOVERY

Phenom-Beta

ESM-2

ESM-Fold

AlphaFold-2

OpenFold

ProtGPT2

NeuralPlexer

DiffDock

MolMIM

MoFlow

GENOMICS

Deep Variant

FQ2BA M

DIGITAL HUMAN

Audio2Face

Riva ASR

LANGUAGE

Code Llama

Cohere

eMotron

Gemma

Jamba

Llama 2

Mistral

Mixtral

Phi-2

CUDA-X LIBRARIES

ACCELERATED COMPUTING INFRASTRUCTURE SOFTWARE

SUPERCOMPUTING SYSTEMS

GPU

CPU

NIC / DPU

SWITCH

- Technology is reshaping biomedical research
- BioNeMo provides a suite to tools for DNA, proteins, small molecules, and single-cell / spatial omics analysis
- Parabricks accelerates genome sequencing analysis to <10 min / WGS
- Join world-class leaders like Genentech, Amgen, National BioBank of Thailand in the AI accelerated biodiscovery journey!