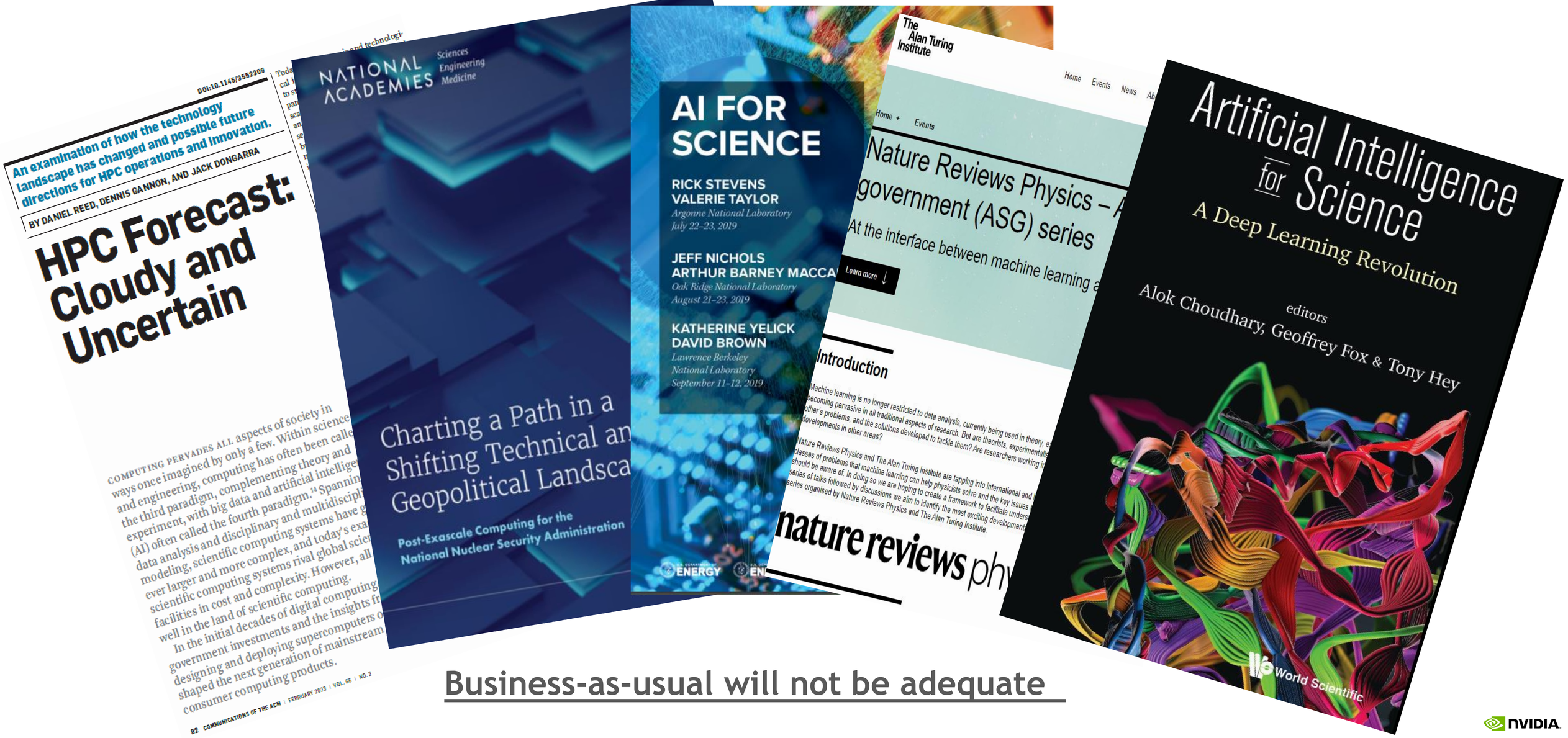
SCIENTIFIC COMPUTING IS EVOLVING

Thought Leaders Map Out the Opportunity and Constraints Given Current Technology and Market Reality

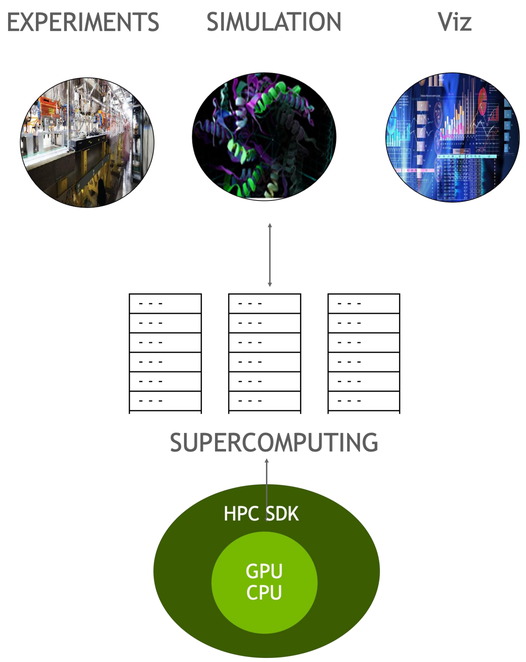


Business-as-usual will not be adequate

NVIDIA

SCIENTIFIC COMPUTING IS EVOLVING

ADAPTING TO THE CONSTRAINTS, MARKET REALITIES AND EXPANDING OPPORTUNITY



ACM HPC Forecast
Reed and Dongarra

Semiconductor Constraints limit the potential increase in scale for legacy algorithms

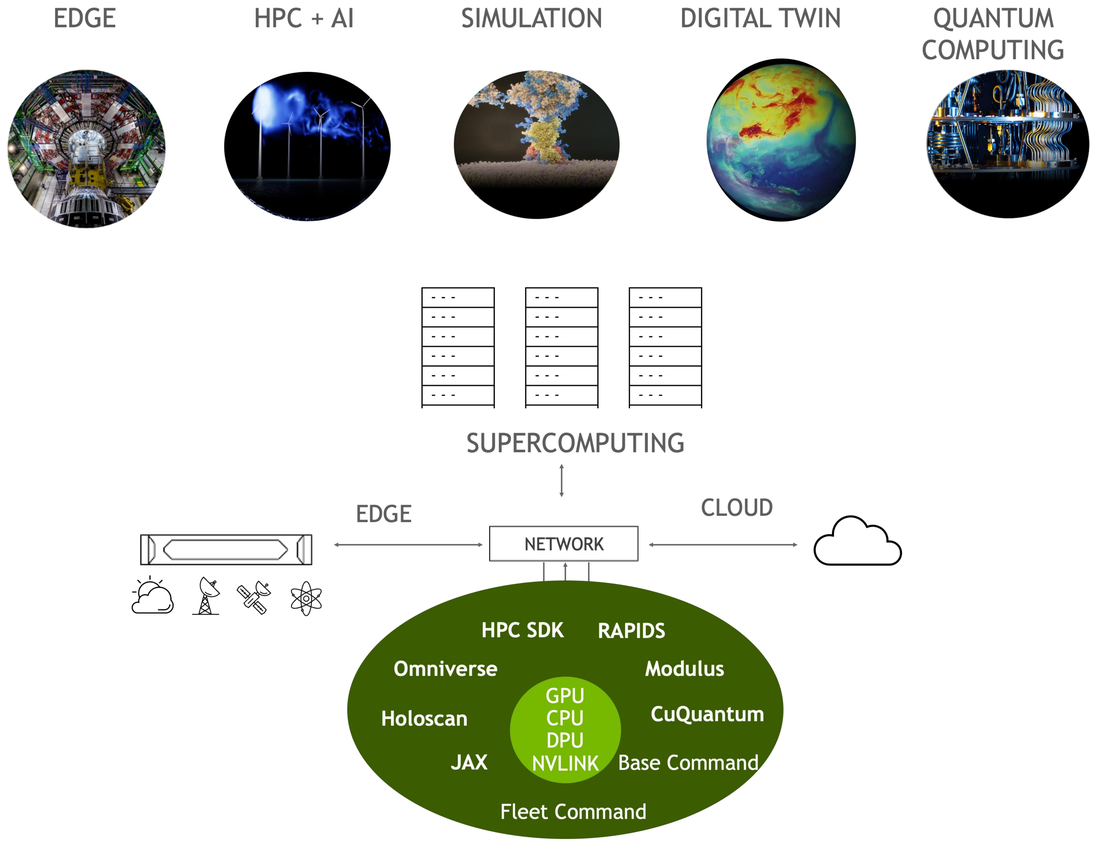
New Algorithms offer potential for dramatic increase in model scale and reduction in latency that is possible within the constraints

Cloud Economics have changed the supply chain ecosystem

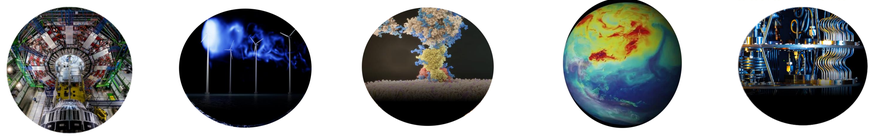
Charting a New Path Post
Exascale Computing NNSA

OVERARCHING FINDING: The combination of increasing demands for computing with the technology and market challenges in HPC requires an intentional and thorough reevaluation of …. algorithms, software development, system design, computing platform acquisition, and workforce development.

Business-as-usual will not be adequate

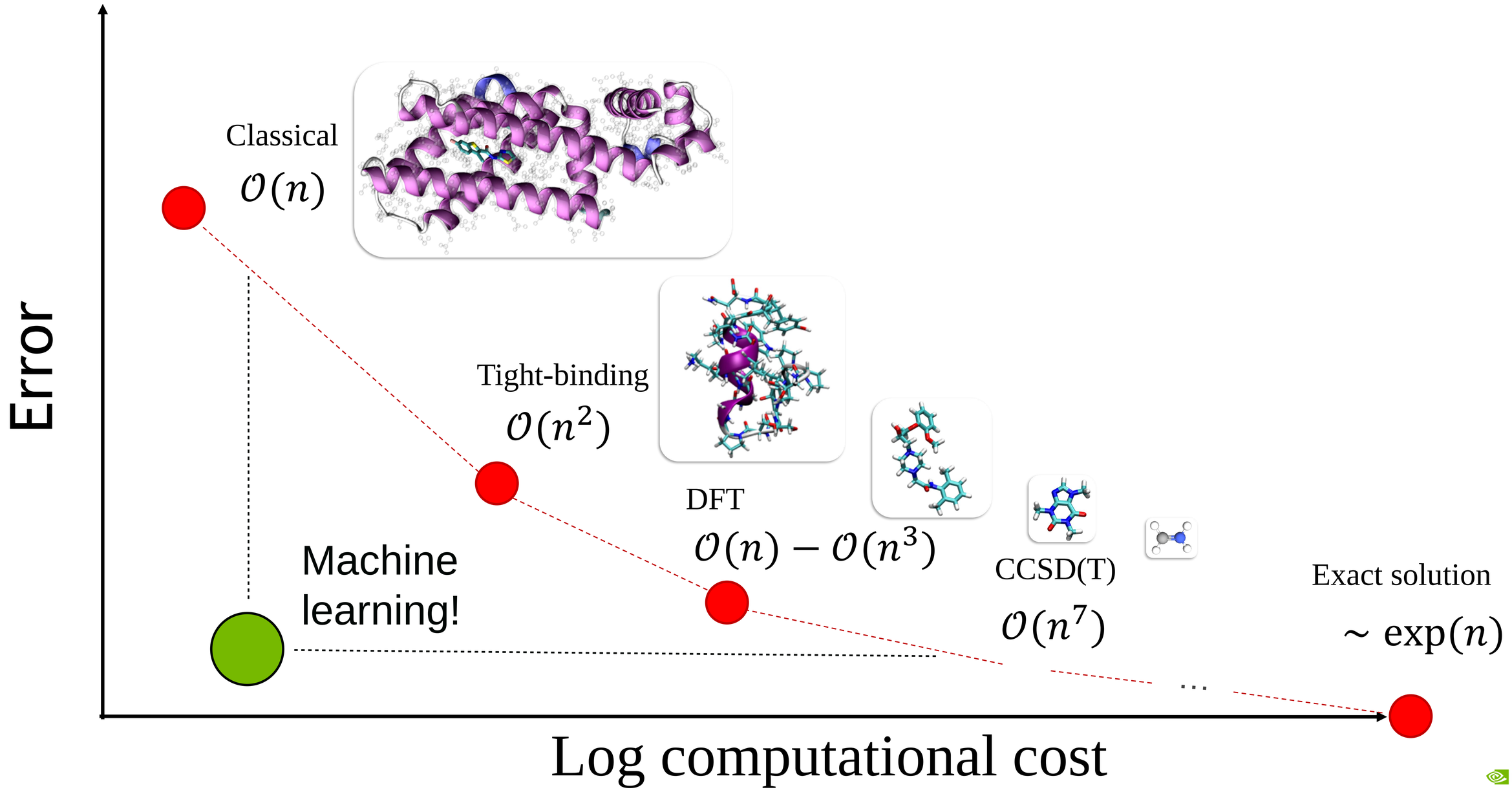


TRANSITION TO POST EXASCALE ERA

FEATURE	TERA THROUGH EXASCALE EXPERIMENTS SIMULATION Viz 	POST EXASCALE EDGE HPC + AI SIMULATION DIGITAL TWIN QUANTUM COMPUTING 
USAGE	BATCH & MOSTLY LOCAL TO A SITE	**INTERACTIVE & DISTRIBUTED WITH MULTIPLE SITES**
WORKLOAD	SINGLE SIMULATION/ENSEMBLE	**WORKFLOW COMPRISED OF SIMULATION ENSEMBLES, AI TRAINING AND INFERENCE, LIVE DATA ANALYTICS**
EXPERIMENTS	OFFLINE DATA ANALYSIS FOR EXPERIMENTS	**MIX OF REAL-TIME ANALYSIS, STEERING AND OFFLINE**
DIGITAL TWINS	**IN-SITU VISUALIZATION OFFLNE**	**INTERACTIVE COMBINATION OF SIMULATION AND OBSERVATIONAL DATA**
QUANTUM COMPUTING	SIMULATION	**SIMULATION PREPARING FOR A HYBRID MODEL**
PROGRAMMING MODELS	FORTRAN, C++, MPI, OPENMP	**STANDARD PARALLELISM SUPPORT IN FORTRAN, C++, MPI, OPENMP, OPENACC, PYTHON, JULIA, PYTORCH, TENSORFLOW**
SYSTEM CONFIGURATION	MONOLITHIC	**MODULAR**
CLOUD	GRID	**BURST CAPABILITIES, FASTER REFRESH CYCLE, ACCESS TO LATEST TECHNOLOGY AT SCALE**

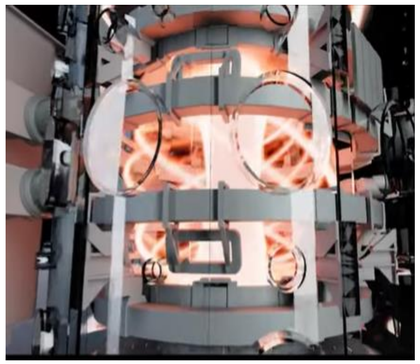
MACHINE LEARNING ENABLES DIGITAL TWINS FOR SCIENCE

Quantum Accuracy at Cost for Physical Scale Models



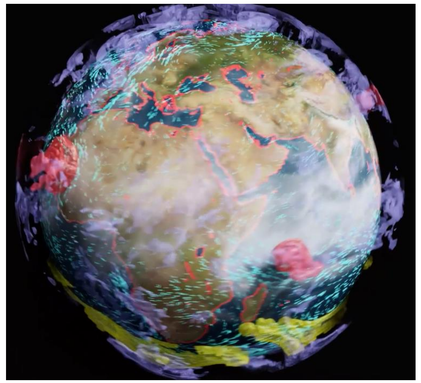
EXAMPLES OF DIGITAL TWINS FOR SCIENCE

Collaborate with the Global Research Community to Pursue Science Discovery that Benefits Mankind



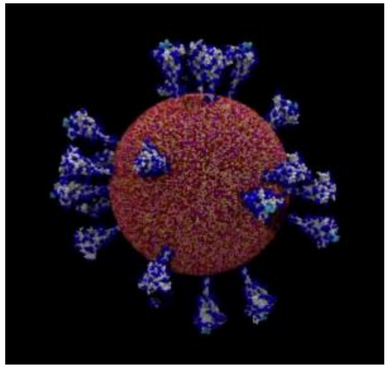
Towards Real time Fusion Reactor Design

Generative AI to Predict Disruption



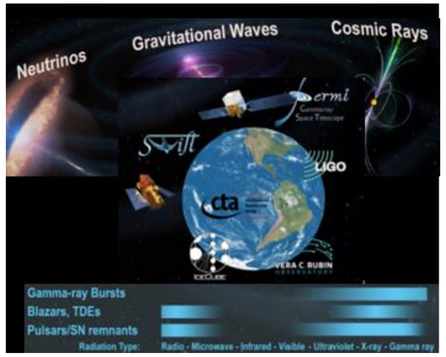
Destination Earth

AI for Good FourcastNEt



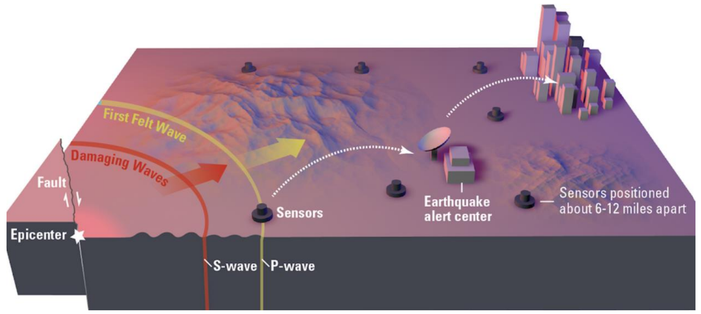
Genome Scale LLMs for Covid

Covid is Airborne



Real Time Multi-Messenger Astrophysics

Multi-Messenger Neutrino Detection



Earthquake Model with Machine Learning

Earthquake Early Warning SCEC

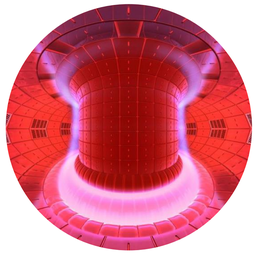
Kubota For Earth For Life

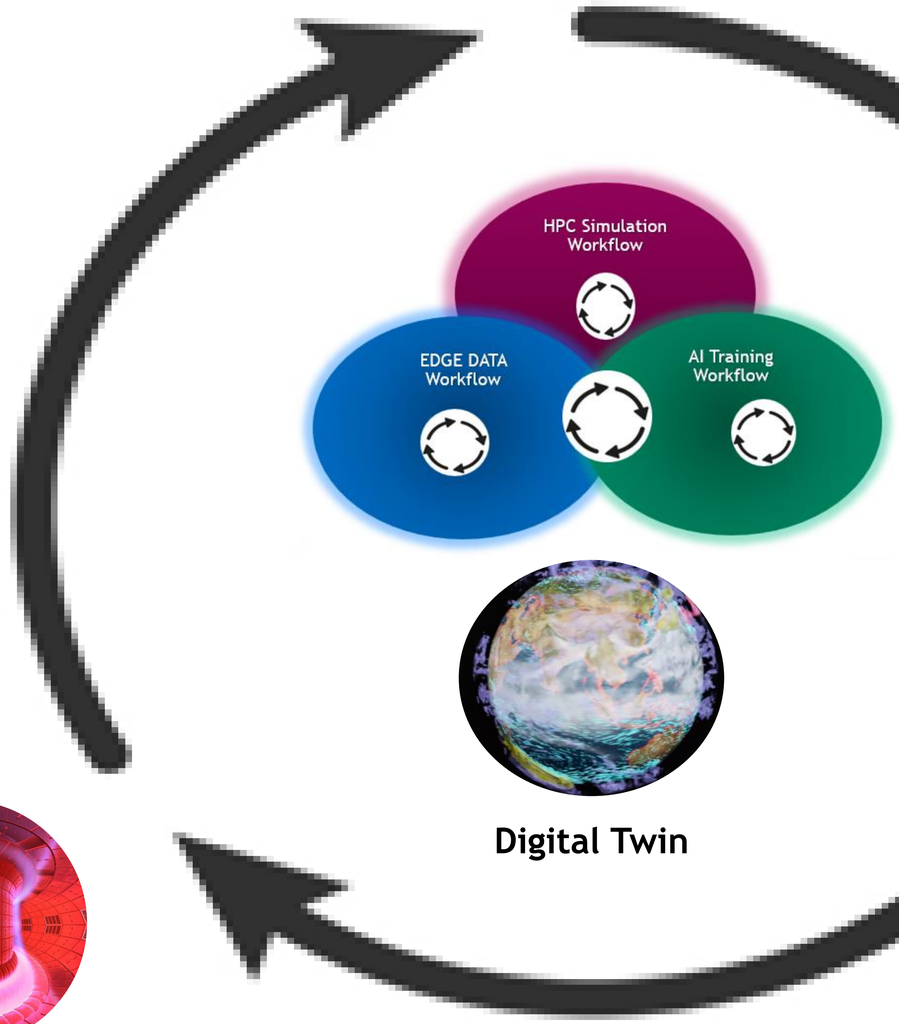
COMPOSITE WORKFLOWS EMERGE AS THE NEW APP

SIMULATION*AI*EDGE COMBINED IN A CONVERGED MODEL THAT CAN DRIVE A DIGITAL TWIN

Data Center

Sensors and Experiment
Data and Operation
At the Edge





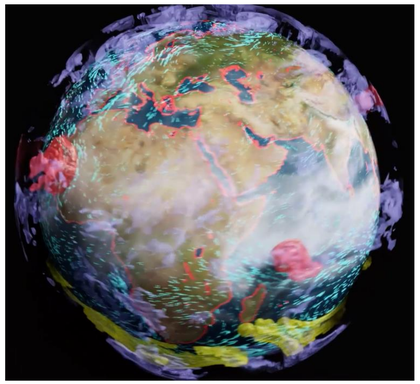
Digital Twin

Cloud

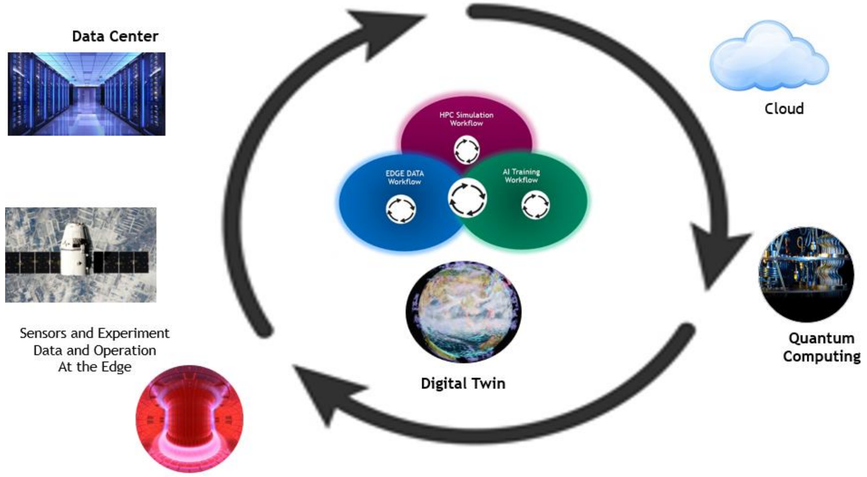
Quantum Computing

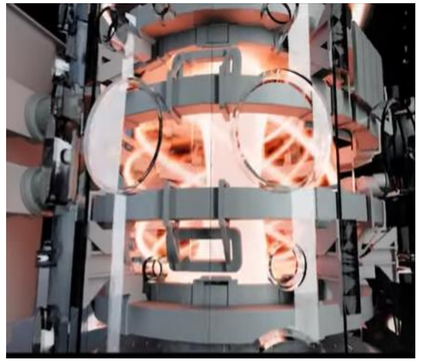
NVIDIA IS EXTENDONG OUR PLATFORM FOR SCIENCE

Collaborate with Early Innovators to Develop POCs and Extend/Enhance the Platform

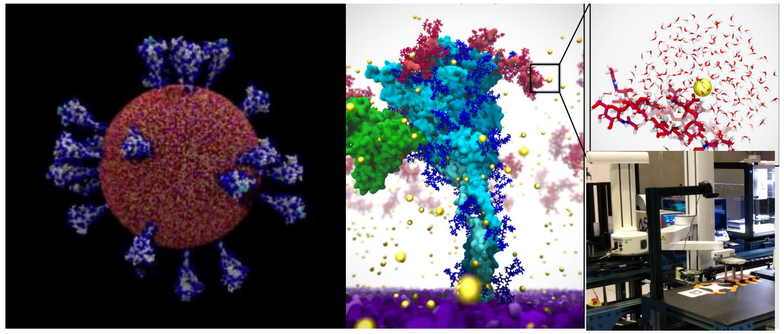


Earth 2 Digital Twin for Climate





Fusion Reactor Design and Control



Covid Models for Improved Drug Development

Platform Extensions

Hopper Transformer Engine improves training by 9x and Inference by 30x

RT cores enhance visualization for interactive digital twins and science apps that use similar algorithms

Base Tools: JAX, Python, Numpy optimized with Legate, TensorFlow, Pytorch, ISO C++ and Fortran

Holoscan: Make it easy to ingest data from an experiment or live instrument with low latency

Omniverse: Digital Twin developer kit that includes distributed data base, library of models and ray tracing tools

Modulus: Library of PINNs and Foundational Models to expedite model development

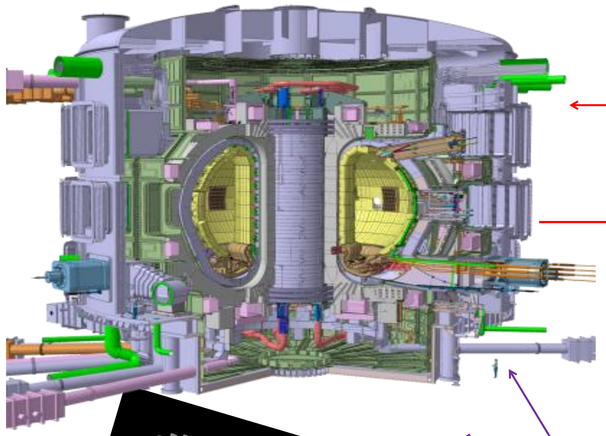
NeMo: Tool set to develop Generative AI at Scale

Enable leading workflow tool sets

Federated workflow could significantly accelerate fusion research

-- Hoping to shorten the fusion development timescale from months/years to days/weeks.

Thousands of sensors



Edge computing

In-stream federated feature detection & feedback

Secondary computer

Smart reduction, while preserving constraints, features & accuracy

Observables discovery

(near)-raw data

Deep steering & Validation, next day or week

Physics feedback for next shots

Newly trained inference codes & surrogates

WAN

Deep Query

Indexed hierarchical data

Smart-reduced data storage

Near real-time simulation trigger

Physics understanding

Reduced-model digital twin

Distributed scientists on many computers at different fidelity level.

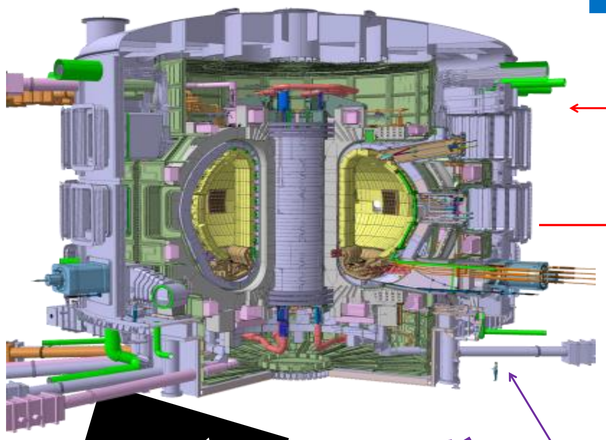
Real time simulation steering can also be triggered from near real-time experimental input

NVIDIA

Federated workflow could significantly accelerate fusion research

Enabling NVIDIA Tools

Thousands of sensors



Holoscan, RiverMax and DeepStream

IGX, EGX and Jetson

In-stream federated feature detection & feedback

Smart reduction, while preserving constraints, features & accuracy

Observables discovery

(near)-raw data

Physics feedback for next shots

Newly trained inference codes & surrogates

Deep steering, validation

WAN

Deep Query

Indexed hierarchical data

Smart-reduced data storage

Physics understanding

DGX, Grace Hopper, Grace and x86 Hosted Hopper Based Servers

Near real-time simulation trigger

Reduced-model digital twin

HPC SDK RAPIDS, JAX, Frameworks Omniverse and Modulus

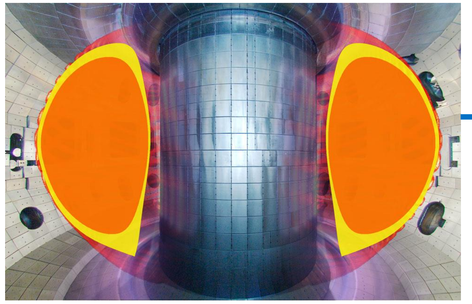
buted scientists on many computers

Rivermax, Magnum IO and DeepStream

tion steering can also be trigge

FUSION DIGITAL TWIN WORKFLOW

File-Based Prototype



DIII-D Fusion Experiment

Shot Data

Equilibrium Fit

Equilibrium Visualization

Gyrokinetic Tokamak Simulation

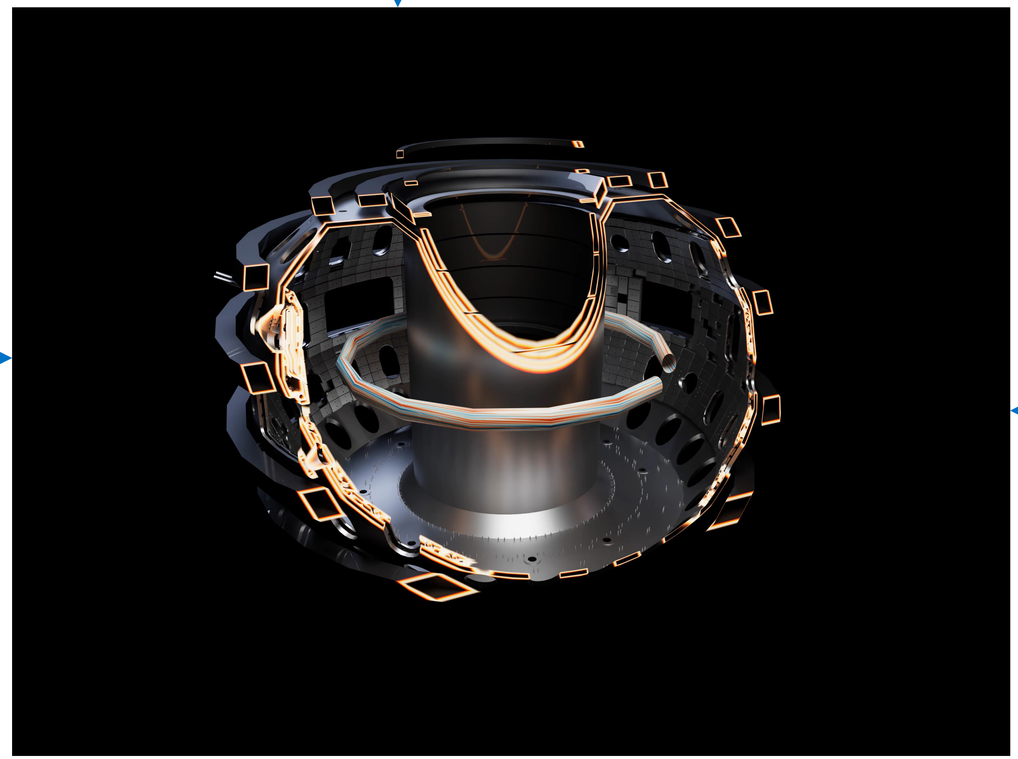
ML Training

Model

Surrogate Gyrokinetic Tokamak Simulation

CAD Model

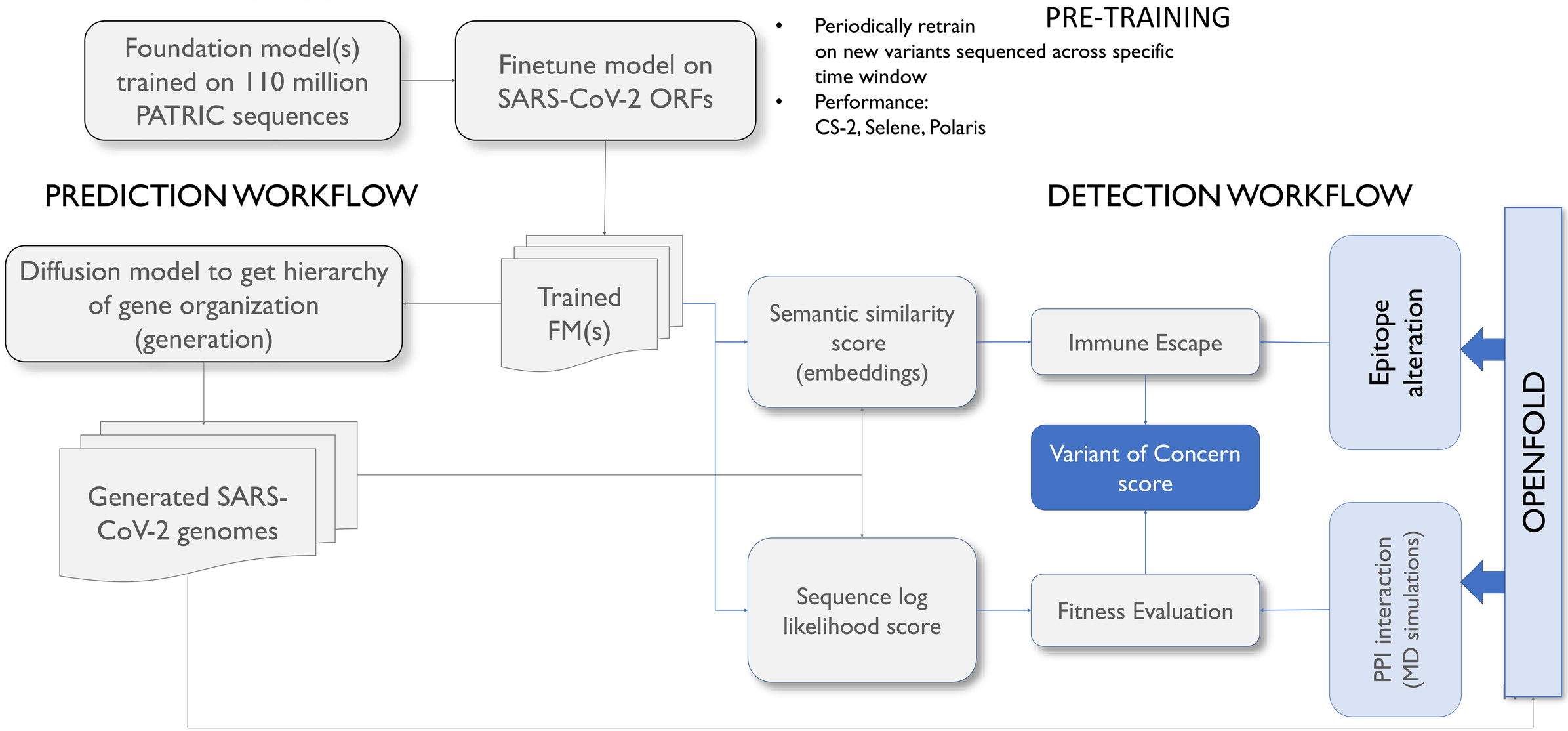
CAD Visualization



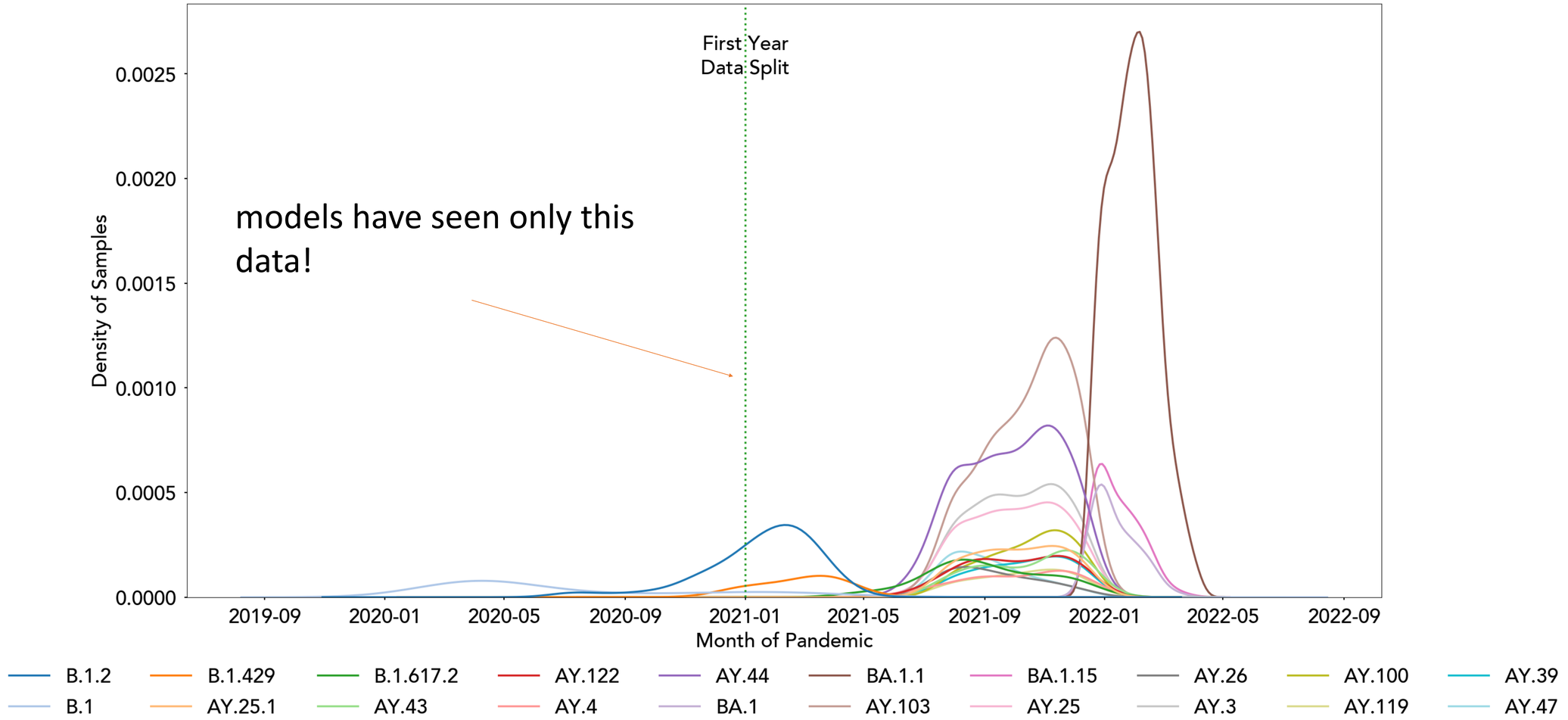
Gyrokinetic Visualization

Paraview Connector

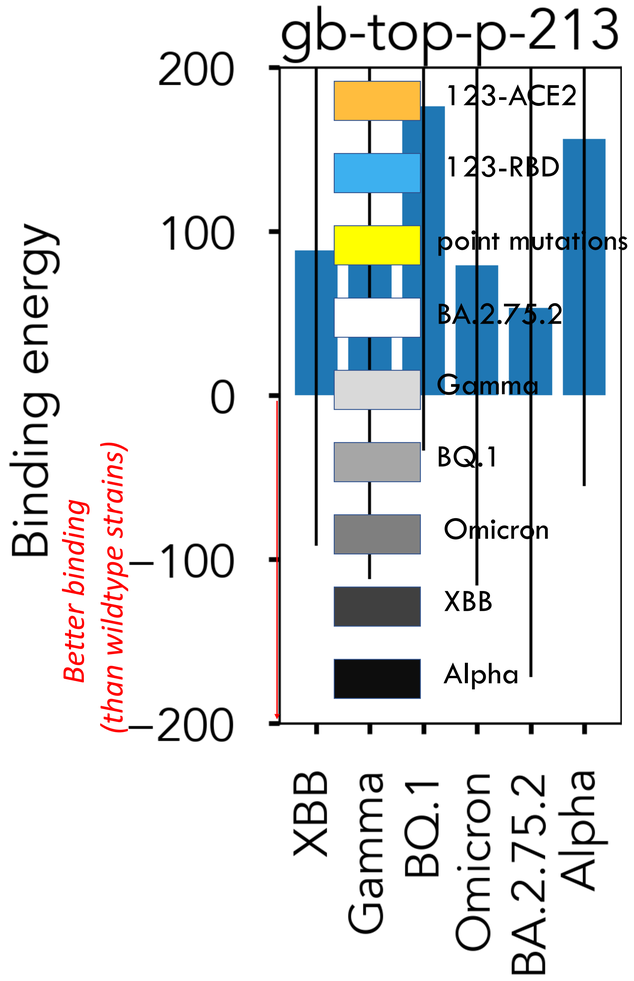
Using foundation models to predict SARS-CoV-2 evolution

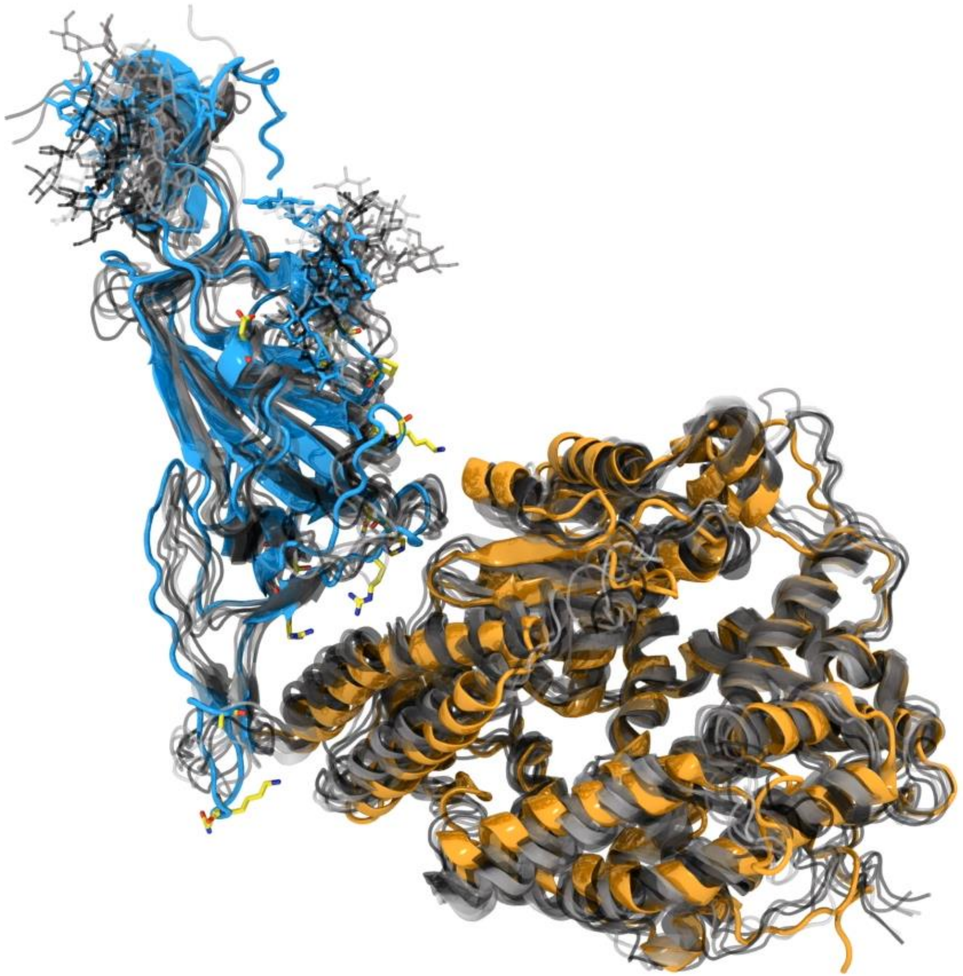


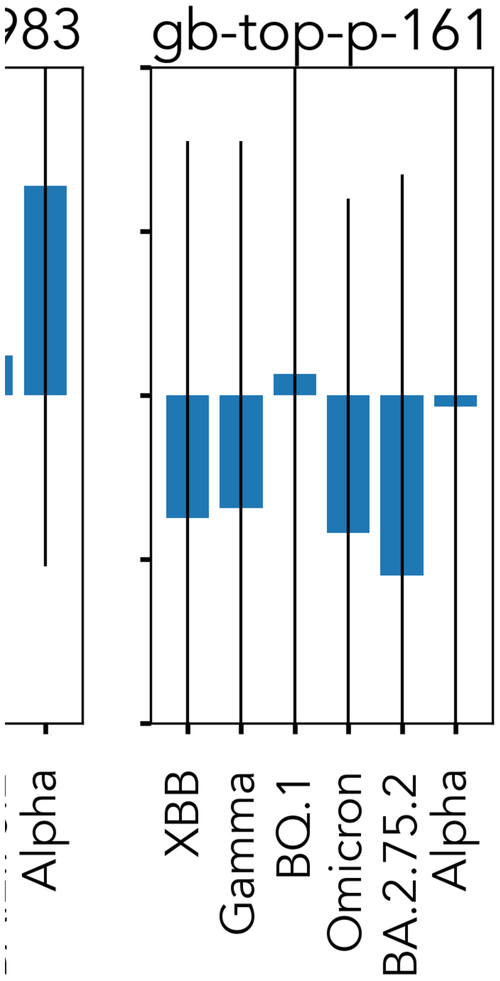
GenSLMs finetuned on SARS-CoV-2 genomes can distinguish variants



A generated variant is evolutionarily close to BQ.1!



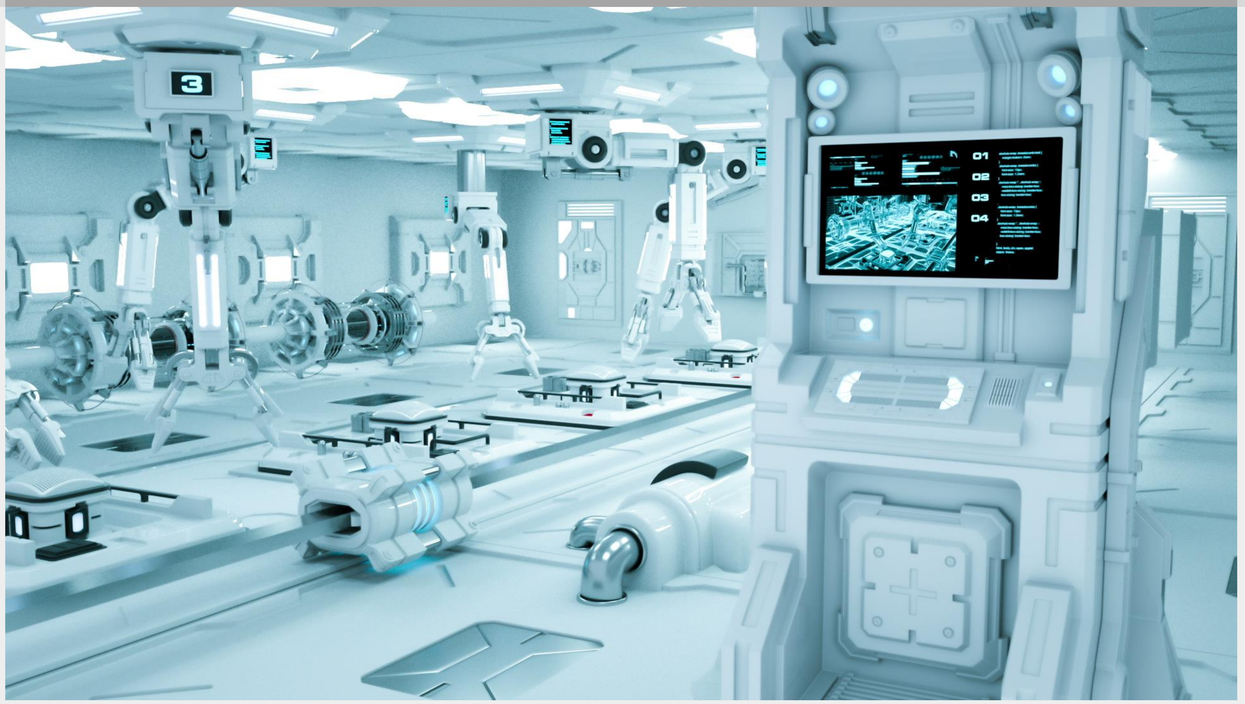


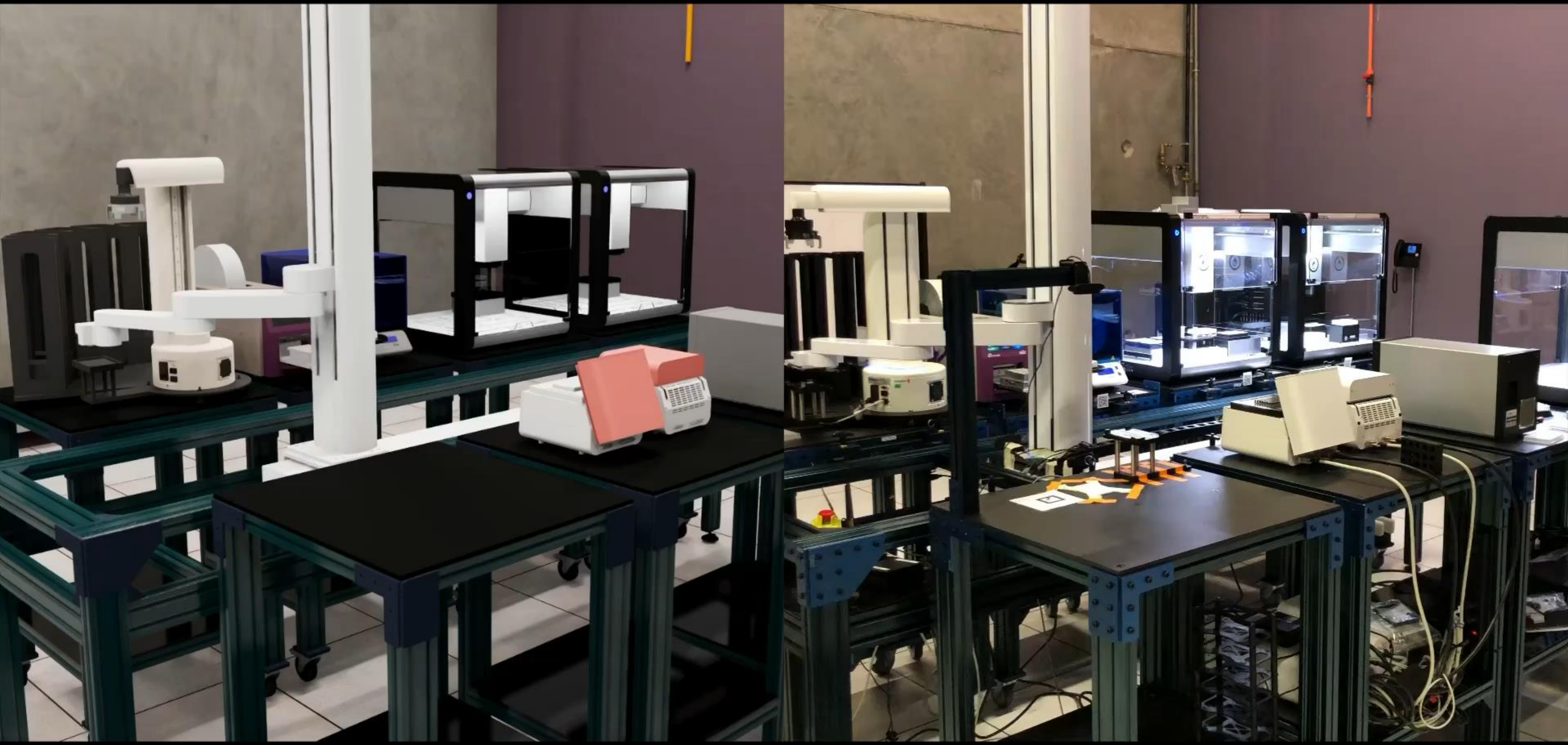


Our Vision: Smart “Factory” for Probing & Designing Complex Biological Systems

- Accelerate the discovery process
- Elevate human creativity to higher level goals
- Democratize biological systems design approaches
- Unbiased data collection and evaluation

ARTIFICIAL INTELLIGENCE
GUIDED, ROBOTICALLY EXECUTED EXPERIMENTS





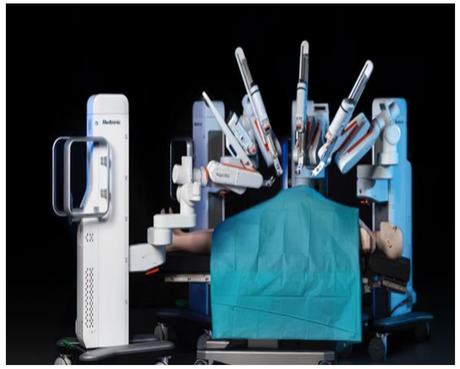
EXAMPLE OF DIGITAL TWIN FOR ASTROPHYSICS

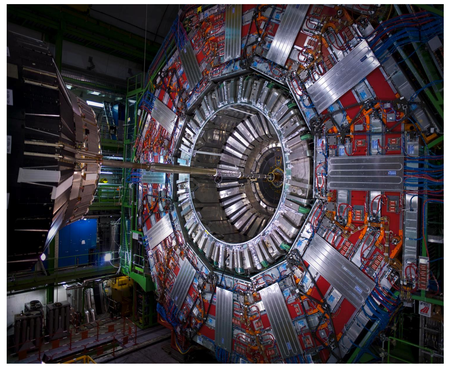
Moonwalker Digital Twin of the South Pole of The Moon

https://www.youtube.com/watch?v=E0Rz0ZbwhJY

Enabling Real Time, AI-Enabled Streaming Analytics at Any Scale

NVIDIA Holoscan

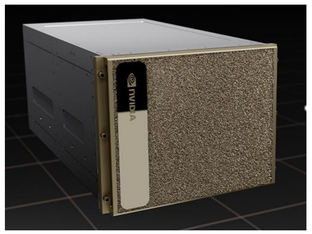




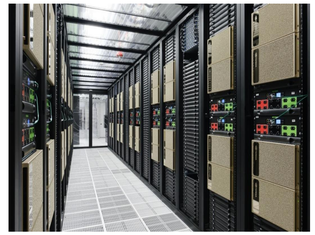
AGX Orin
Embedded

IGX Orin
Enterprise Edge

NVIDIA AI



MGX / DGX
HPC



Grace Hopper
Simulation

Sensor and Domain Agnostic

Software Defined

Low Latency, High Throughput

Scalable from Edge to Datacenter

Abandoned Apollo 17 lunar lander module is causing tremors on the moon

“Moonquakes”

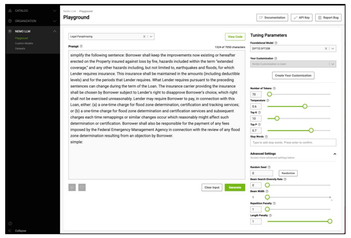
- The Apollo 17 lunar lander module left behind by US astronauts on the moon’s surface could be causing moonquakes, or small tremors, a new study revealed.
- Seismic Detection w/ NASA for the Hackathon went from 110 seconds of processing / AI inferencing to 4 seconds
- Follow up is to work with Holoscan to get the app real time.

Source: https://www.cnn.com/2023/09/14/world/moonquakes-apollo-17-scn/index.html

NVIDIA NeMo Service

Enterprise Hyper-Personalization and At-Scale Deployment of Intelligent Large Language Models

AI Development



GUI

API

NeMo Service

MODEL CUSTOMIZATION

P-tuning

Teach it Skills

Private Usage Data

RLHF

Improve from Interactions

INFERENCE

Guardrails

Remain in Operating Domain

Information Retrieval

Latest Business Knowledge

MODELS

GPT-8B

GPT-43B

GPT-530B

API

Enterprise Applications

Supply Chain Forecasting

Financial Modelling

Sales Pipeline Analysis

Legal Contract Discovery

NVIDIA DGX Cloud

Your Enterprise AI
Customize state-of-the-art pre-trained language models

Easily Develop & Connect Applications
GUI-based Playground and Scalable Cloud API

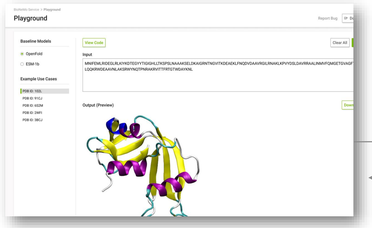
Deploy Anywhere
In the Service, Across Public Clouds, or On-Premises

Enterprise Support
Fully supported by NVIDIA AI Experts from Customization to Deployment At-Scale

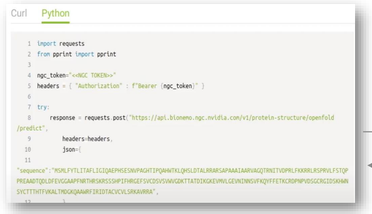
BioNeMo Service

For Generative AI in Biology Enterprise-Scale Training & Deployment of Biomolecular AI Models

BioNeMo



WEB



API

Drug Discovery Applications

- Protein Structure Prediction
- Biomolecule Representation
- Biomolecule Generation
- Molecular Docking

BioNeMo Models

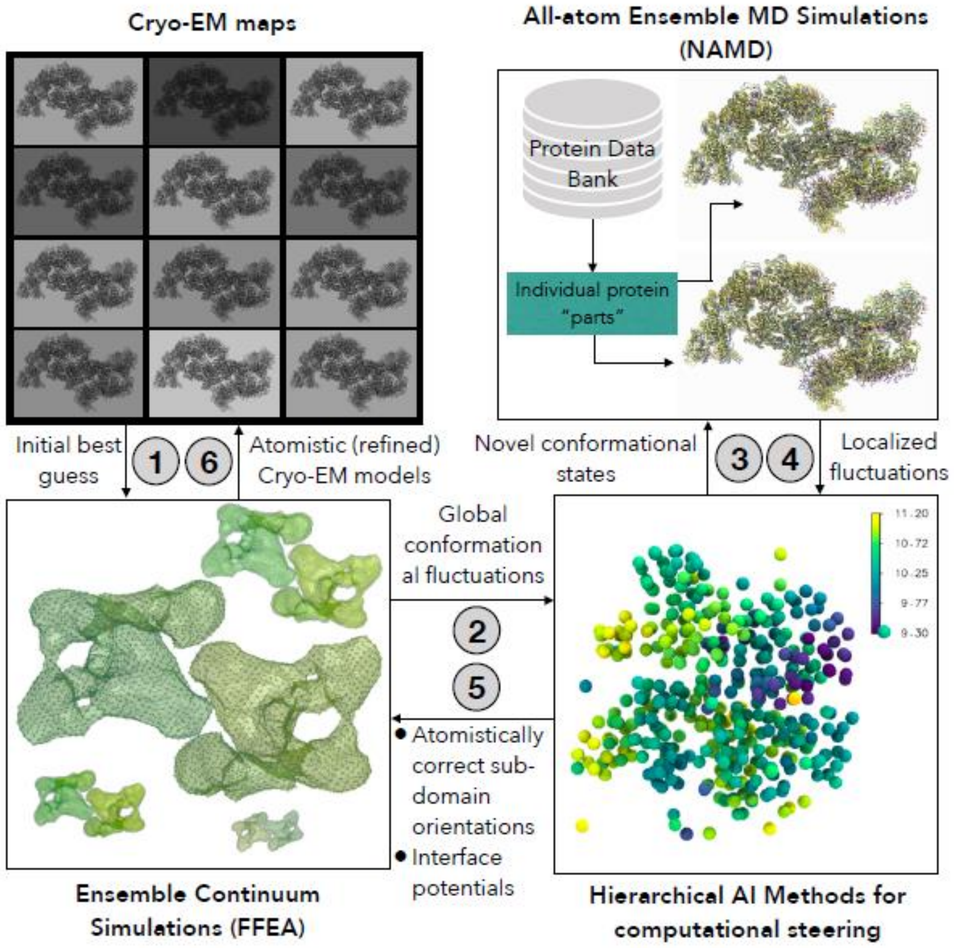
- AlphaFold2
- OpenFold
- ESMFold
- ESM1
- ESM2
- ProtGPT2
- MegaMolBART
- MoFlow
- DiffDock

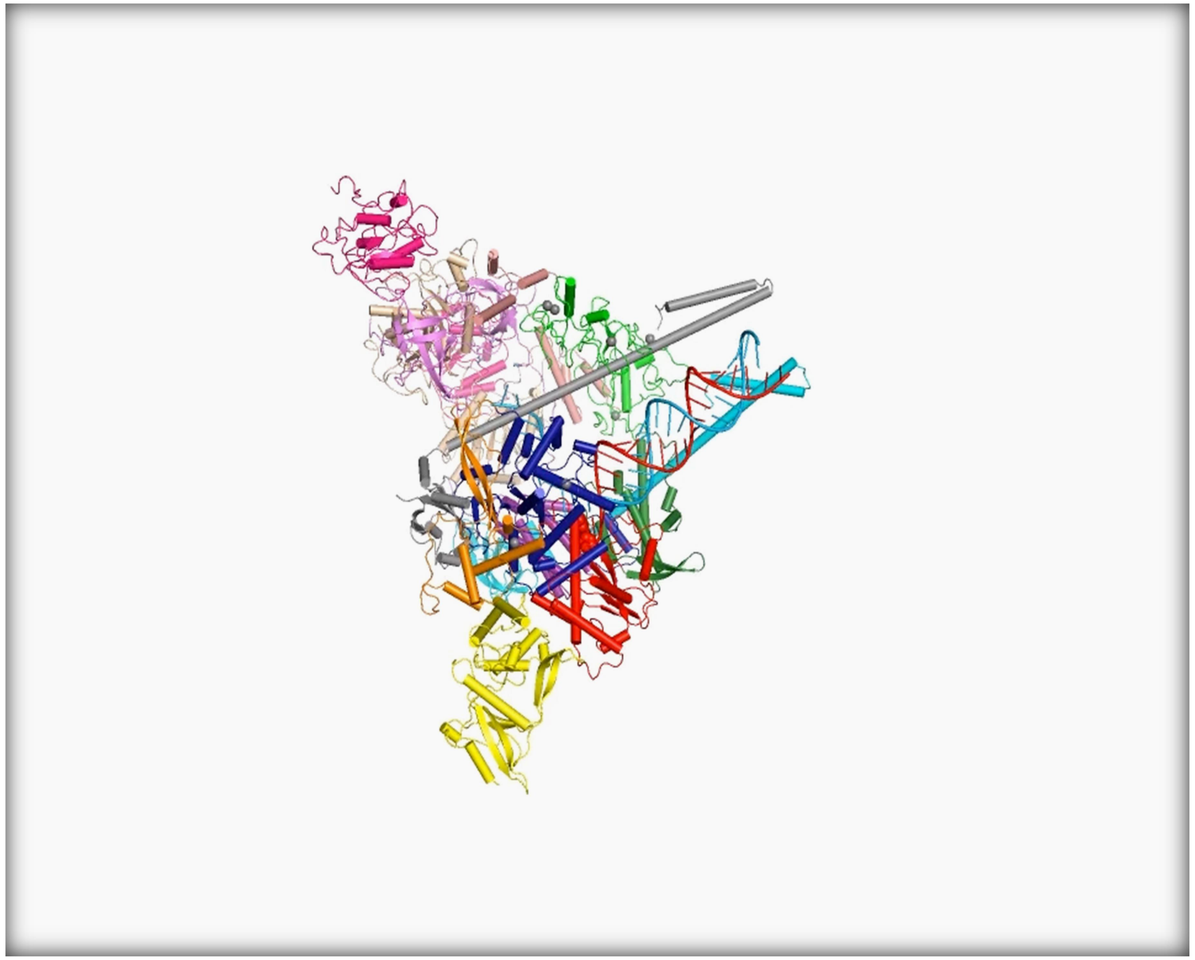
NVIDIA DGX Cloud

Training, fine-tuning, inferencing and deploying

NVIDIA

SCIENCE GOAL: UNDERSTAND HOW TO “TRAP” THE SARS-COV-2 REPLICATION-TRANSCRIPTION COMPLEX (RTC) USING LOW RES CRYOEM



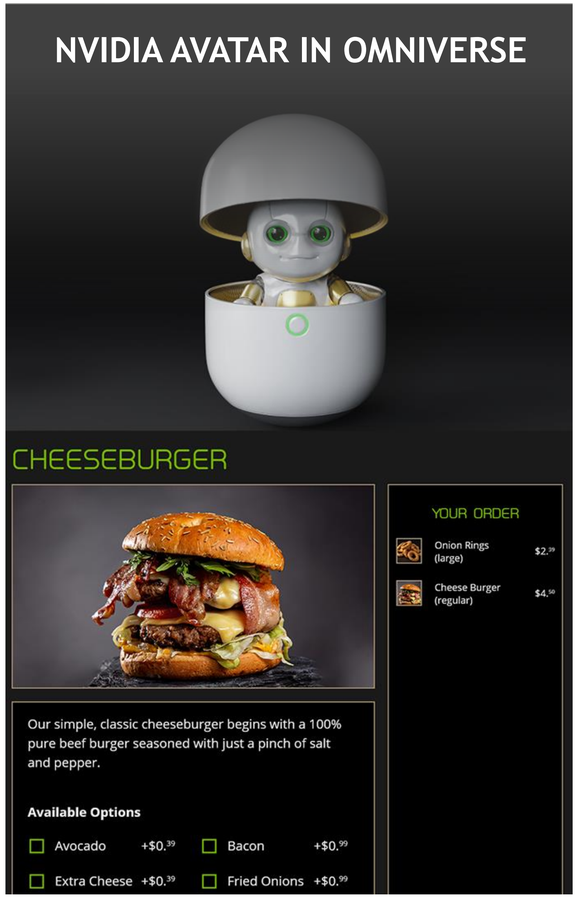


NVIDIA BUILT OMNIVERSE TO ENABLE VIRTUAL WORLDS

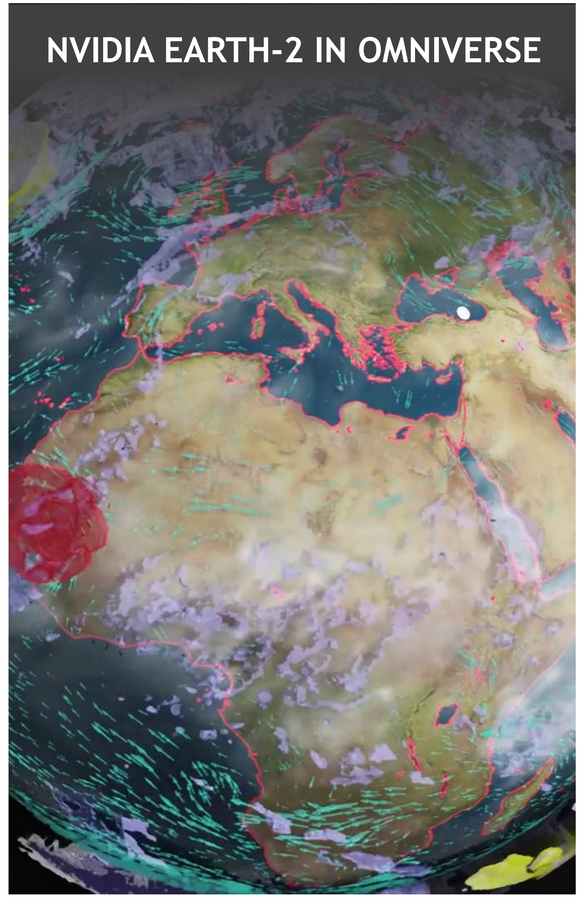
NVIDIA DRIVE IN OMNIVERSE

NVIDIA ISAAC IN OMNIVERSE

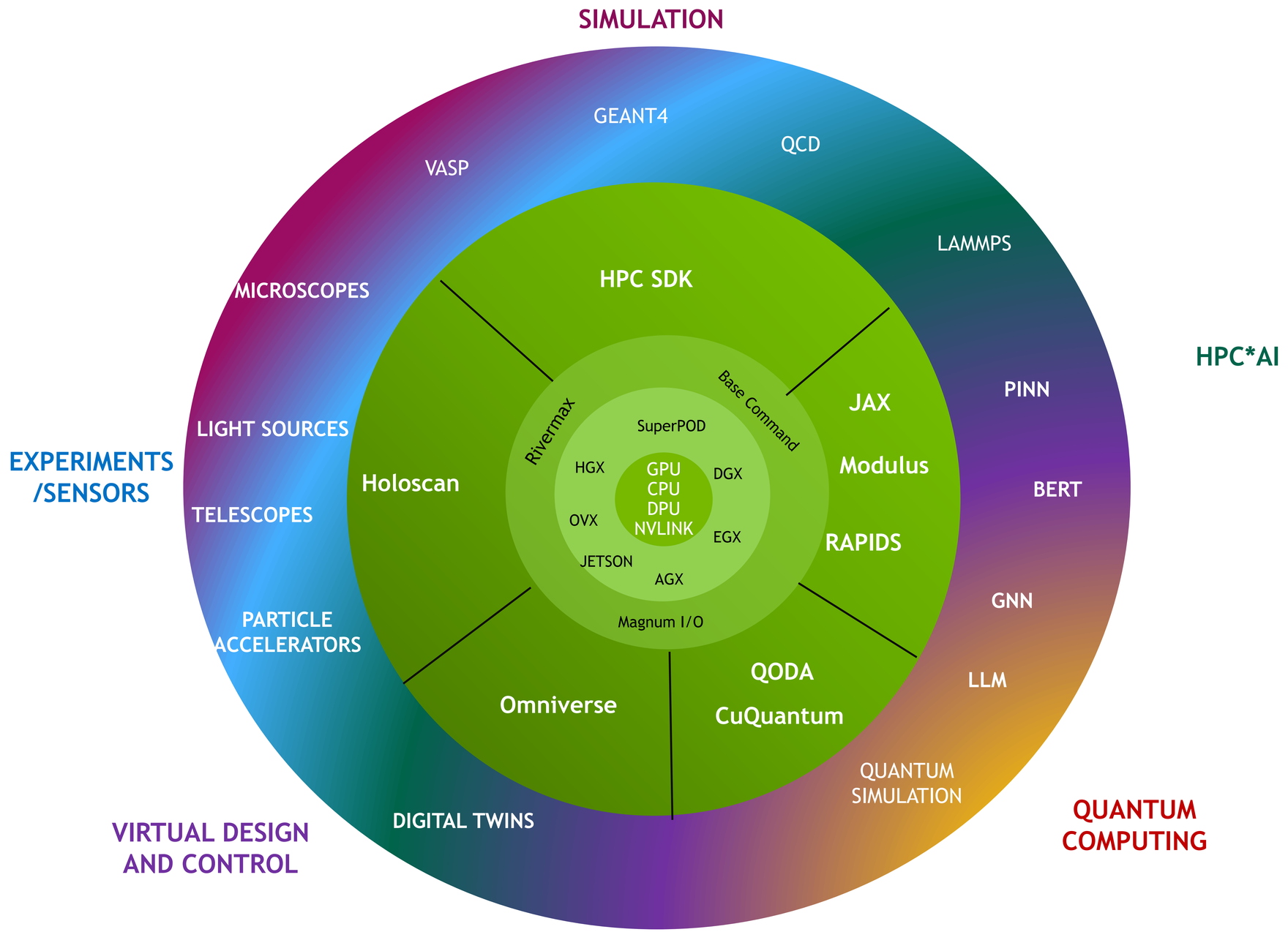
NVIDIA AVATAR IN OMNIVERSE

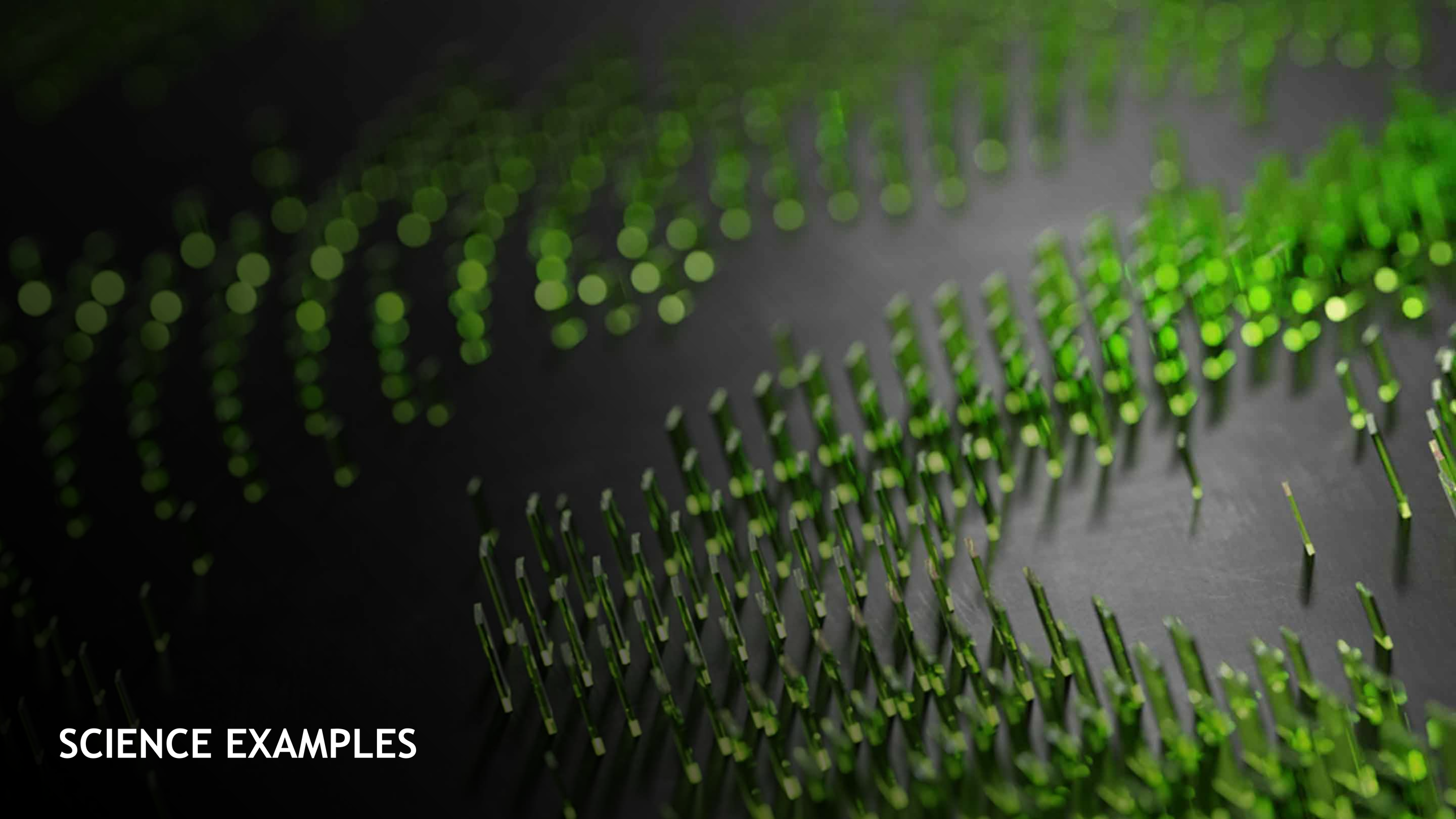


NVIDIA EARTH-2 IN OMNIVERSE



ROADMAP EVOLVING TO MEET THE CHALLENGE





SCIENCE EXAMPLES

NVIDIA LLM Software Ecosystem Offerings and Choices

NVIDIA's LLM Ecosystem is deep & wide, offering choices for researchers across the entire stack

Offerings

Domain Specific / LLM Application Systems

Model Customization, Evaluation, Safety & Explainability

Model Architecture & Techniques

Systems Optimization

LLM Systems & Applications
Ecosystem: LangChain
NVIDIA: NeMo Guardrails

LLM Services
Ecosystem: OpenAI APIs, Cohere, AI21, AWS Bedrock
NVIDIA: NeMo Service

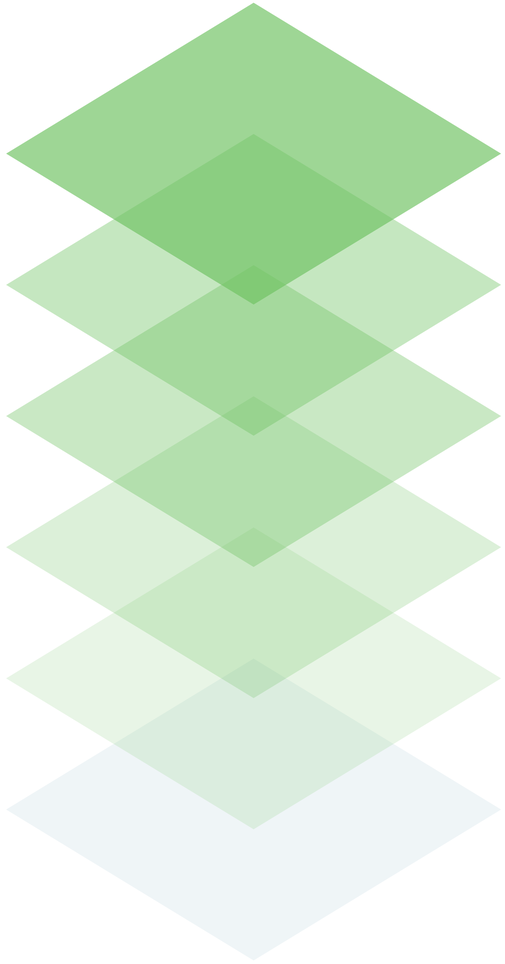
LLM Models
Ecosystem: LLaMa, MPT, PaLM, OPT, BLOOM
NVIDIA: NeMo Models

SDKs & Frameworks
Ecosystem: PyTorch, Colossal-AI, HuggingFace Transformers, PaxML, JAX
NVIDIA: NeMo Framework, Megatron-LM

Libraries
Ecosystem: XLA
NVIDIA: CUDA, NCCL, TensorRT, RAFT, Transformer Engine, CUTLASS

Management & Orchestration
Ecosystem: Kubernetes, Slurm, Nephele,
NVIDIA: Base Command Platform

Accelerated Infrastructure
NVIDIA: GPUs, CPUs, InfiniBand

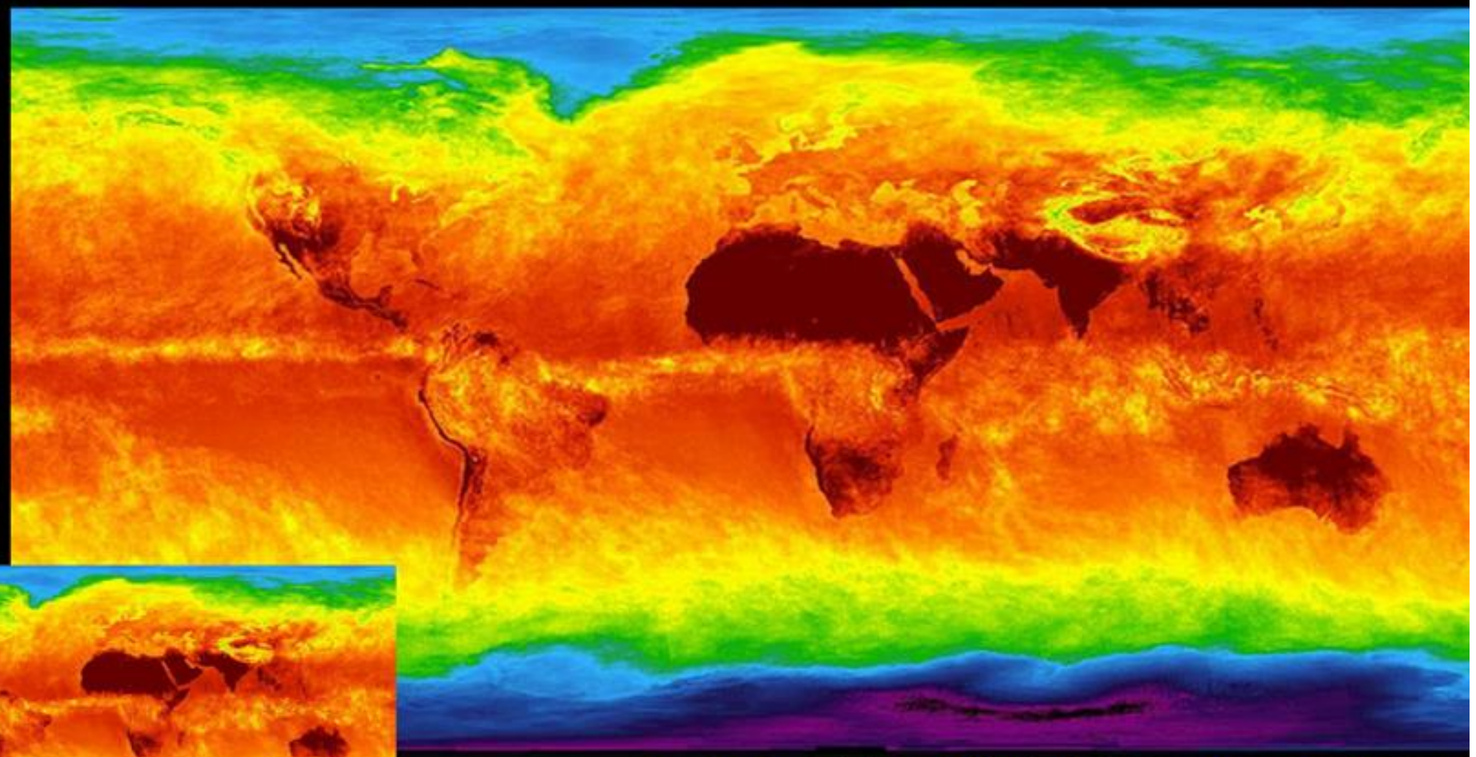


ALGORITHMS EVOLVING AT UNPRECEDENTED PACE

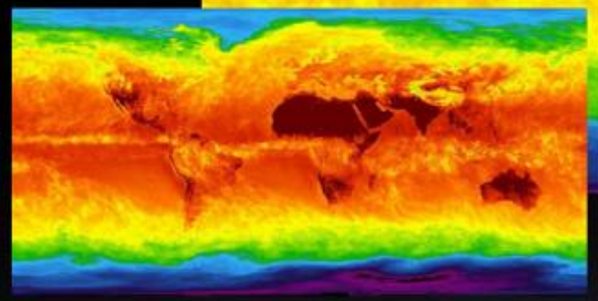
FourCastNet High Resolution for Data-Driven Weather Models

Comparison of resolutions for data-driven weather models since 2018 (Dueben & Bauer)

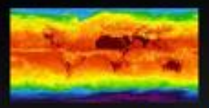
SOTA evolving rapidly
Recent Pre-print Kang Chen et al (2023) extend forecast to 10 days with 0.25° resolution using "cross modal Transformer"



FourCastNet, Pathak et al. (2022), 0.25°, ~1,000,000 Pixels, ViT+AFNO



GNN, Keisler et al. (2022), 1°, 64,000 Pixels, Graph Neural Networks



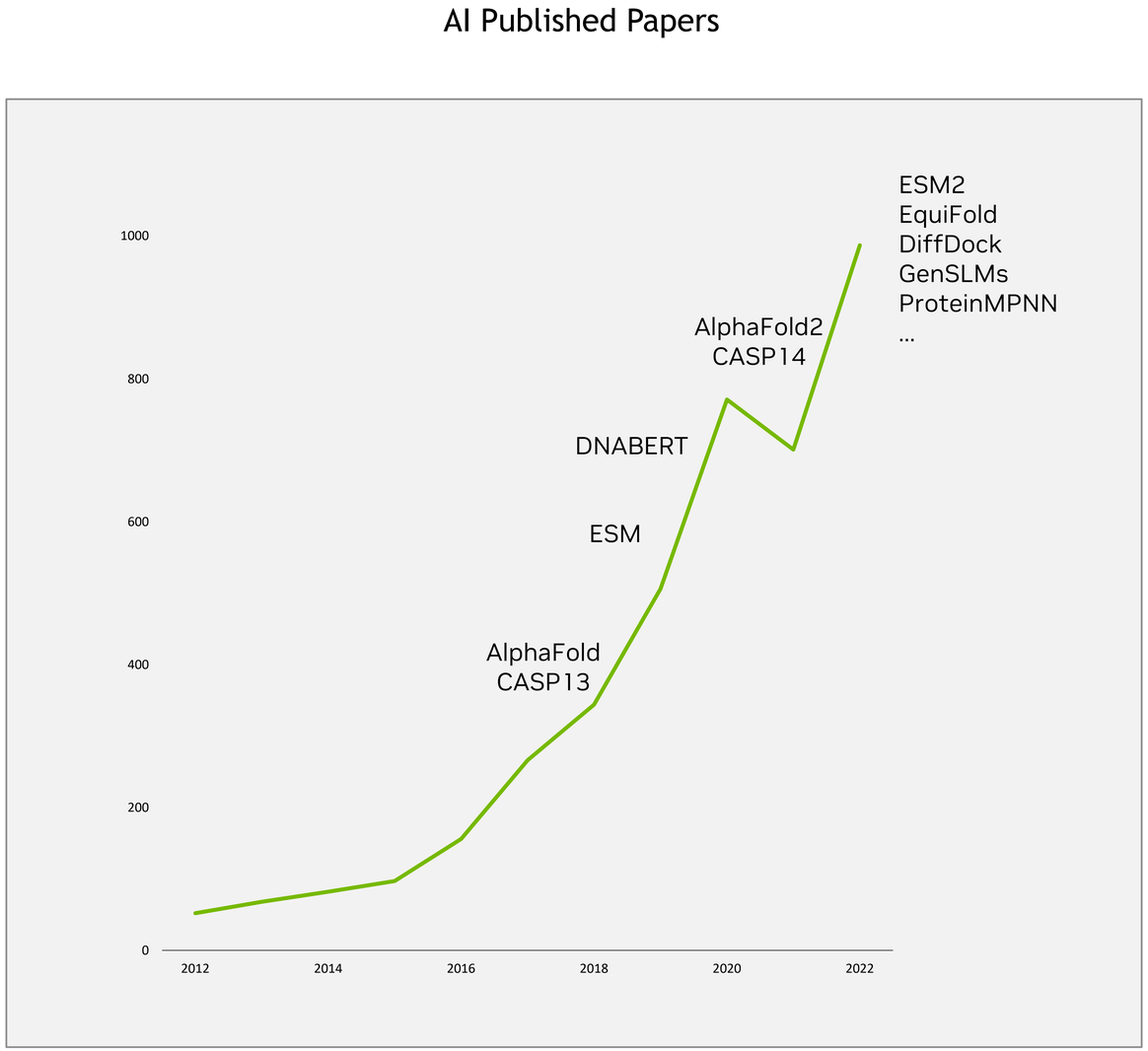
DLWP, Weyn et al. (2020). 2°, 16K pixels, Deep CNN on Cubesphere/(2021) ResNet

Weyn et al. (2019), 2.5° N.H only, 72x36, 2.6k pixels, ConvLSTM

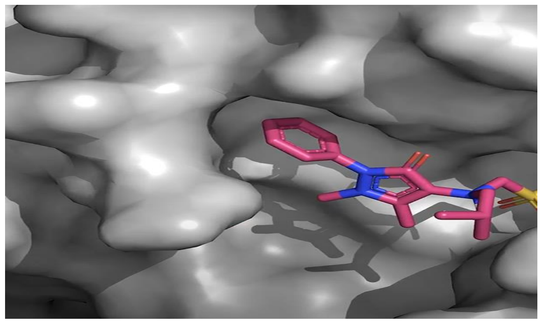
WeatherBench, Rasp et al. (2020). 5.625°, 64x32, 2K pixels, CNN

Deuben & Bauer (2018), 6° , 60x30, 1.8K pixels, MLP

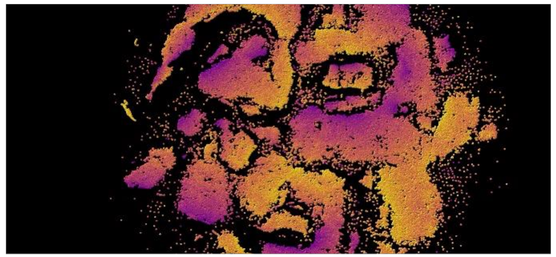
Generative AI making headway into Biology and Drug Discovery



Lab Automation: Sensors & Robotics

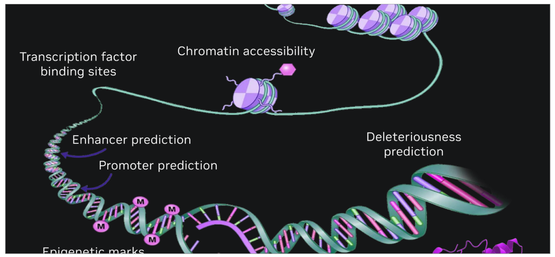


In Silico Drug Discovery: AI & Computing



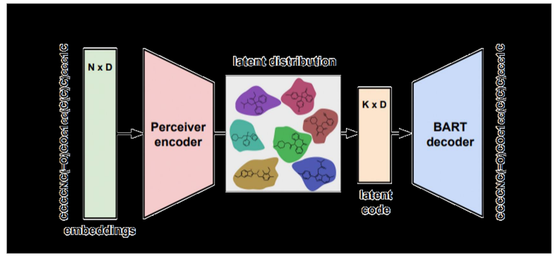
GensLMs

Genome-scale language models reveal SARS-CoV-2 Evolutionary Dynamics



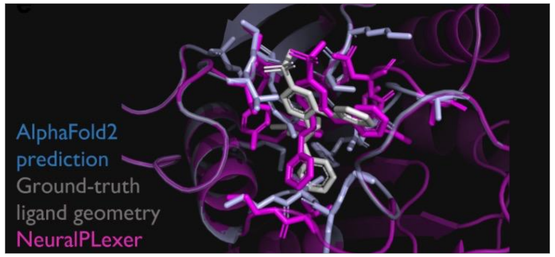
Nucleotide Transformer

Building and Evaluating Robust Foundation Models for Human Genomics



MolMIM

Improving Small Molecule Generation using Mutual Information Machine



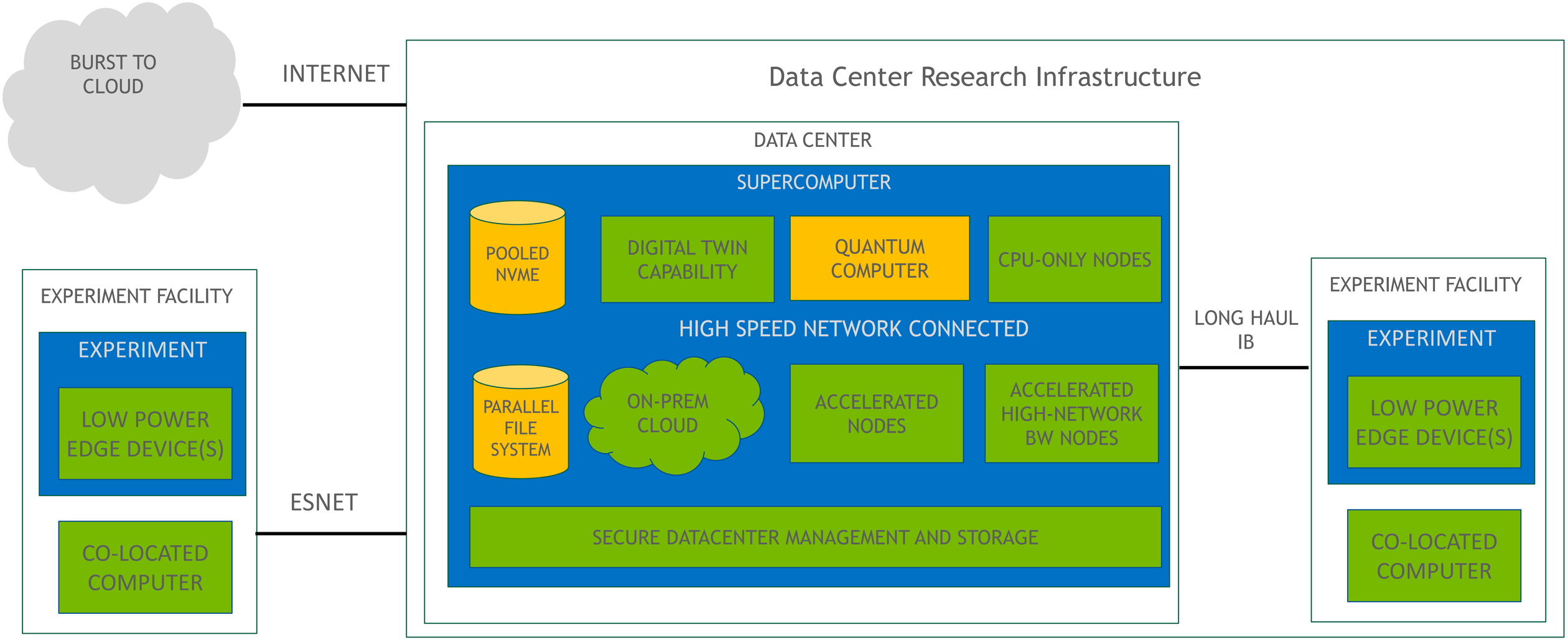
NeuralPlexer

Dynamic-Backbone Protein-Ligand Structure Prediction with Multiscale Generative Diffusion Models

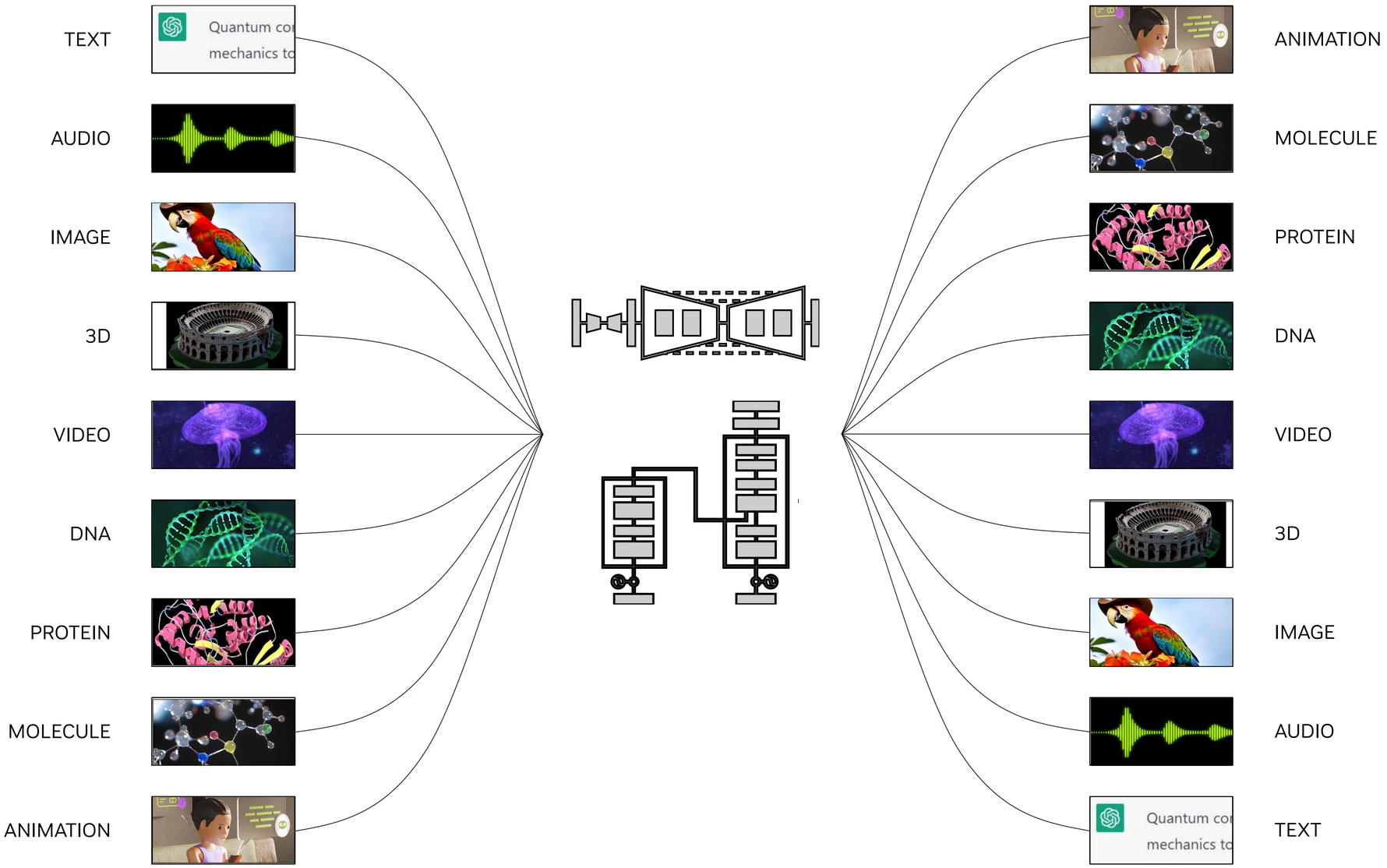
Source: arXiv.org Q-bio: AI, ML, DL, NN

NVIDIA

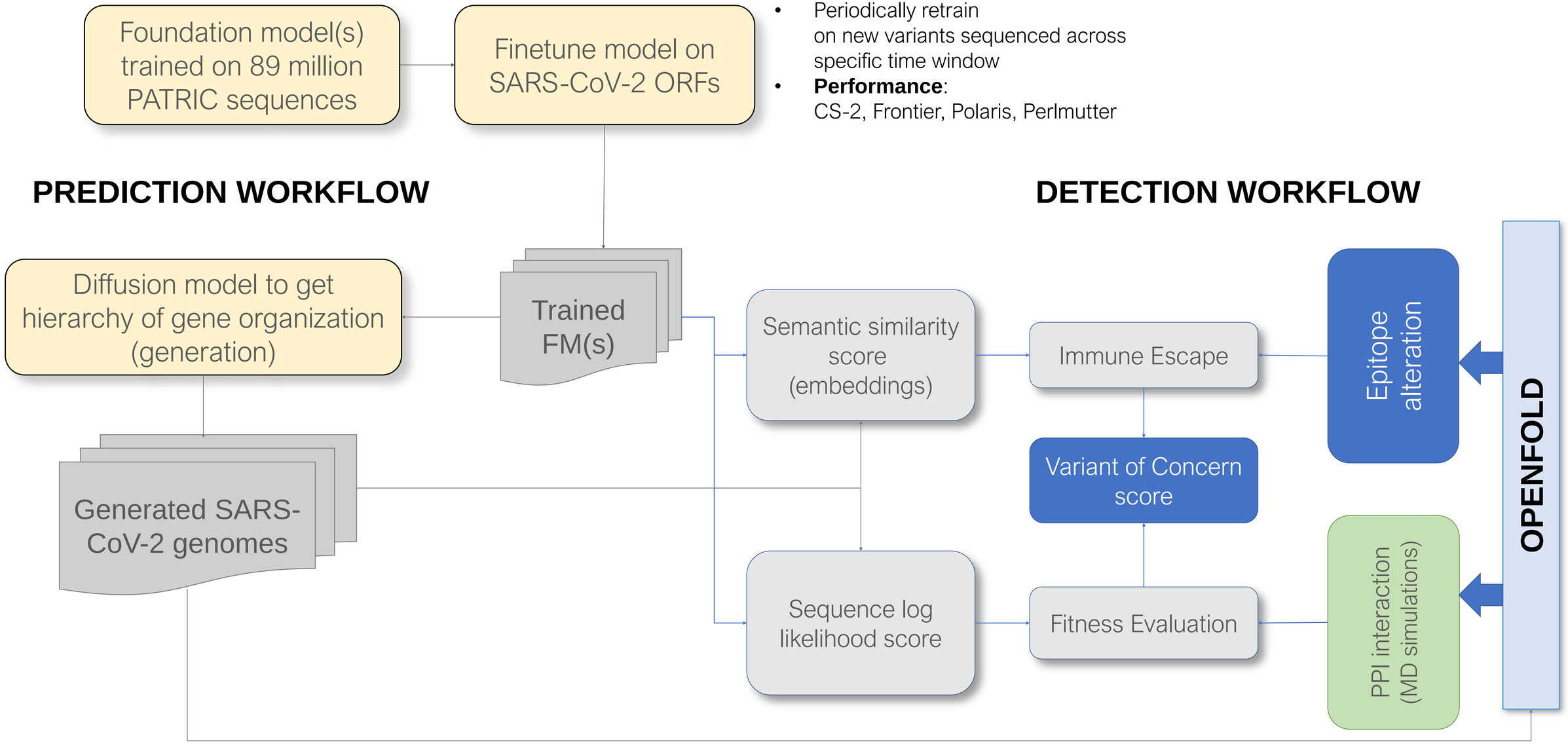
INTEGRATED DATA CENTER OF THE FUTURE



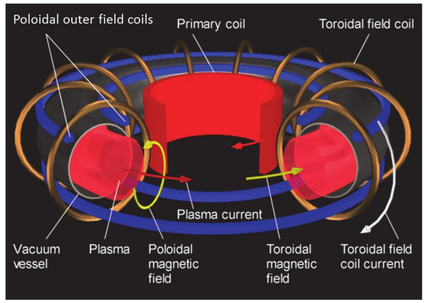
WHAT IS GENERATIVE AI?

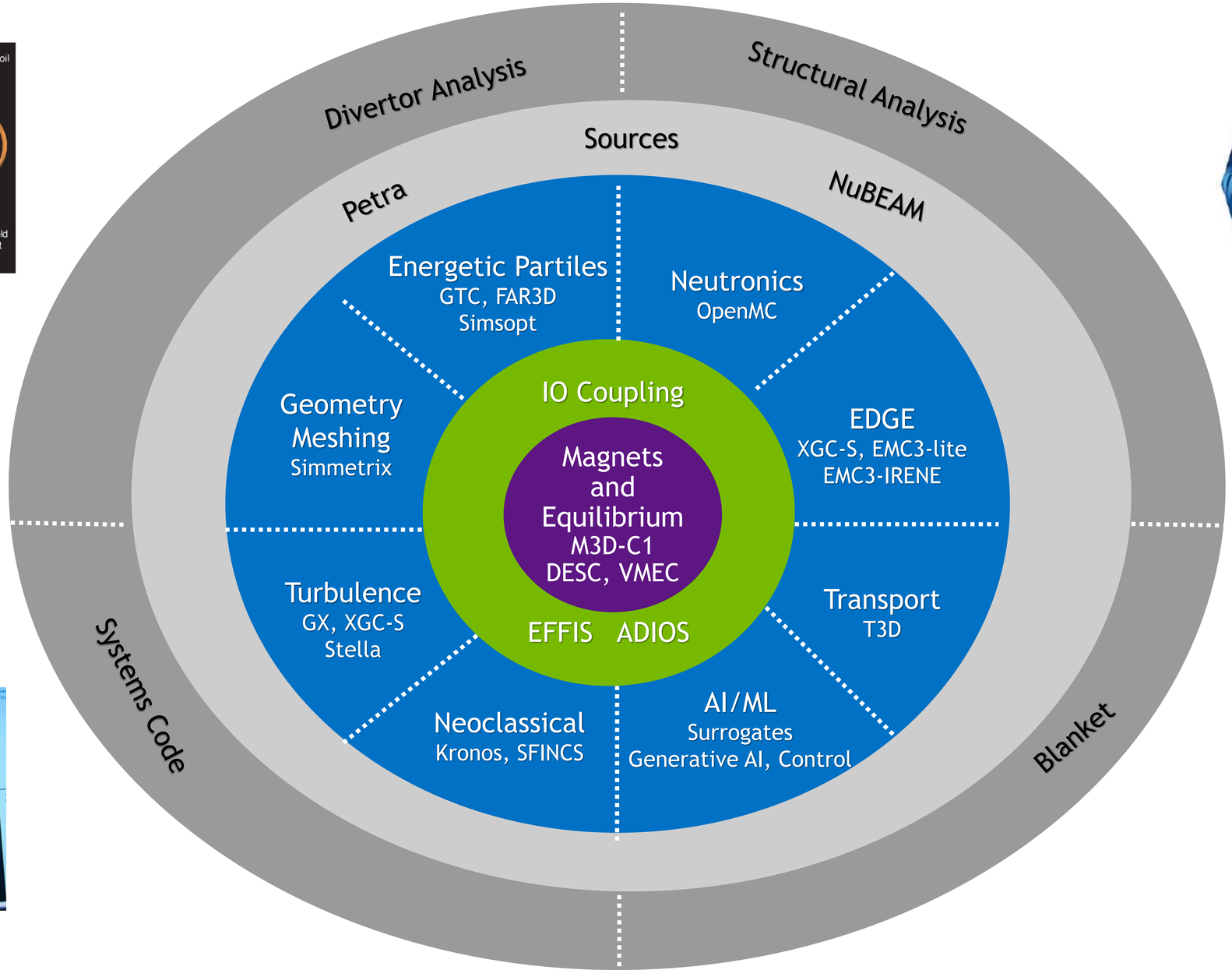


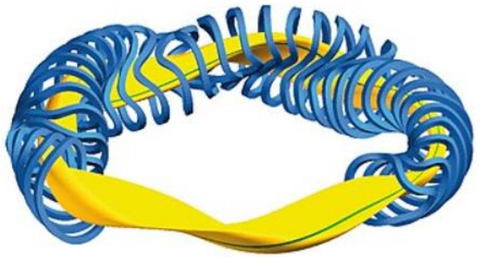
TRAINING

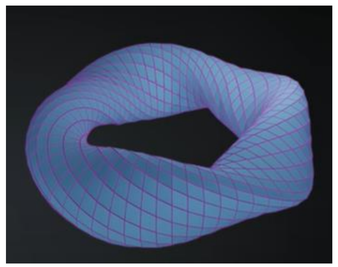

Foundation model(s) trained on 89 million PATRIC sequences
Finetune model on SARS-CoV-2 ORFs
Periodically retrain on new variants sequenced across specific time window
Performance: CS-2, Frontier, Polaris, Perlmutter
PREDICTION WORKFLOW
DETECTION WORKFLOW
Diffusion model to get hierarchy of gene organization (generation)
Trained FM(s)
Semantic similarity score (embeddings)
Immune Escape
Epitope alteration
Variant of Concern score
Generated SARS-CoV-2 genomes
Sequence log likelihood score
Fitness Evaluation
PPI interaction (MD simulations)
OPENFOLD

WORKFLOW FOR FULL SCALE GYROKINETIC FUSION









NVIDIA

CONVERGED WORKFLOW FOR GYROKINETIC FUSION

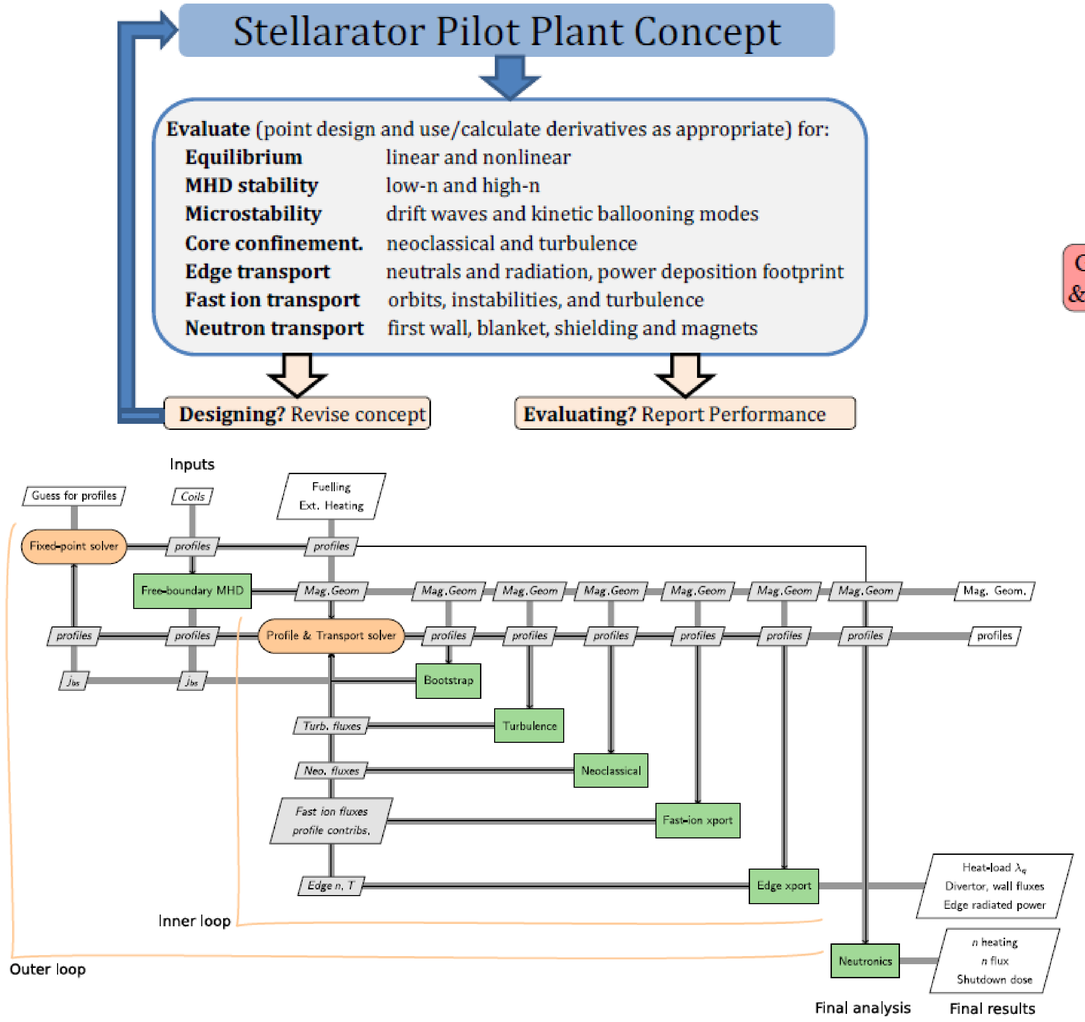
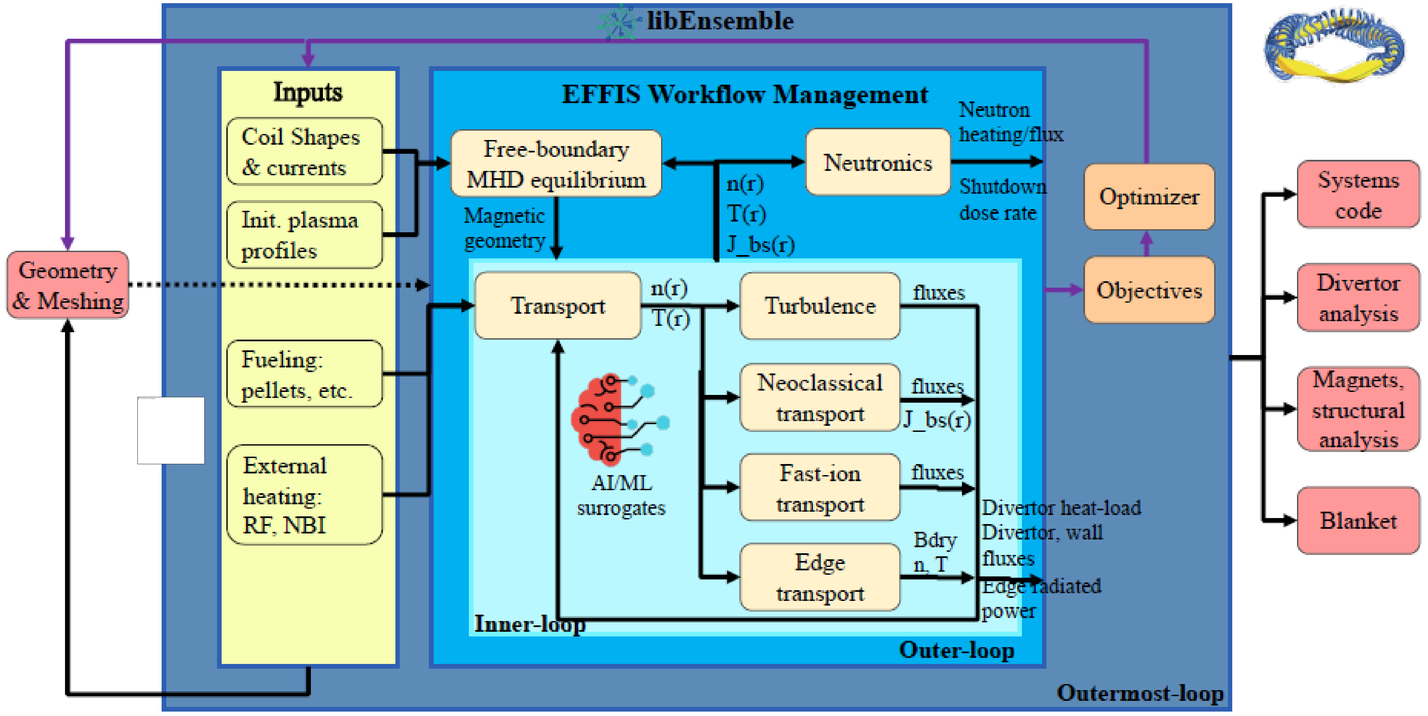
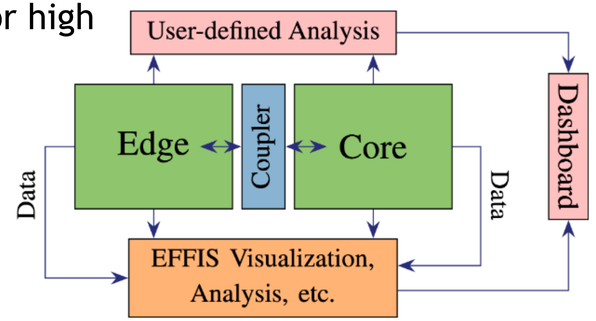


Figure 3: Example workflow for computing self-consistent plasma state. Green rectangles on the diagonal indicate the physics calculations, while the off-diagonal parallelograms show the data passed between codes.



Exascale Framework for high Fidelity Simulation



NVIDIA


NVIDIA
BACKUP