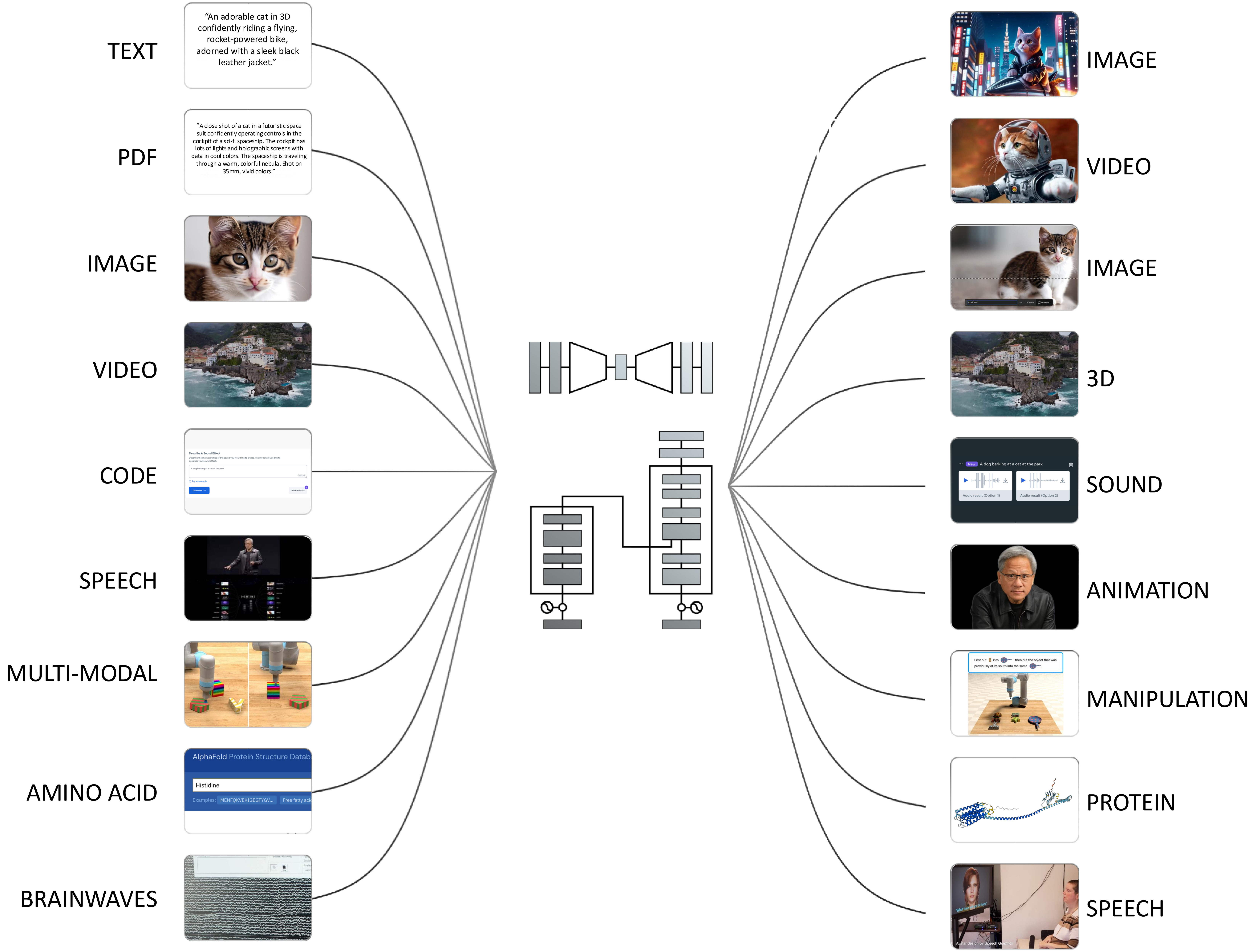
Generative AI for Scientific Research and Discovery

Generative AI Adoption Across Industries



Media & Entertainment

Character Development
Video Editing & Image Creation
Style Augmentation

Finance

Enterprise Search / Doc AI
AI Banking Assistant
Investment Insights

Retail

Automated Catalog Descriptions
Automatic Price Optimization

Manufacturing

Factory Simulation
Product Design

Telecommunications

Network Performance Tuning
Remote Support Capabilities

Healthcare

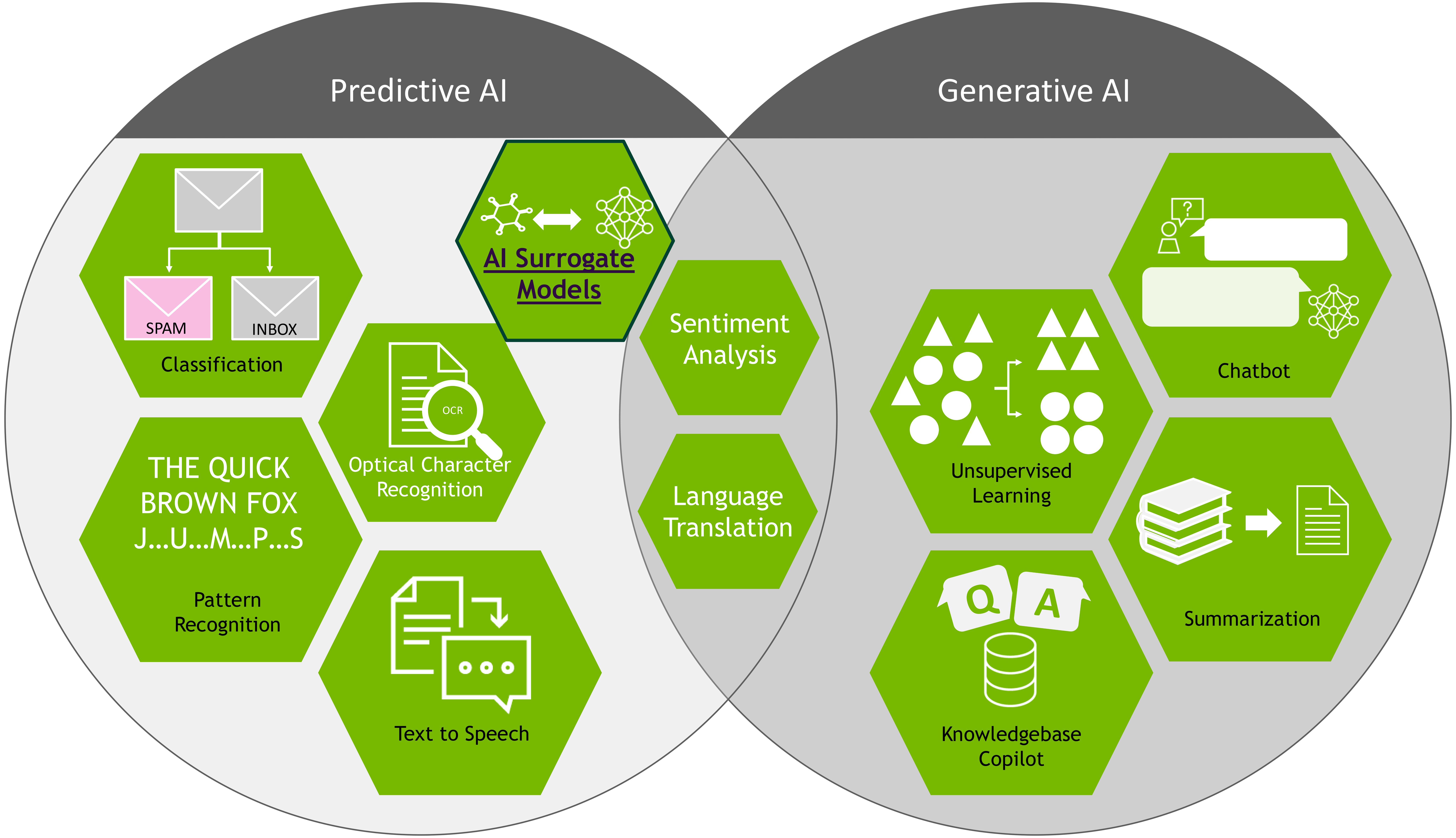
Molecule Design
AI Virtual Assistant

Energy

Knowledge Base Q&A
Predictive Maintenance

Multiple Approaches for Applying AI

Both have Their Strengths



Predictive AI focuses on understanding historical data and making accurate predictions

Generative AI creates new data based on patterns and trends learned from training data

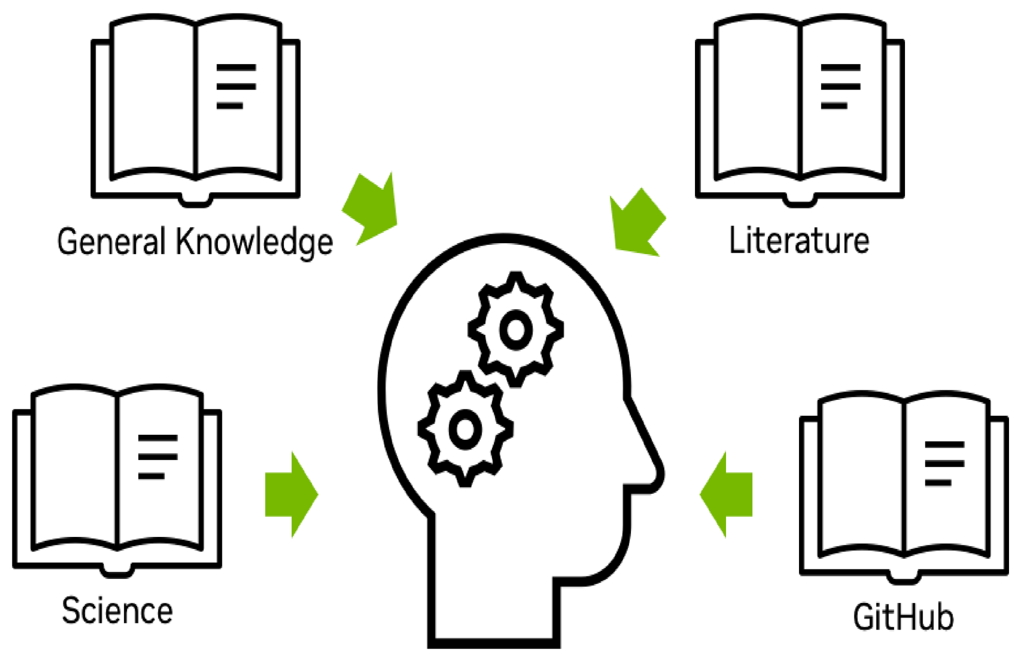
Where GenAI makes sense?

- Painful & Impractical to get a large corpus of labelled data
- Models can learn new tasks
- A single model can serve all use-cases

Foundation Model vs Generative AI Model

There are subtle differences

Predict next token (word)



Foundation Model
(trained on raw data from internet to predict the next word)

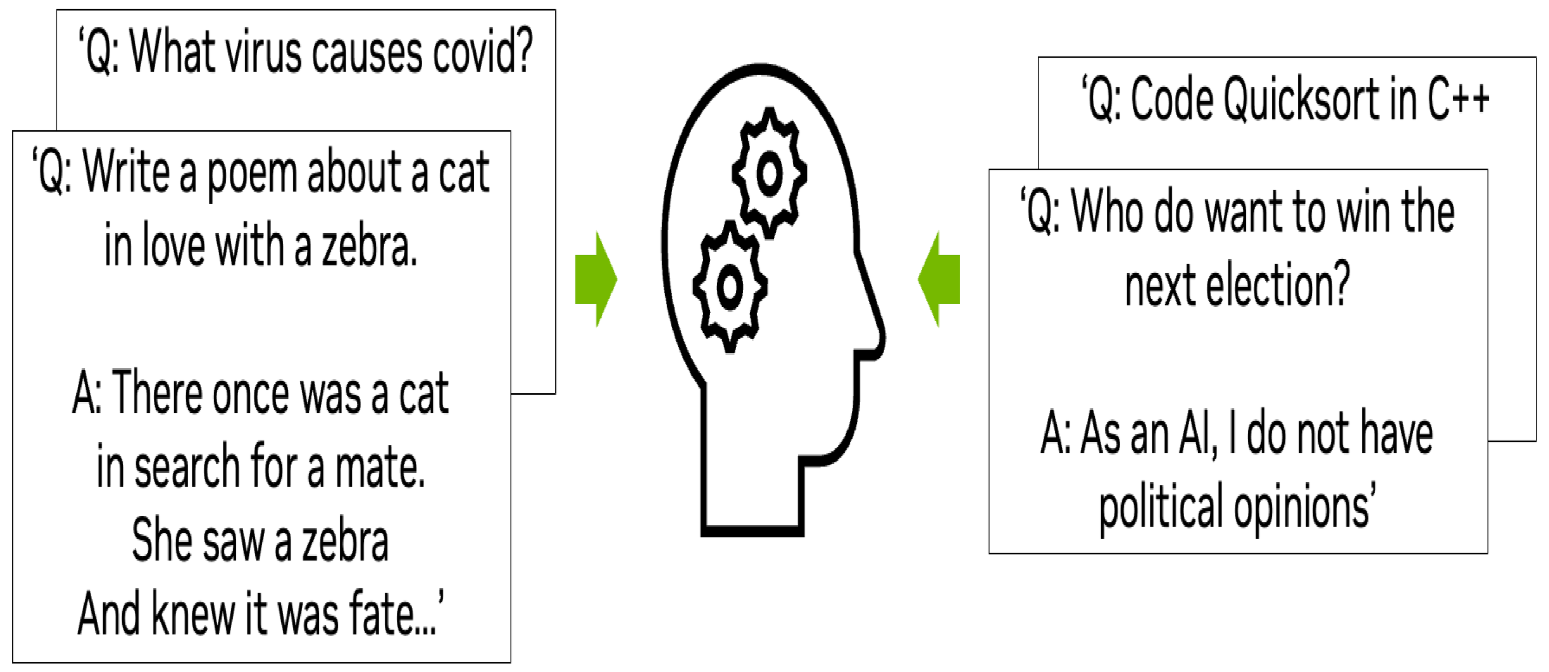
Foundation Model	Generative Model
Serve as the "bedrock" on which generative AI models are built. Foundation models can be Generative AI models, but not all Generative models are "Foundation models"	Discern patterns and relationships within the data Capable of generating new content that can resemble in style and content on training data.

General purpose	Specialization	Task Specific
High	Versatility	Medium
High	Adaptability	Medium
Wide	Range of Tasks	Narrow
Medium	Accuracy for specific tasks	High

Generates, paragraphs, poems based on prompts

Generates code examples

Answer/not Answer questions



Intersection of Gen AI and Science

LLaMa | GPT | Falcon | Starcoder | ChatGLM | MPT | ...

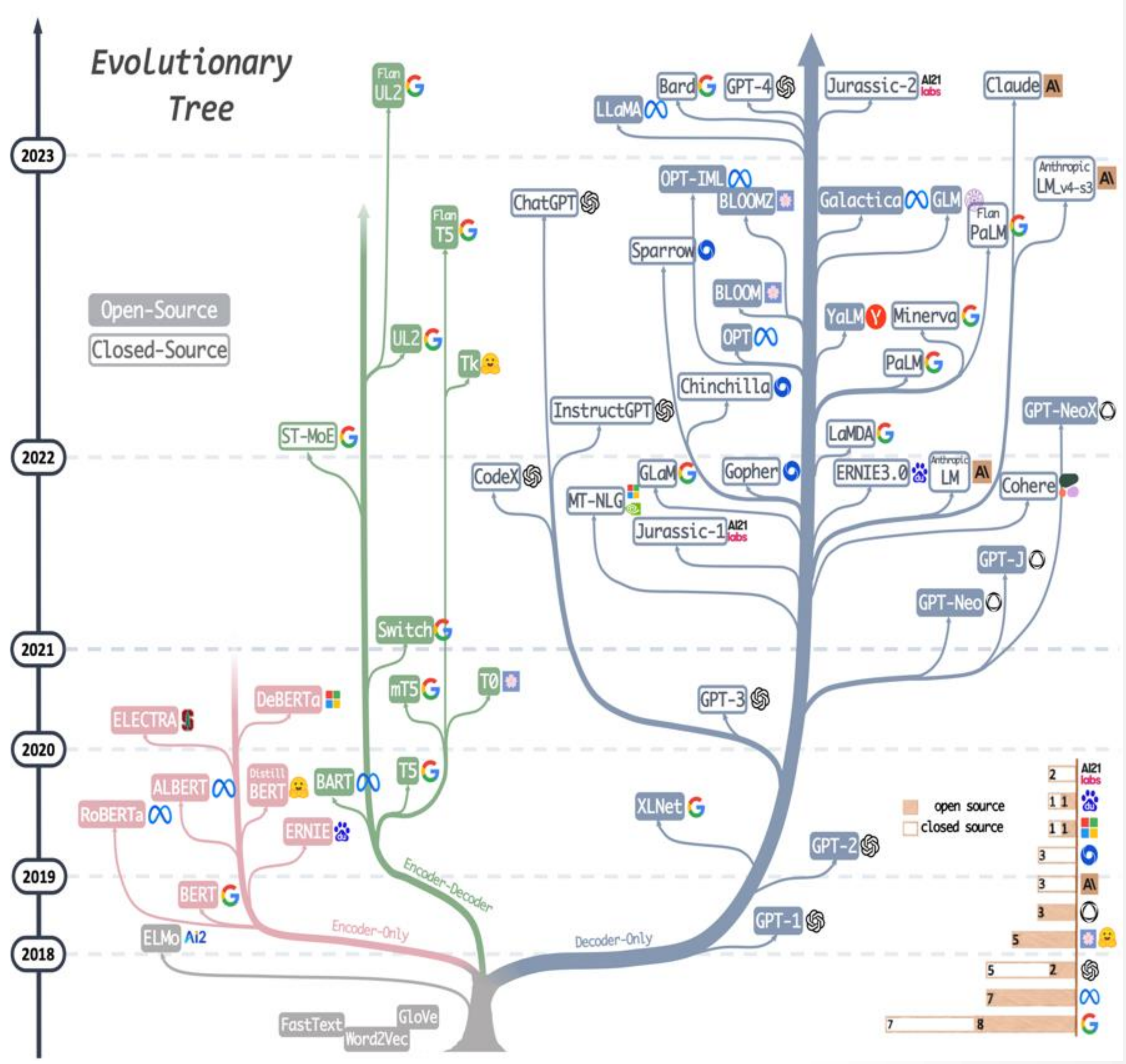
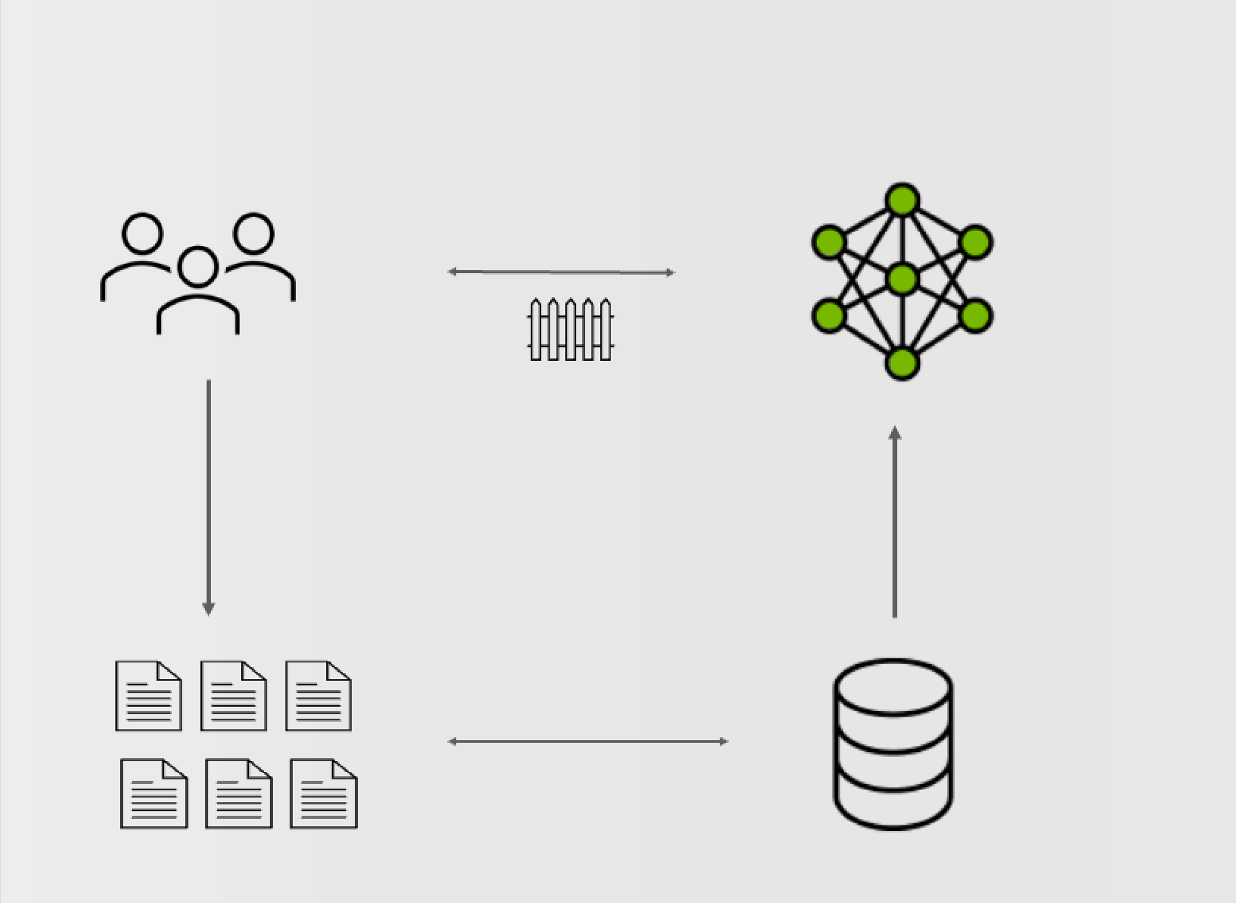


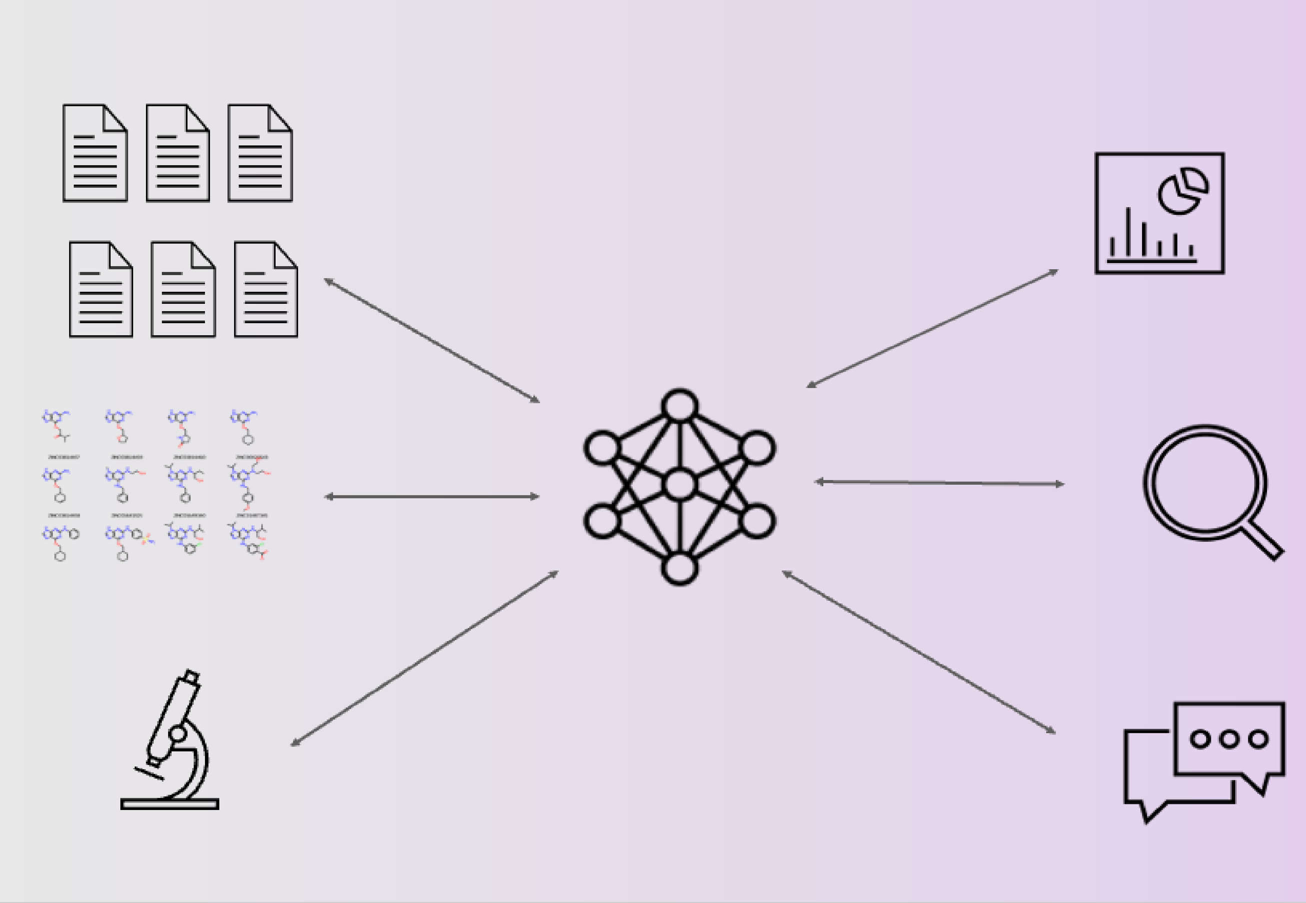
Image from [Mooler0410/LLMsPracticalGuide]()

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., … Hu, X. (2023). Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2304.13712

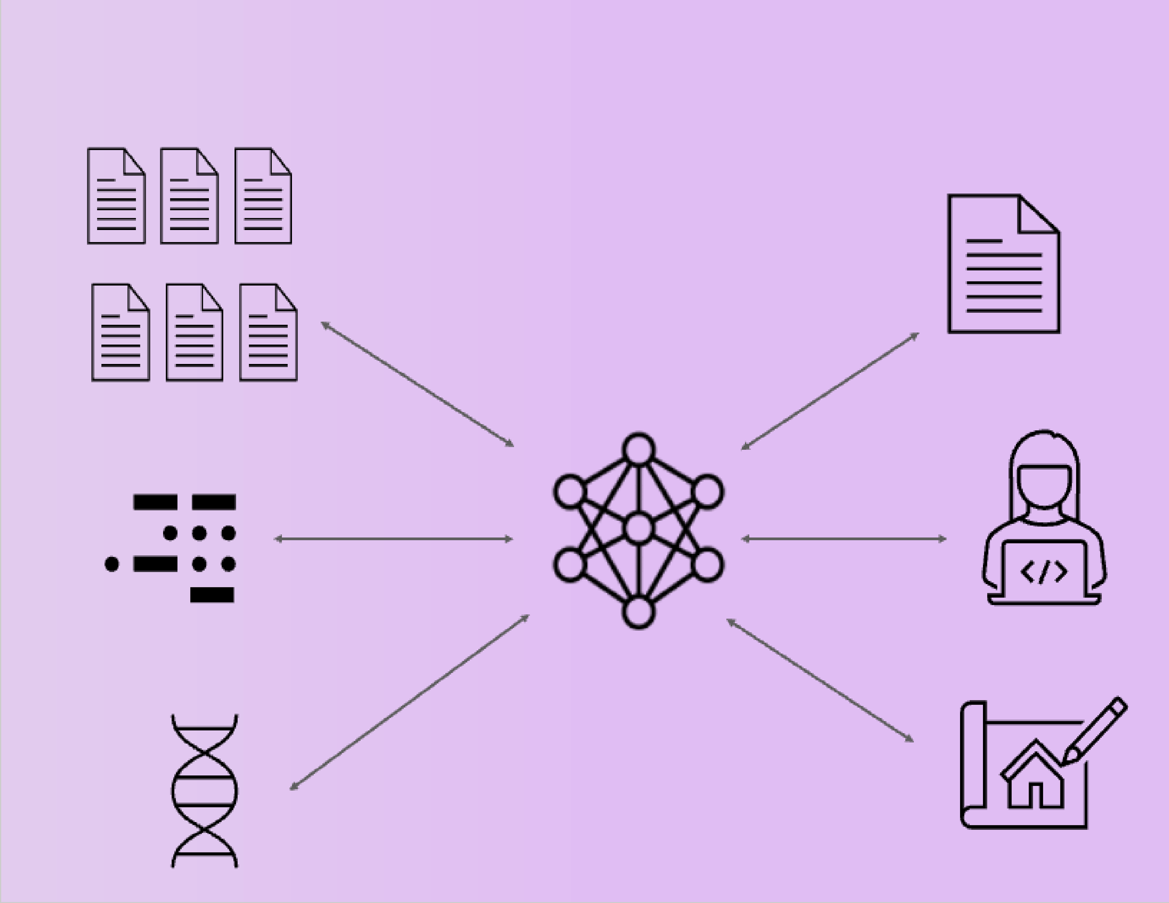
Summarize



Synthesize



Generate



Things to consider when applying GenAI for Science

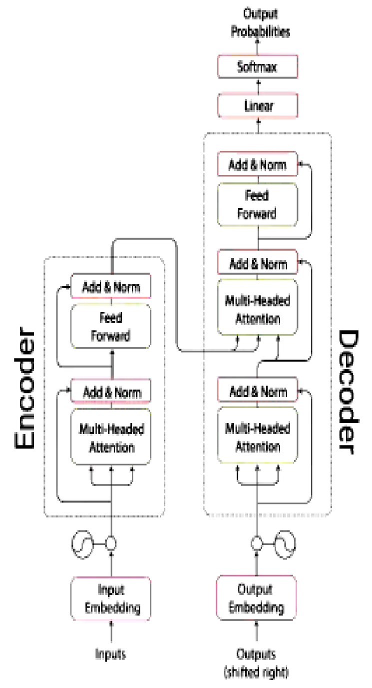
"Some" Challenges for Scaling & Using LLMs in Science

Model

Model Elements
- Ø **Tokenization** for arbitrary data types, formats, and dimensions
- Ø **Encodings**: vs. positional encoding
- Ø **Customized decoders** to generate interpretable output

Model Versatility
- Ø **Extremely long (input) sequences**
- Ø **Reconcile** out of order data
- Ø **Multi-modal** content generation in multiple modalities
- Ø **Instill** physics knowledge to model

Perf.
- Ø **Cost of inference**
- Ø **Serving** w/ batching



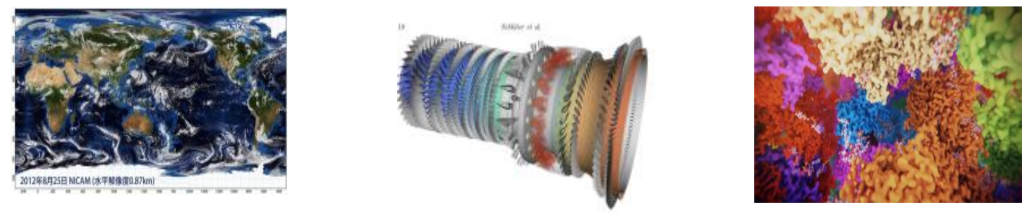
Data

Nature of Data
- Ø **Massive datasets**: O(1 exabytes) datasets
- Ø **Semi-structured data**
- Ø **High dimensional data**
- Ø **Different views of same data**

Handling
- Ø **Tooling for data curation**: 10x nightmare!
- Ø **Streaming vs. accumulate then train**
- Ø **Data managers and repositories**

Human Factor
- Ø **Human in the loop**
- Ø **Attribution and Provenance**

Integration into Simulations



Reliability of AI as a Tool
- Ø **Correctness and UQ**: tooling to assuring correctness and UQ
- Ø **Software engineering**: seamless integration of AI into simulations

Composability
- Ø **Tandem** simulation and training
- Ø **Interacting models:** models of different sizes interact in different phases of simulation
- Ø **Ensembles** of light-weight models as smart agents

6

- Data Types differ vastly
- Custom Tokenizers
- Extremely long sequence length
- Un-ordered/unstructured data
- Not just data, also meta-data
- Multiple modalities
- Interaction with Simulation applications
- Data management and attribution

Source : ISC2024 session on GenAI for Science by Dr Wahib from Riken

Biology and Life Sciences

Transformers Meets Biology

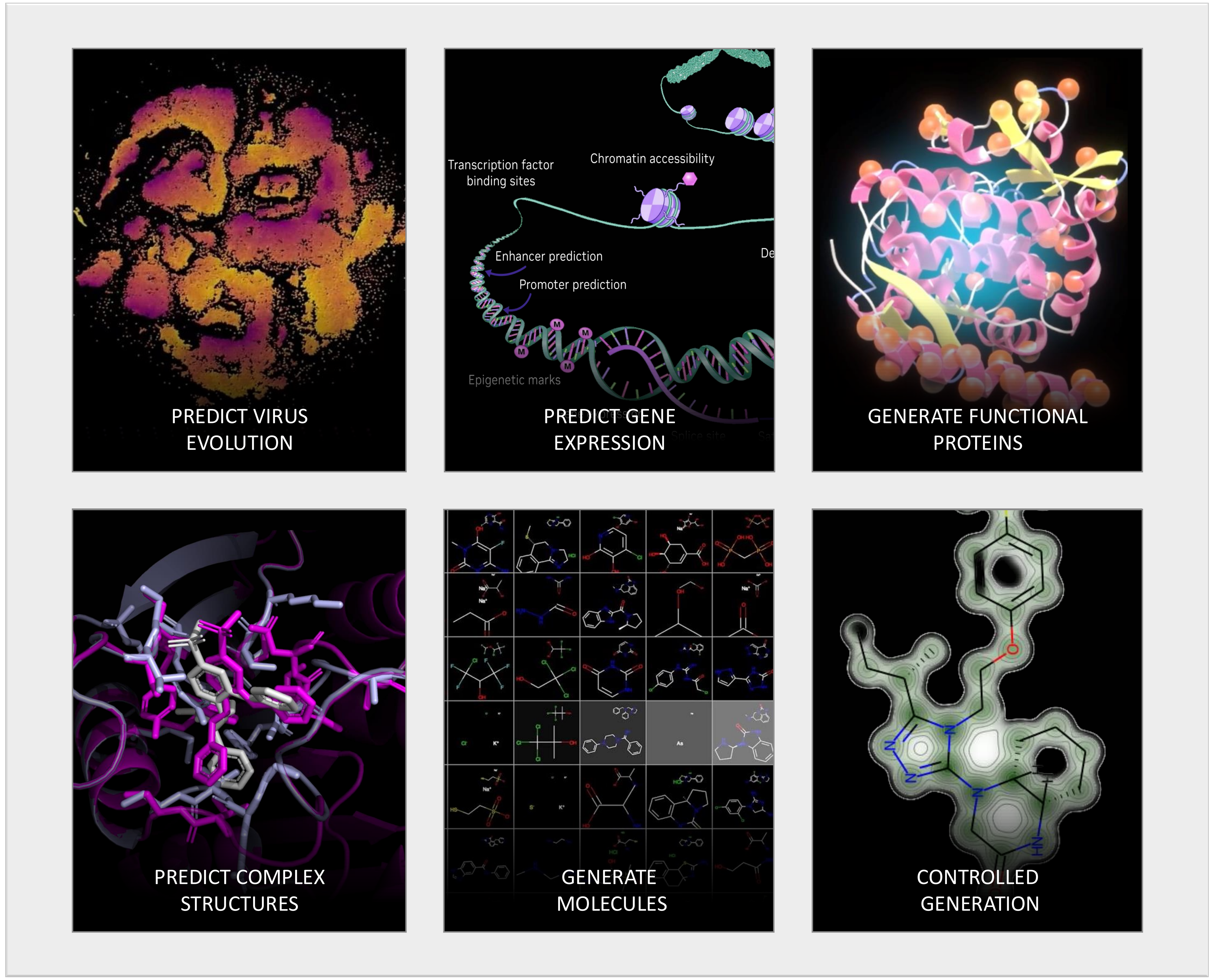
2024
Alphafold3
NeuralPlexer

2022
ESM2
EquiFold
DiffDock
OpenFold
ProteinMPNN
ProtGPT2
...

2020
AlphaFold2

2018
AlphaFold

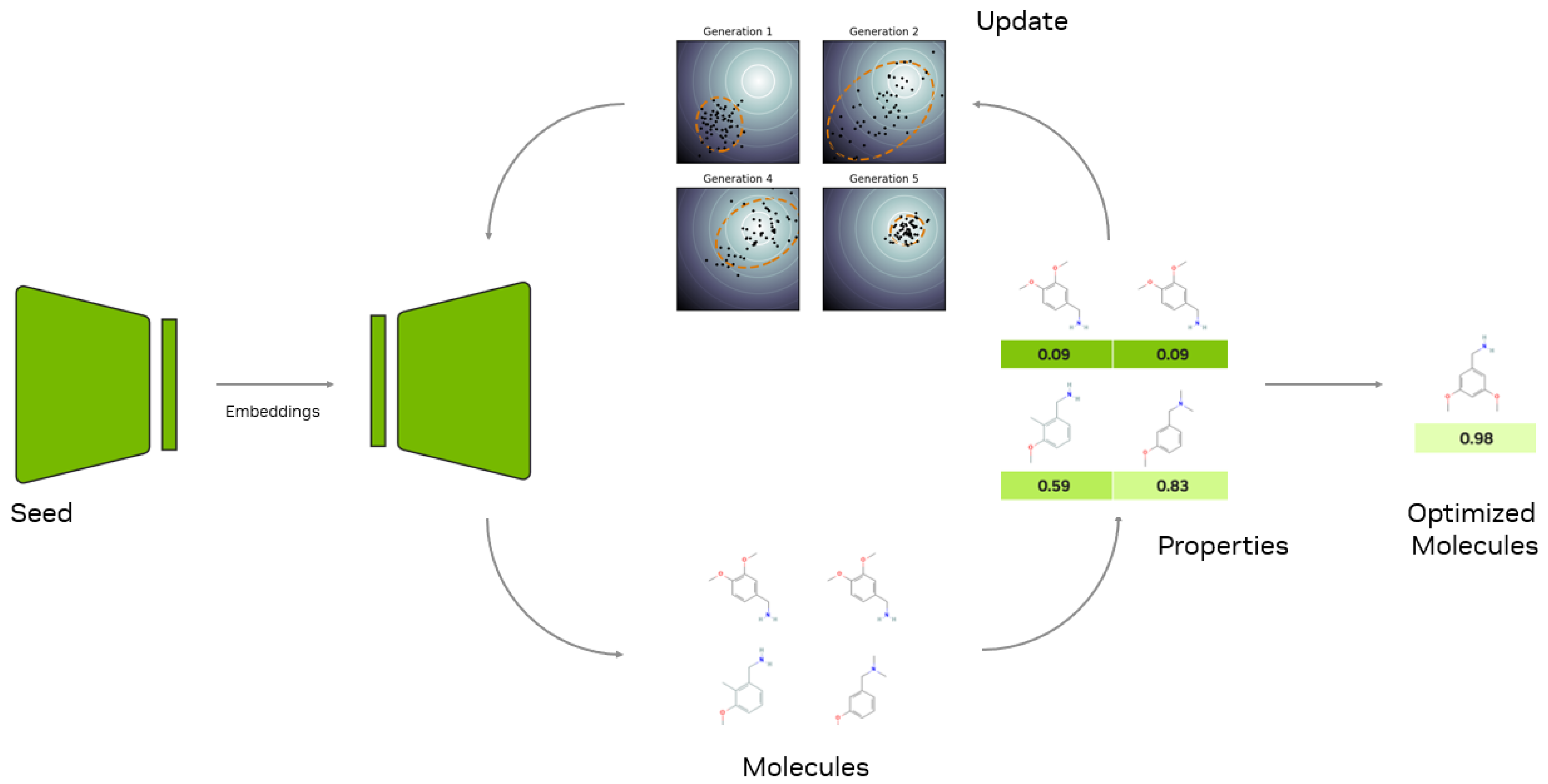
2017
Attention is all you need



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

MolMIM

GenAI model for small molecule drug discovery



What does it do

- Controlled molecular generation
- Multiparameter Optimization
- Accepts User-Defined Oracles (user-specified scoring function)

Input

- Hit molecule (SMILES format)

Output

- New, optimized molecules

Architecture

- Encoder-decoder

Where MolMIM can be used

- Virtual screening
  - *Rapid computational evaluation of large chemical libraries to identify potential hits that can bind to a target*
- Lead optimization
  - *Improve molecular hits from initial experiments to improve biochemical qualities needed for a drug (e.g. low toxicity, proper distribution)*

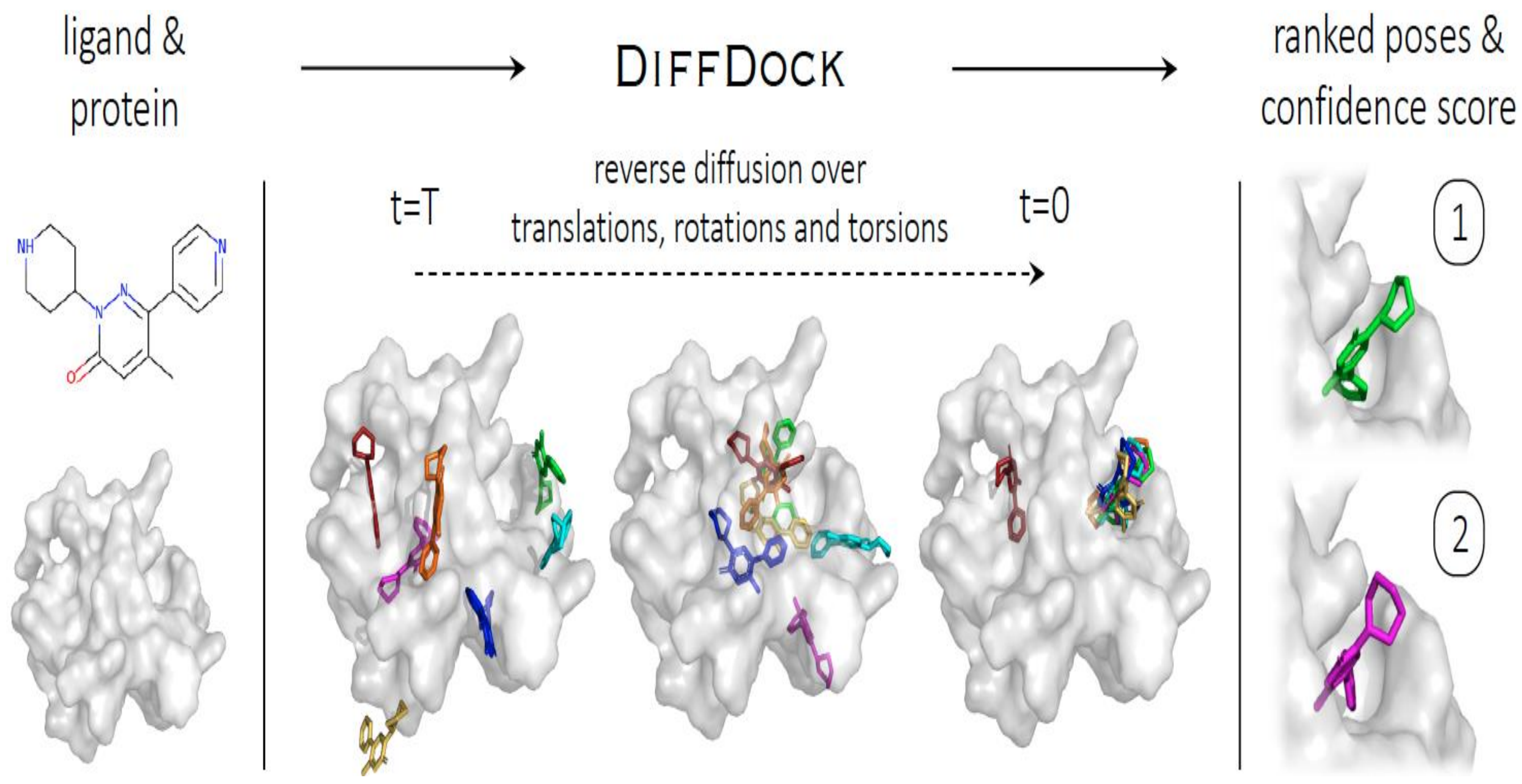
MolMIM : https://arxiv.org/abs/2208.09016

https://docs.nvidia.com/bionemo-framework/latest/models/molmim.html

NVIDIA Aug 2022

DiffDock

Diffusion generative model for molecular blind docking



DiffDock
MIT https://arxiv.org/abs/2210.01776
https://github.com/gcorso/DiffDock?tab=readme-ov-file

What does it do

- Predict protein-Ligand binding
- Generate binding poses

Input

- Molecule and protein structure(PDB, SDF, MOL2, SMILES)
- **Output**
- 3D pose prodiction
- **Architecture**
- Score-Based Diffusion Model (SBDM)
- GCNN

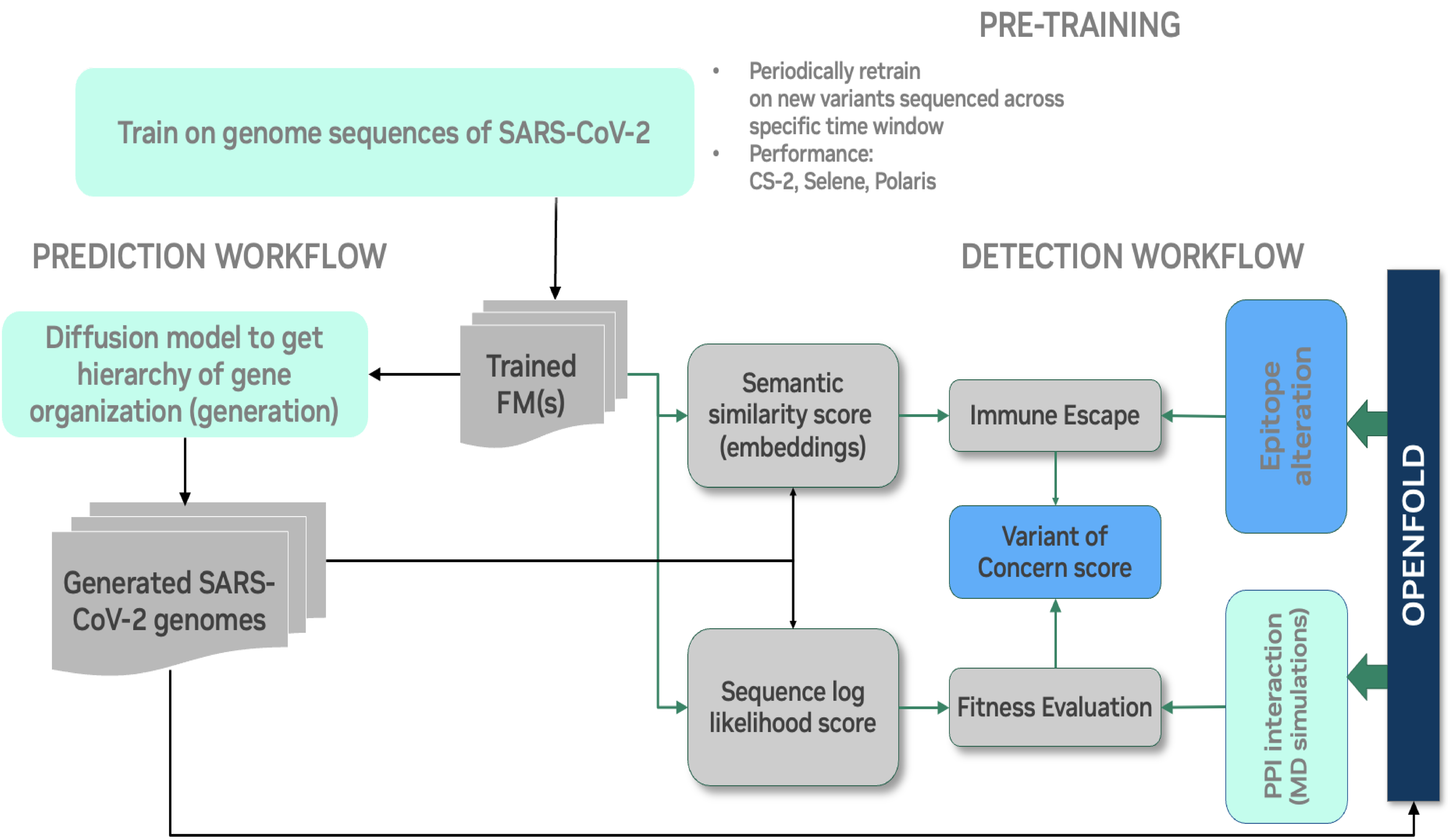
Where DiffDock can be used

- Rapidly screen large libraries of compounds against target proteins, identifying potential drug candidates

GenSLM

Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics

ACM Gordon Bell Special Prize
for High Performance Computing-Based
COVID-19 Research



- Base model trained on more than 110 million gene sequences from prokaryotes, which are single-celled organisms like bacteria.
- Fine-tuning using 1.5 million high-quality genome sequences of COVID virus.
- Once trained, GenSLM was able to:
  - Distinguish between genome sequences of the virus' variants.
  - Generate its own nucleotide sequences, predicting potential mutations of the COVID genome that could help scientists anticipate future variants of concern.
- Learnings
  - Handle long sequence length

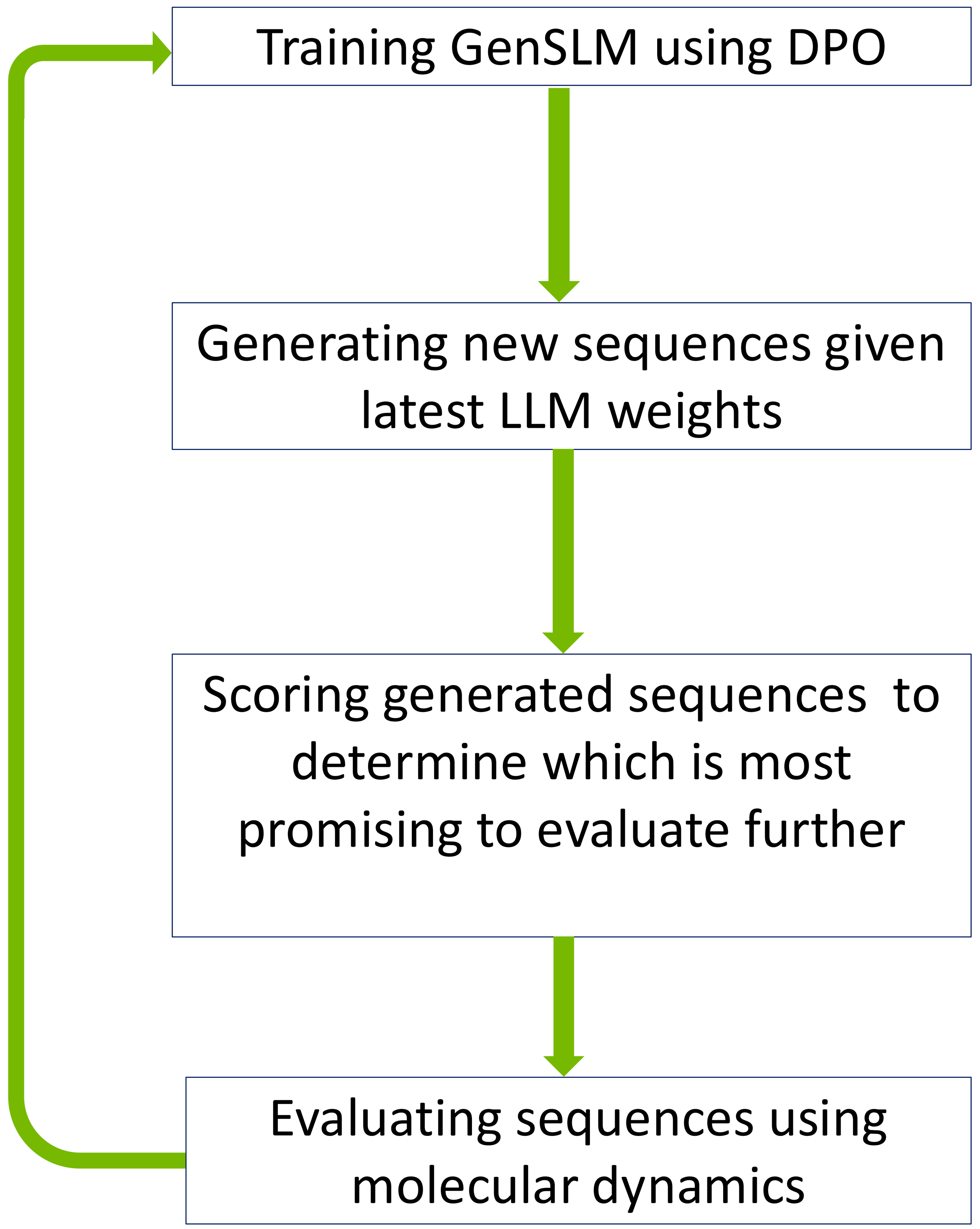
https://github.com/ramanathanlab/genslm

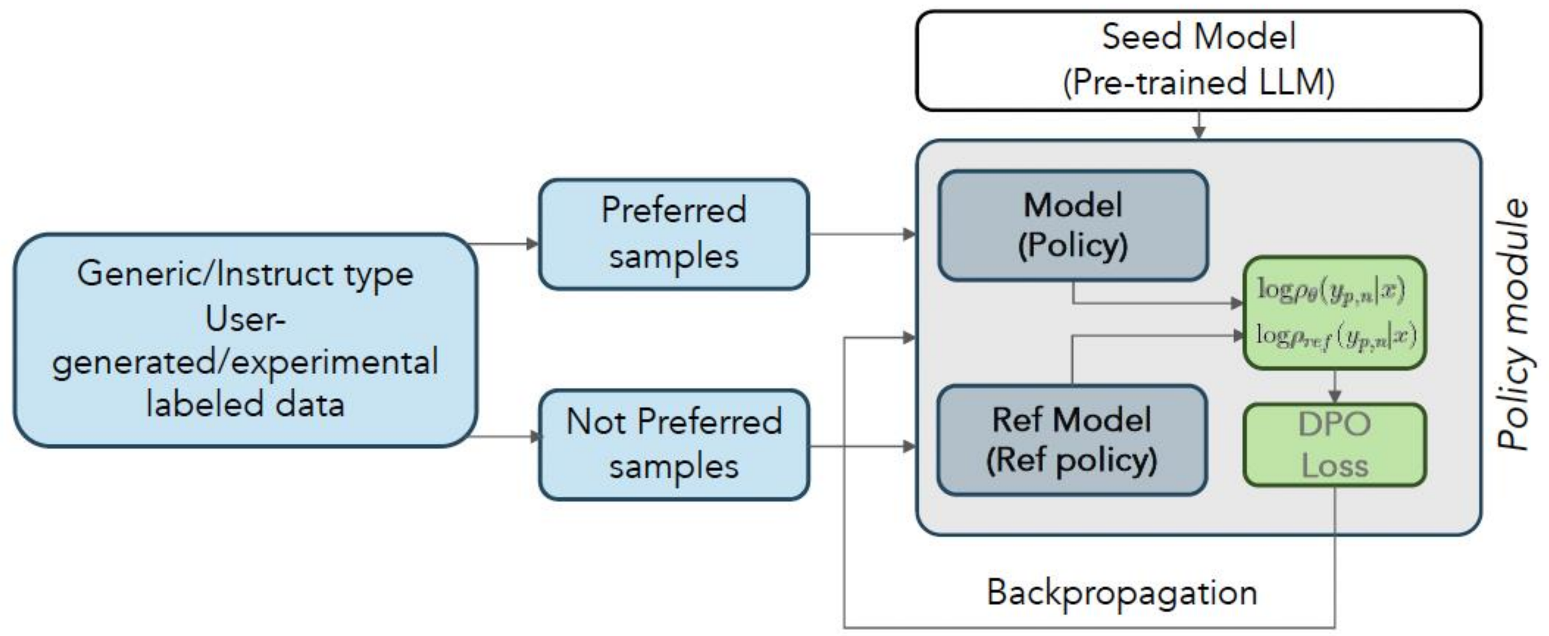
Zvyagin, Maxim, et al. "GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics." *The International Journal of High Performance Computing Applications* 37.6 (2023): 683-705.

Combining Multi-Modality With GenSLM

Integrating Experimental Observations into Generative Modeling Workflows

- Incorporating biophysically-informed fine-tuning schemes to explore a range of generative tasks
- Input data : Protein Sequence and text/knowledge-based description of the protein sequence
- Applied Direct Preference Optimization (DPO) for fine tuning the model.
- DPO generates protein sequences with fitness tuning
  - Ability to steer the multimodal generative model to sample new protein sequences with natural language prompting.





The protein designated by the unique identifier A0A140D2T1_ZIKV_M1632, has a PropertyName=<Deep Mutational Scanning (DMS) score> of PropertyVal=<-2.34> indicating it is Fitness=<unfit>. It has a composition of Alanine, Lysine, and Glutamic Acid, accounting for 56.41% of its total 39 residues. This molecule has a mass of 4357.98 Da. Further analysis reveals the following physicochemical properties: an instability index of 53.96 suggesting its instable disposition, an aromaticity of 2.56,and an isoelectric point computed to be 9.22. The protein's average flexibility is documented at 1.02, with a standard deviation of 0.000561. Additionally, its hydropathicity, as measured by a GRAVY score, is -0.64. Its sequence is <SSEQ>M I S N A K I A N I N E L A A K A K A G V I T E E E K A N Q Q K L R Q E Y L K<ESEQ>

Fine Tuning of GenSLM using PPO and DPO

Ensuring that Model's output is aligned with desired output

Proximal Policy Optimization

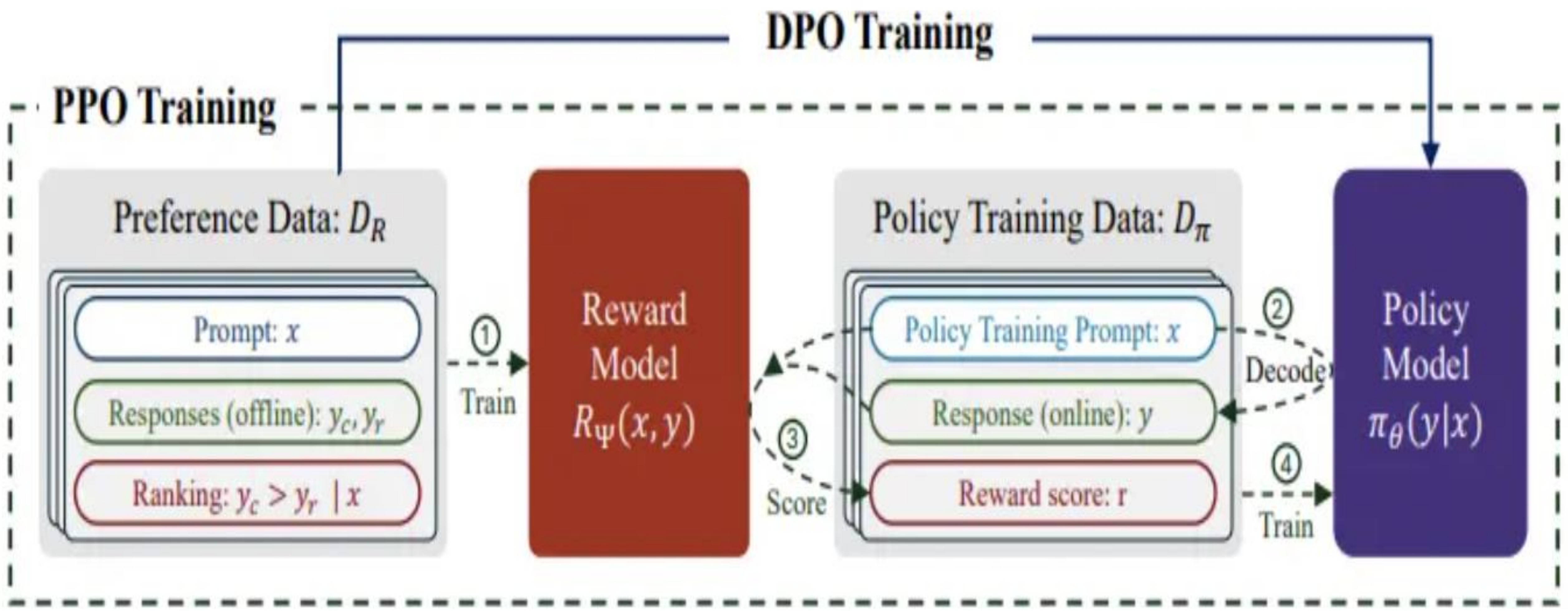
Policy Initialization
(a function that maps states to actions)

Data Collection

Estimate Advantage

Update Policy

Iteration



Source: https://arxiv.org/pdf/2406.09279
Unpacking DPO and PPO: Disentangling Best Practices for Learning from Preference Feedback

Direct Proximity Optimization

Data Collection

Model Initialization

Preference Modeling

Parameter Update

Iteration

Adept at handling complex reward structures and exploring a wider range of potential solutions

Well-suited for complex tasks. E.g Code generation, Autonomous driving

DPO simplifies the training process, also enhances stability and reduces computational overhead

Well-suited for simpler-narrow focus tasks

Generative AI in Climate/Weather

Foundation Model And Generative Models for CWO

Prominent Examples

FOUNDATION MODELS

- ClimaX: A foundation model for weather and climate – (UCLA- Nguyen, Grover, Microsoft, Scaled Foundations)
- Stormer – Scaling transformer neural networks for skillful and reliable medium-range weather forecasting (UCLA, DOE Argonne)
- AURORA: A FOUNDATION MODEL OF THE ATMOSPHERE (Microsoft)
- Prithv-WxC - NASA MSFC(Marshall Space Flight Center) IBM
- ORBIT – Oak Ridge Base Foundation Model for Earth System Predictability
- AtmoRep – ECMWF; Juelich SC; CERN AtmoRep: A stochastic model of atmosphere dynamics using large scale representation learning (ECMWF; Juelich SC; CERN)
- HClimRep – Juelich; AWI; KIT; Hereon

GENERATIVE MODELS

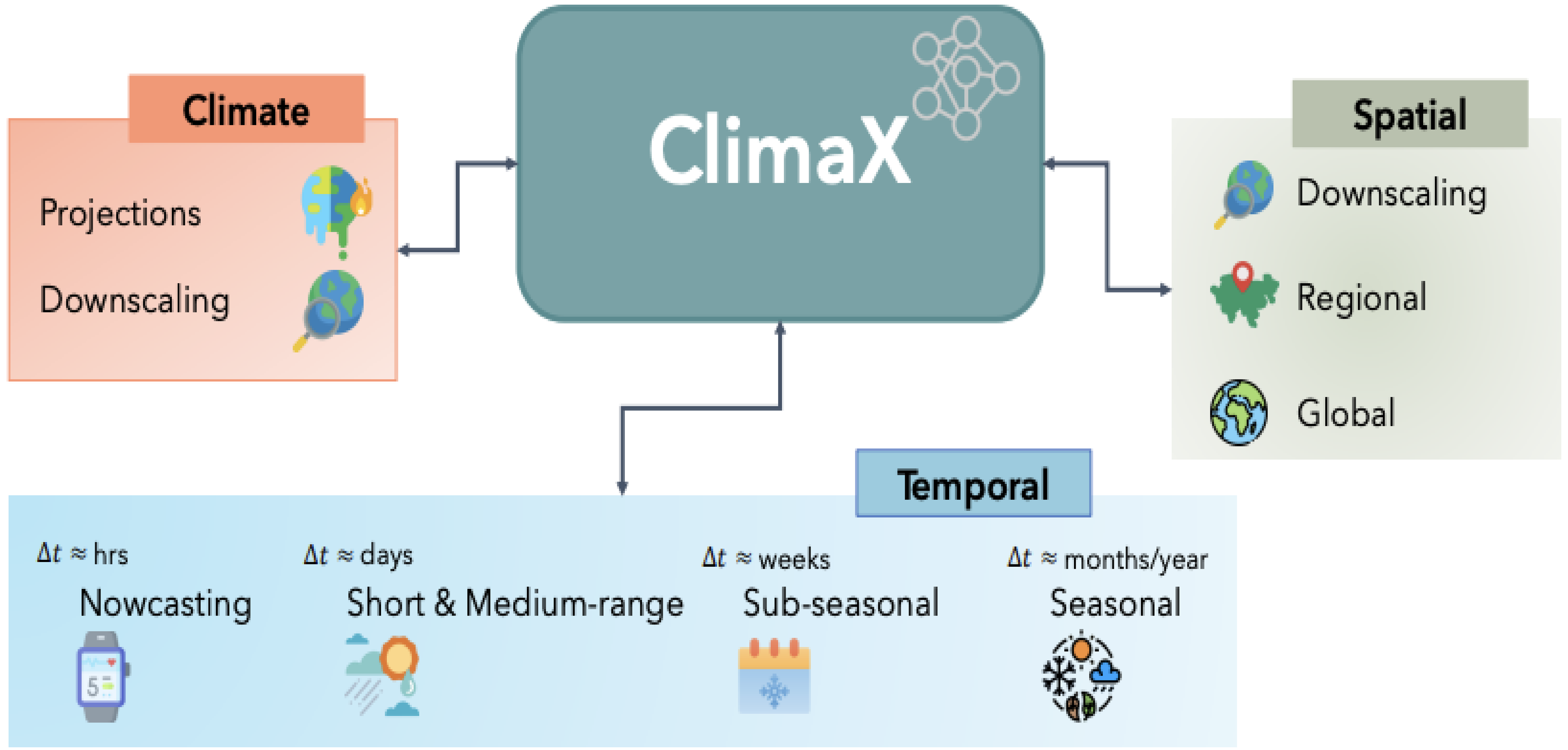
- CorrDiff: Generative diffusion modeling for regional km-scale downscaling (NVIDIA)
- StormCast– Scaling transformer neural networks for skillful and reliable medium-range weather forecasting (NVIDIA)
- GenCAST: Diffusion-based ensemble forecasting for medium-range weather (Google Deepmind)

GEN AI + Data Assimilation

- Deep generative data assimilation in multimodal setting (Columbia University)
- DiffDA: A diffusion model for weather-scale data assimilation (ETH, ECMWF)
- Generative data assimilation of sparse weather station observations at kilometer scales (NVIDIA, University of Oxford, UC-Irvine

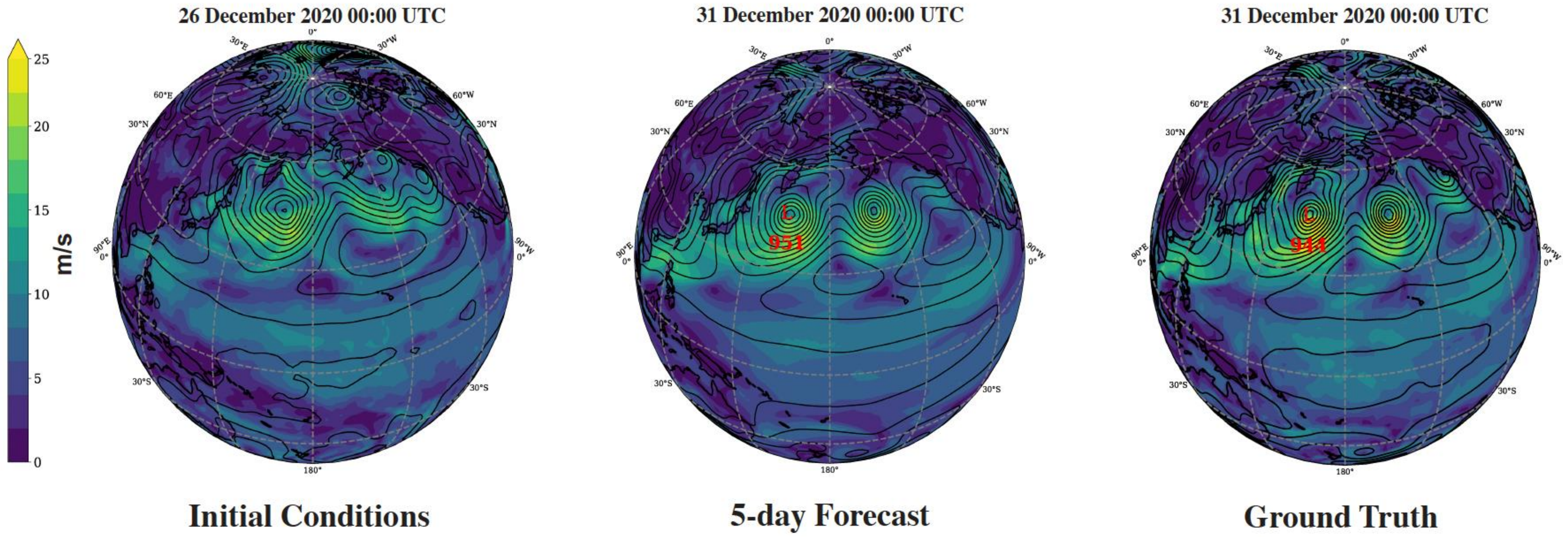
From ClimaX to Stormer

Competitive performance at short to medium-range forecasts with less training data and compute



A foundation model for weather and climate – (UCLA- Nguyen, Grover, Microsoft, Scaled Foundations) arxiv.org/pdf/2301.10343 [Submitted on 24 Jan 2023 *(v1), last revised 18 Dec 2023 (this version, v5)]*

Dataset : Trained on CMIP6 and fine tuned with ERA5
Trained on : 80x V100



Stormer – Scaling transformer neural networks for skillful and reliable medium-range weather forecasting,(UCLA, DOE Argonne) [Submitted 6 Dec 2023]

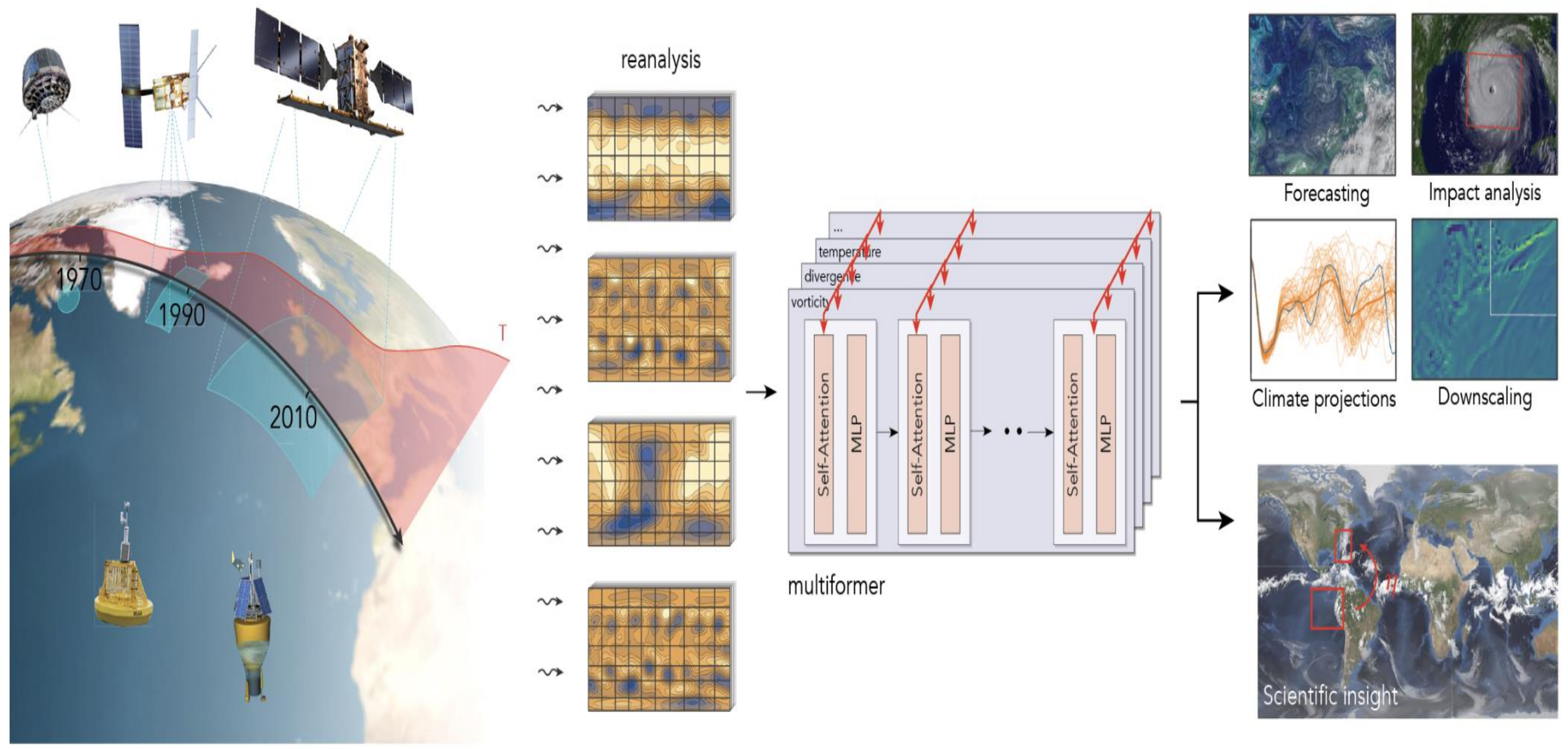
Dataset : ERA5 reanalysis data ECMWF
Trained on. : 128 40GB A100

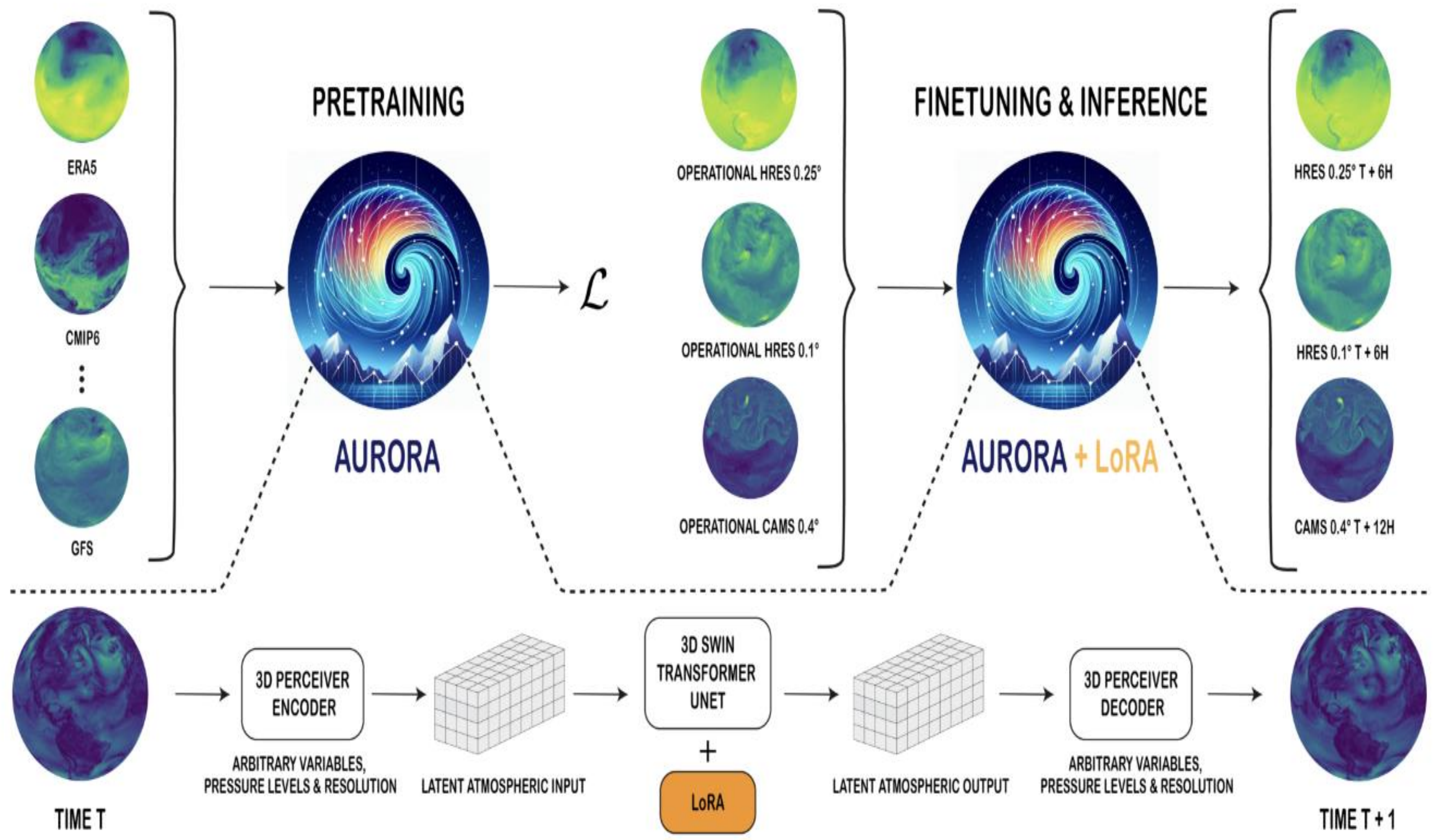
Adapting Vision Transformer to Weather Data

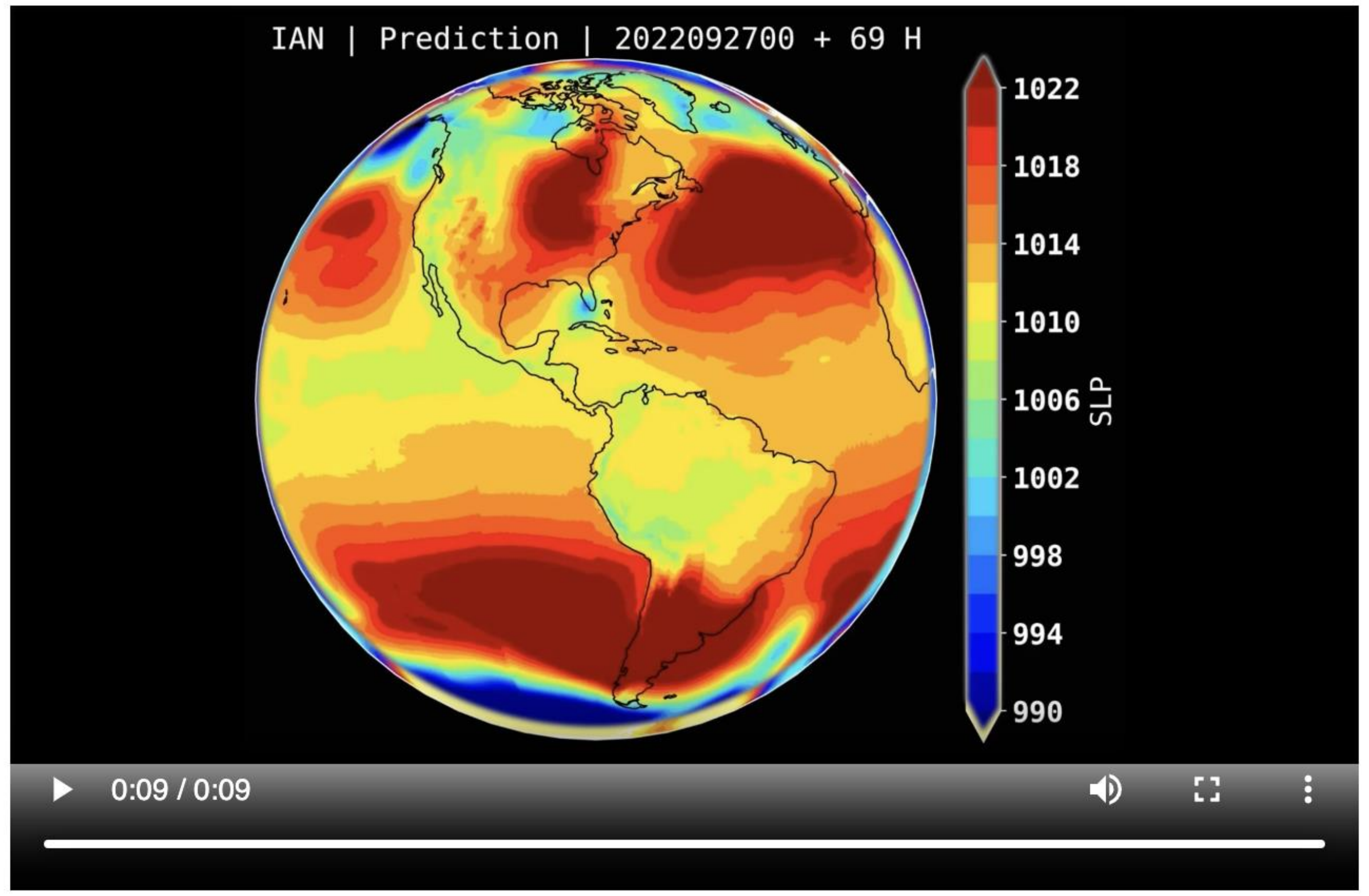
- Uses Randomized iterative forecasting objective
- Added Weather – specific embedding layer
- Pressure-weighted loss-function
- Multi-step fine-tuning

Notable Foundation Models 2023-2024

What is unique about them?







AtmoREP, ECMWF; Julich SC; CERN

Sept 2023 https://arxiv.org/abs/2308.13280

AURORA, Microsoft,

May 2024 https://arxiv.org/pdf/2405.13063v1

Prithvi-WxC, NASA, IBM

May 2024 NASA Blog

Diffusion Models for Downscaling and Evolution of Thunderstorms

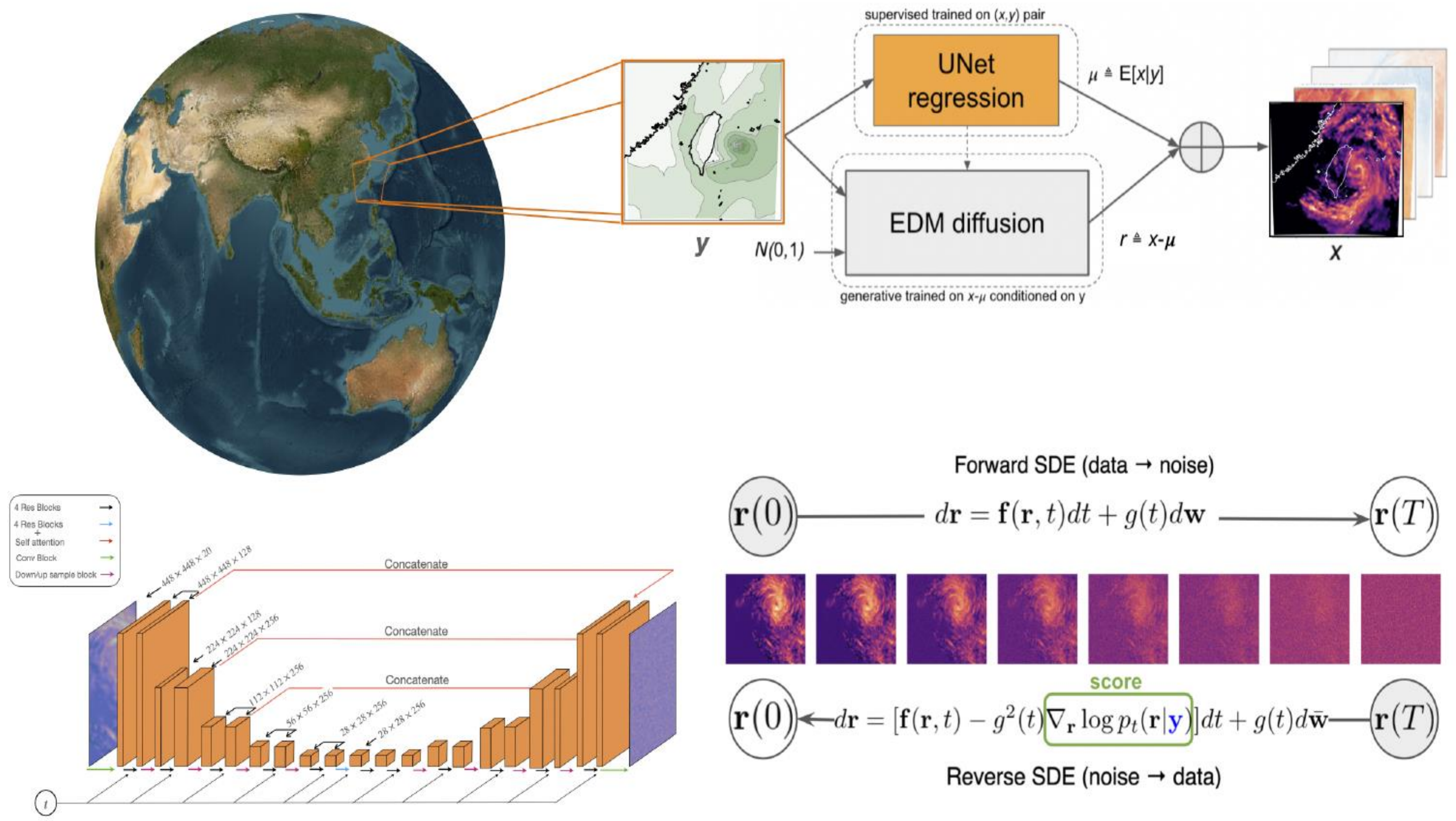
Combination of UNet + Diffusion Model

CorrDiff

12.5x super resolution + radar channel synthesis

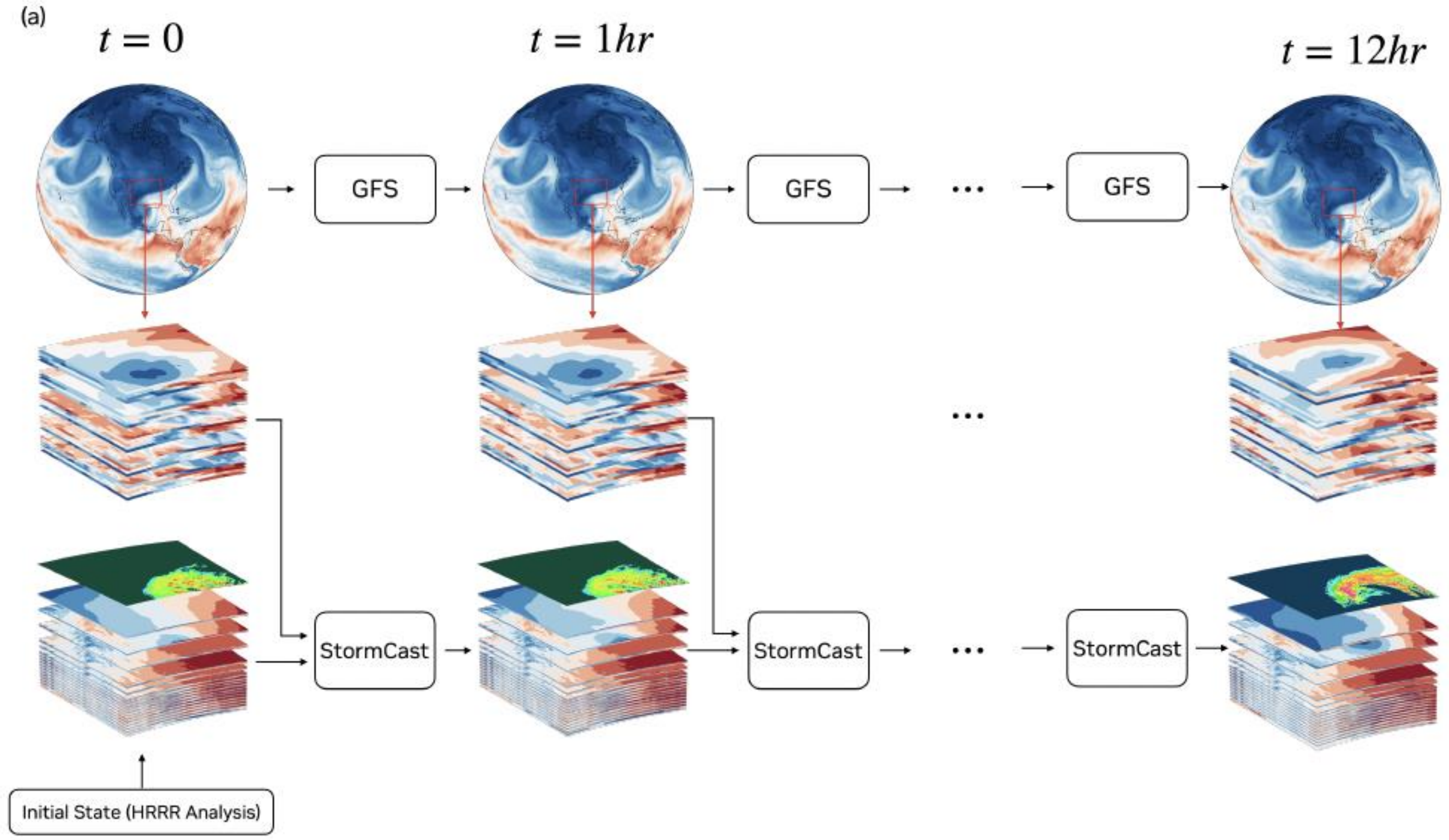
Features: ERA5 : 36x36 (20-ch)

TARGETS: Radar-assimilating WRF 448x448(4-ch)



[Mardani et al. (2023), Generative Residual Diffusion Modeling for Km-scale Atmospheric Downscaling]()

StormCAST



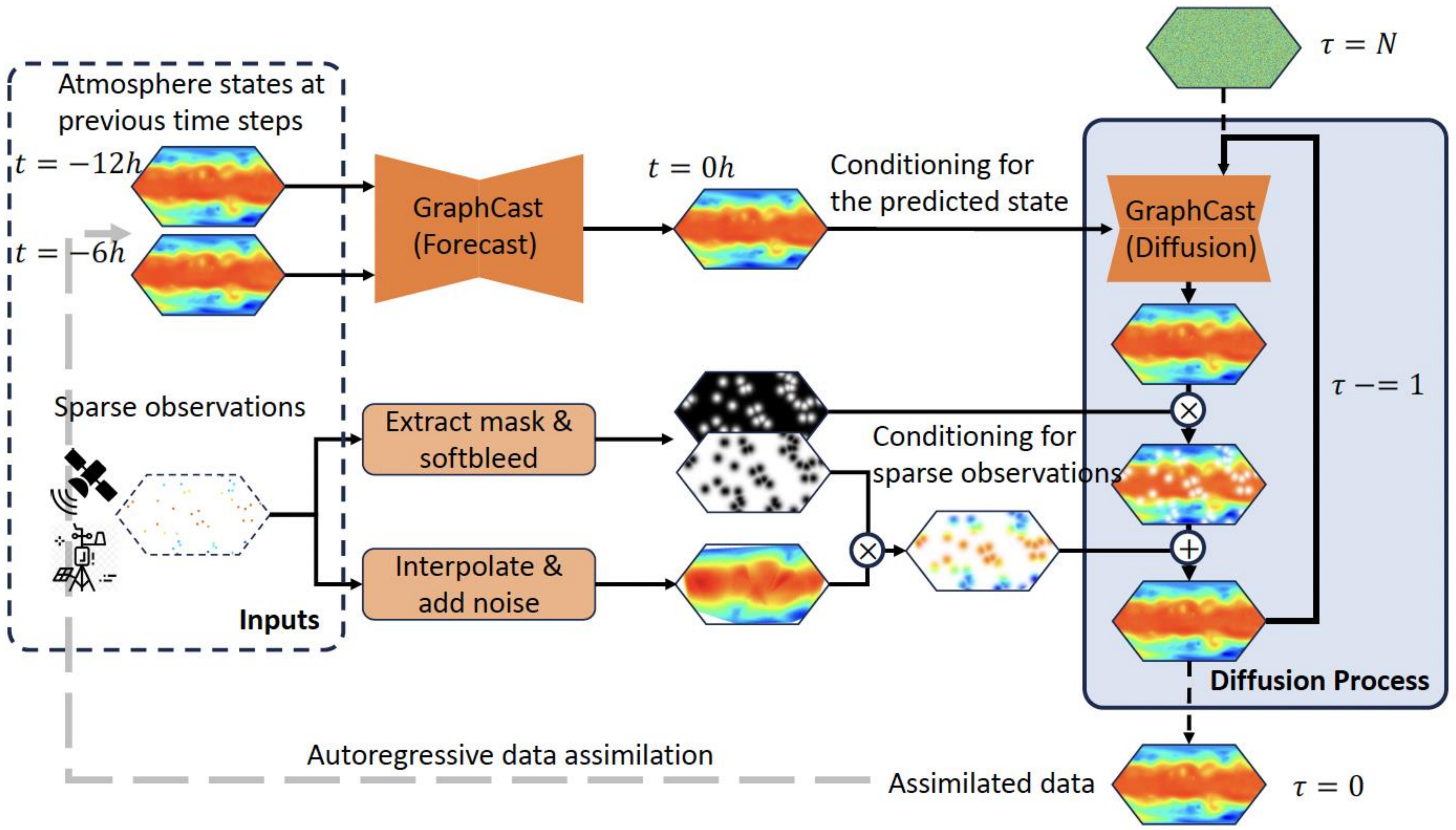
[Kilometer-Scale Convection Allowing Model Emulation using Generative Diffusion Modeling]()

Super-resolution in natural images is much simpler as it involves local interpolation and does not require accounting large spatial shifts, correct biases in static features like topography, and synthesize entirely new channels like radar reflectivity

GEN AI + Data Assimilation

Addressing the Data Assimilation bottleneck needed for the simulation pipeline

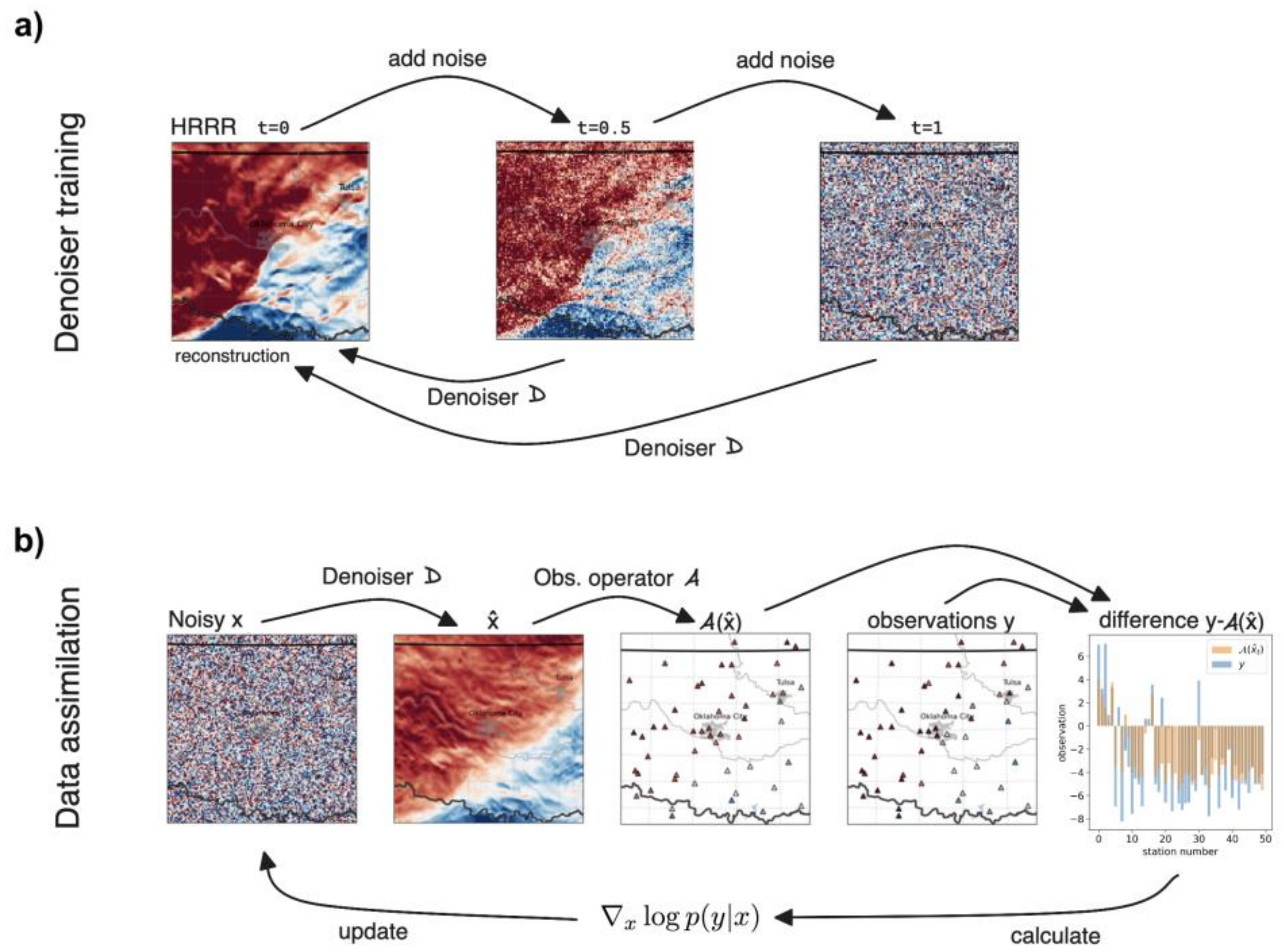
DiffDA: A Diffusion Model for Weather-scale Data Assimilation



Implemented a denoising diffusion model capable of assimilating atmospheric variables using predicted states and sparse observations

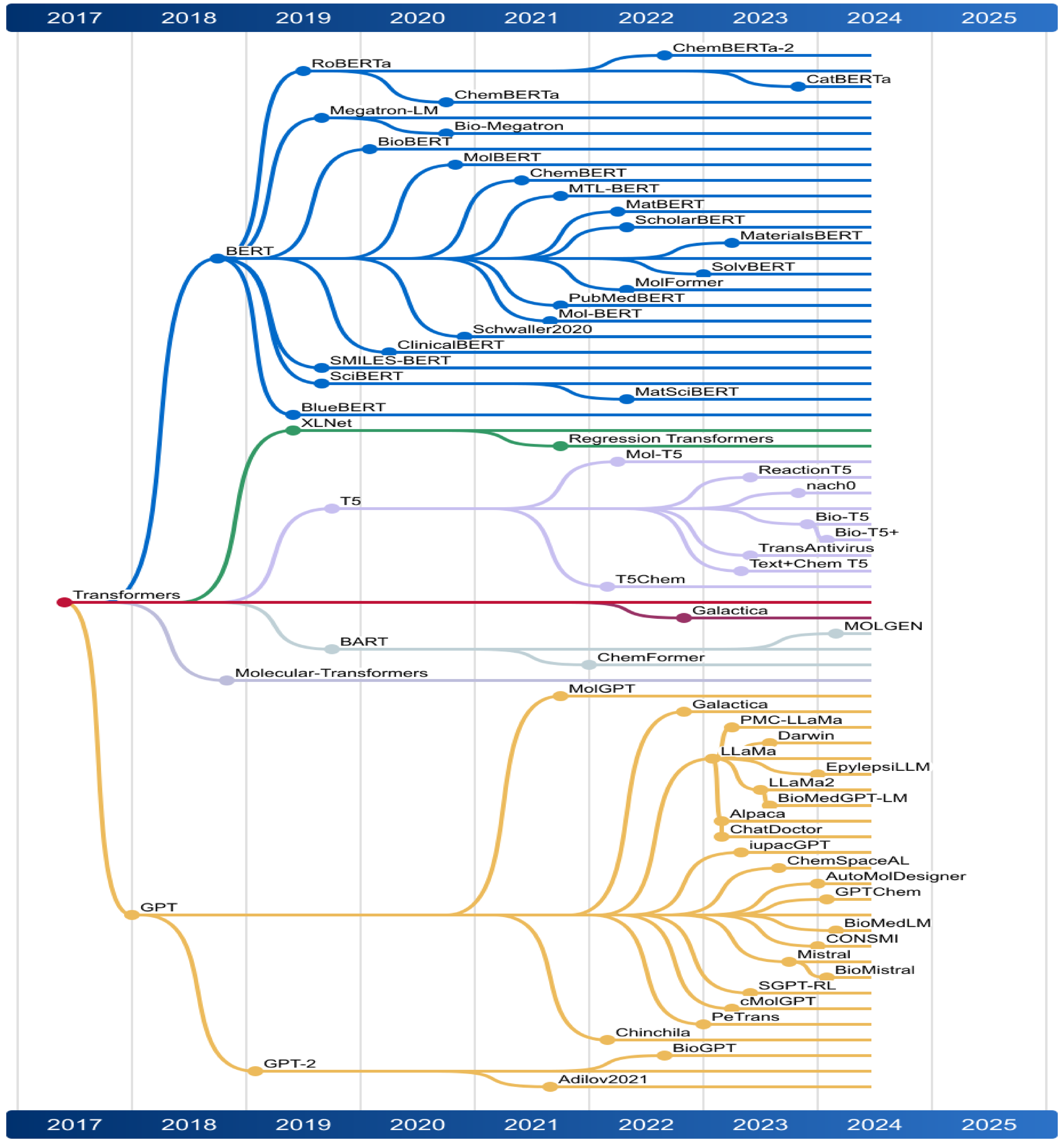
Adapted the pretrained GraphCast neural network as the backbone of the diffusion model.

Generative Data Assimilation of Sparse Weather Station Observations at Kilometer Scales

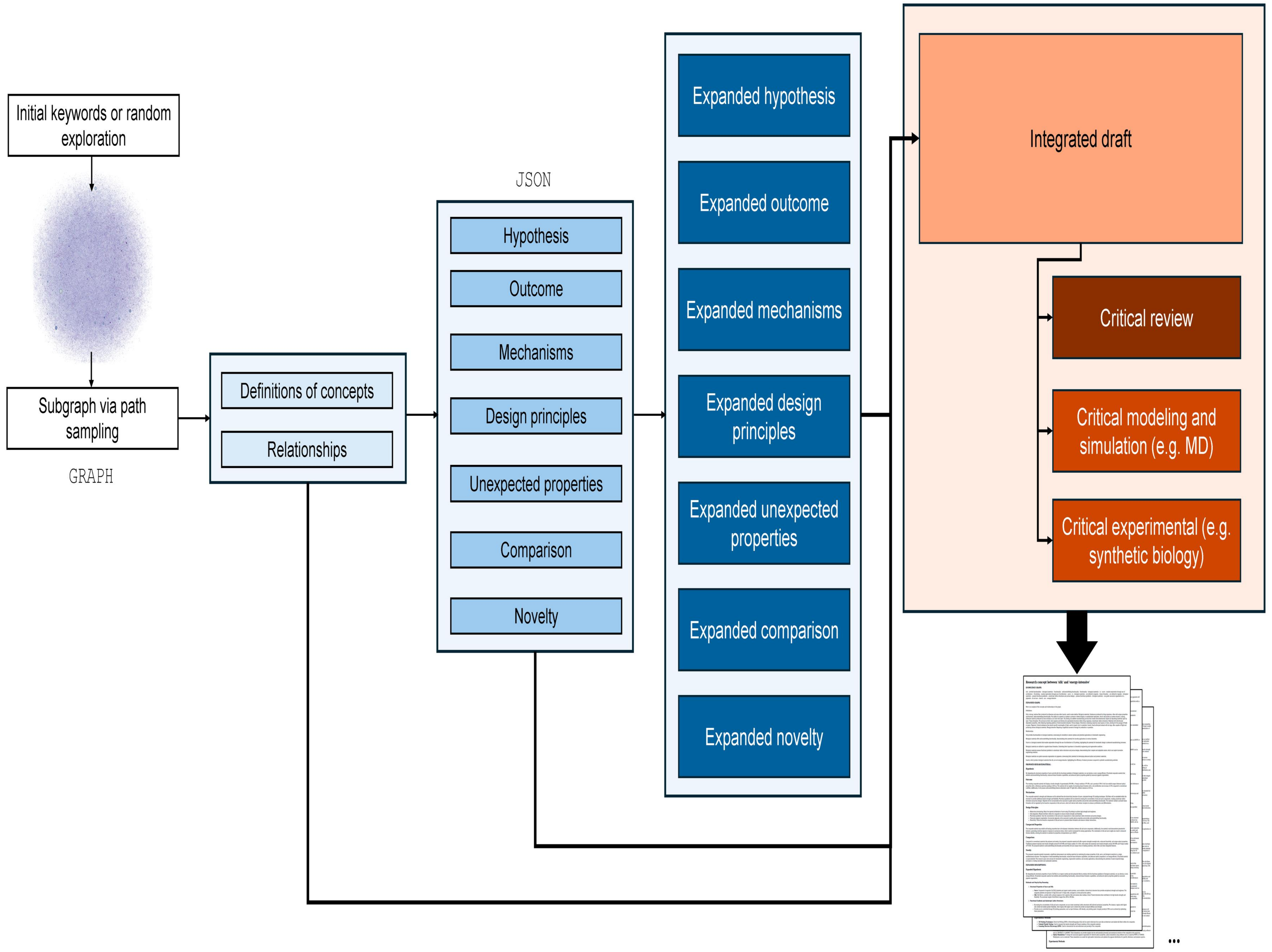


Examples from Material Science and Chemistry

LLMs Virtual Assistants For Chemistry



LARGE LANGUAGE MODELS AND AUTONOMOUS AGENTS IN CHEMISTRY

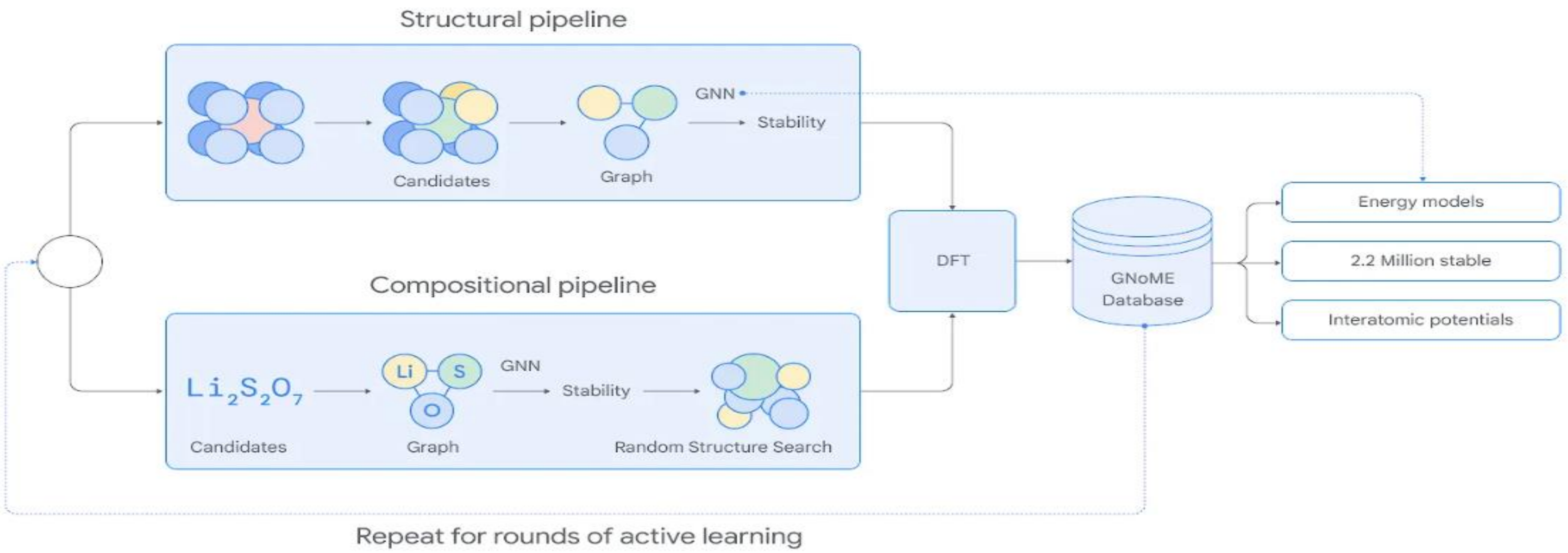


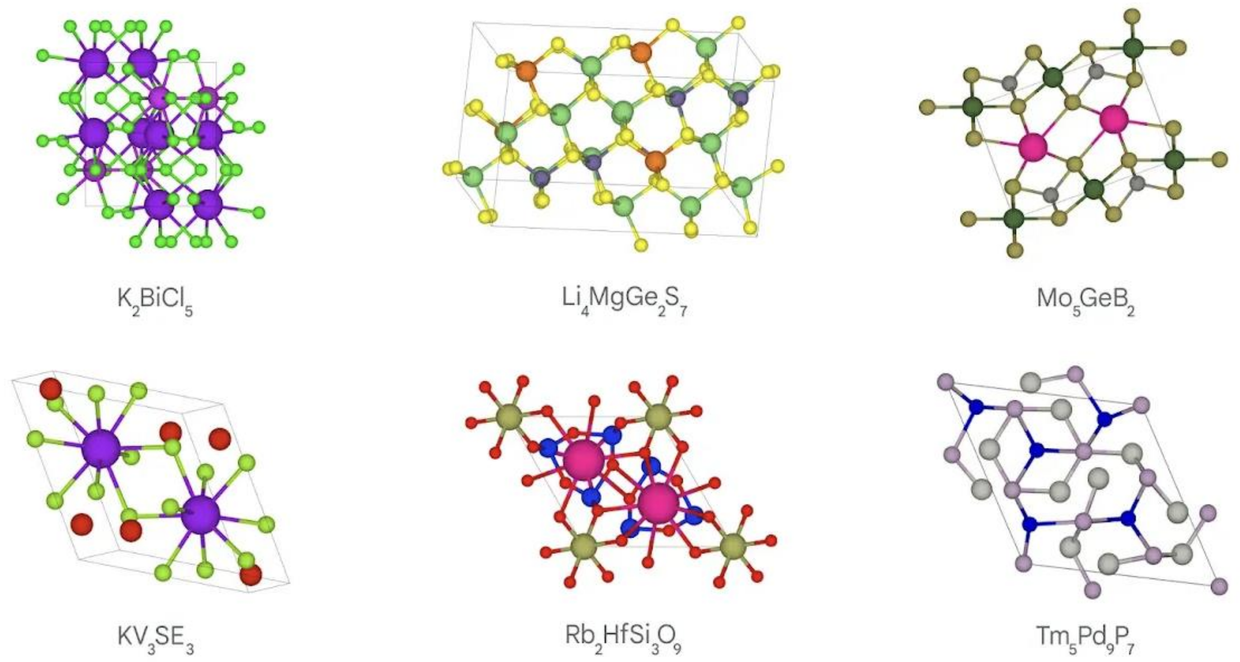
SCIAGENTS: AUTOMATING SCIENTIFIC DISCOVERY THROUGH MULTI-AGENT INTELLIGENT GRAPH REASONING

Source: A REVIEW OF LARGE LANGUAGE MODELS AND AUTONOMOUS AGENTS IN CHEMISTRY

GNoMe : Using GNNs for Materials Exploration

An AI tool that dramatically increases the speed and efficiency of discovery by predicting the stability of new materials.





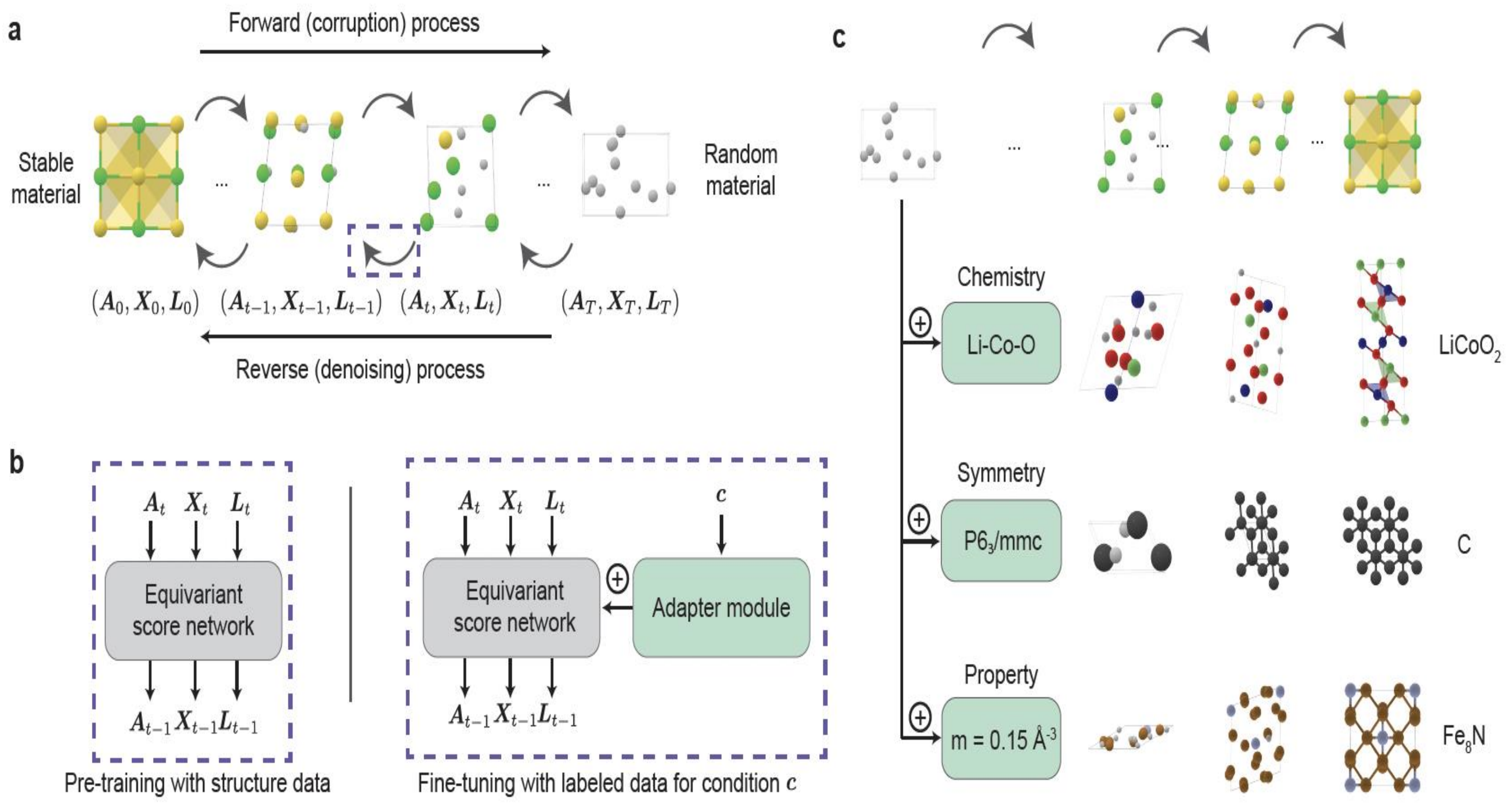
Six examples ranging from a first-of-its-kind Alkaline-Earth Diamond-Like optical material (Li4MgGe2S7) to a potential superconductor (Mo5GeB2)

A-Lab, a facility at Berkeley Lab where artificial intelligence guides robots in making new materials. Photo credit: Marilyn Sargent/Berkeley Lab

- Trained on data on on crystal structures and their stability, openly available through the Materials Project
- Used GNoMe to generate novel candidates and predict stability
- Used DFT simulations to periodically cross-checked the performance via active learning
- Achieved materials stability prediction from around 50%, to 80% - based on MatBench Discovery

MatterGen

A Generative model for inorganic materials design

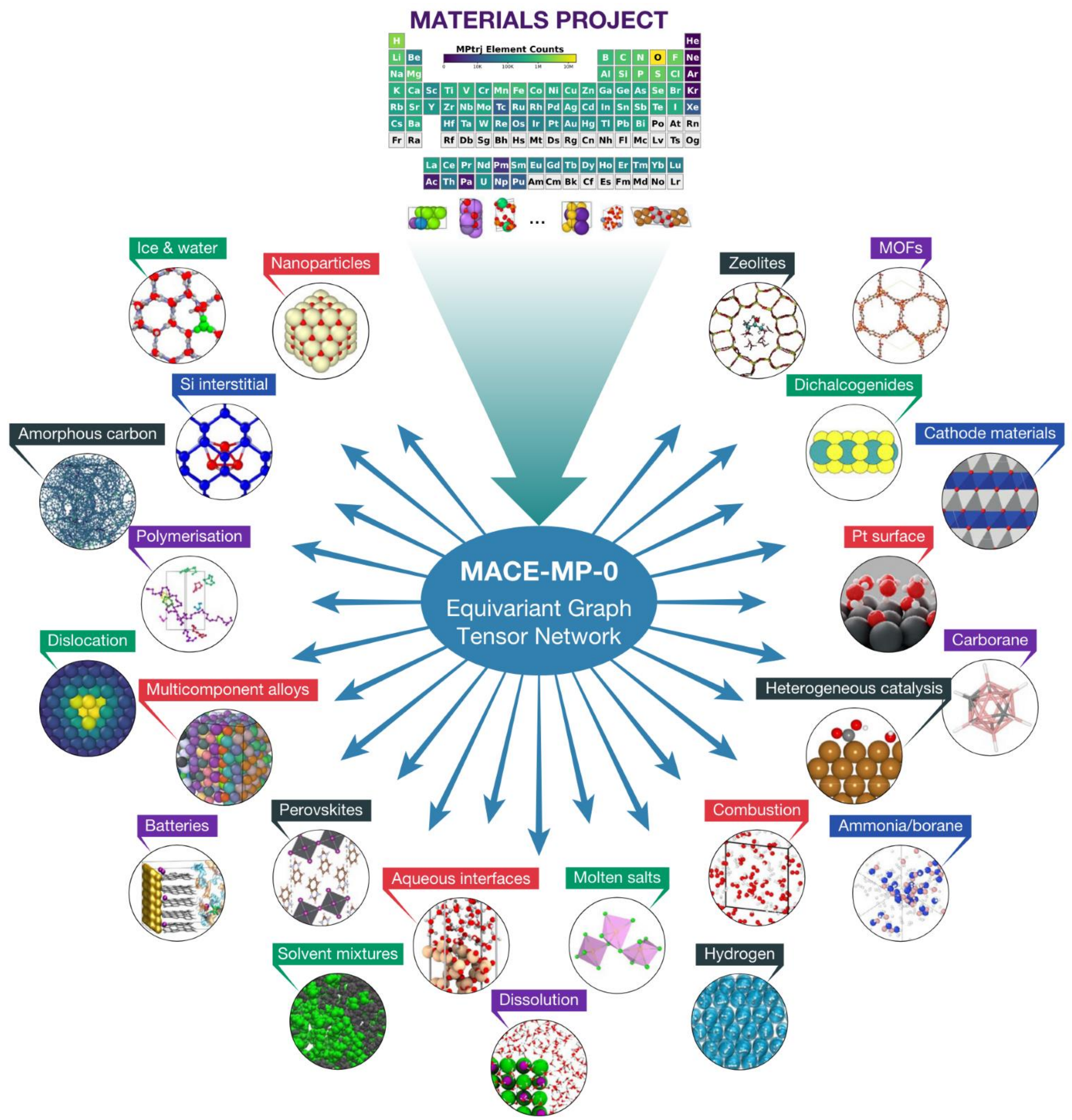


- A diffusion based generative model for designing stable inorganic materials across the periodic table
- With adapter modules it can be steered to generate materials with desired properties.
- Doubles the rate of S.U.N. materials that15x closer to ground truth structures at the DFT local energy minimum.

Dataset used for base model: *607,684 stable structures with up to 20 atoms recomputed from the Materials Project*
Energy per atom after relaxation: below 0.1 eV/atom (via DFT)
Structure is novel if it's not in the MP, Alexandria, and Inorganic Crystal Structure Database (ICSD) datasets

MACE-MP

A foundation model for atomistic materials chemistry



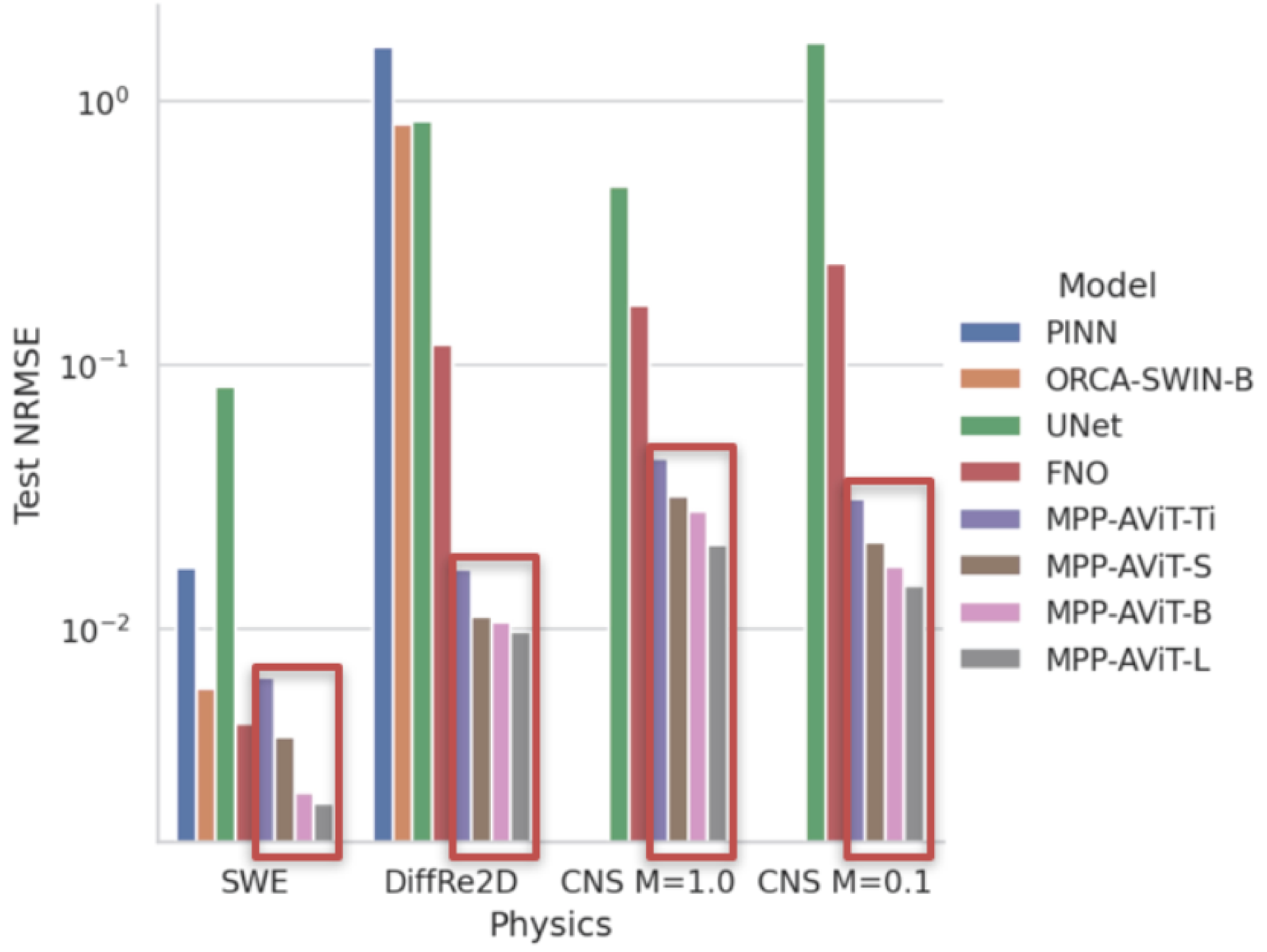
- A single general-purpose ML model, trained on a public database of 150k inorganic crystals, that is capable of running stable molecular dynamics on molecules and materials.
- #1 open model on MatBench
- Using MACE-MP0, A single NVIDIA A100 GPU with 80GB of RAM, it can do several nanoseconds per day for 1000 atoms.
  - The performance depends on the atomic density, hardware floating point precision, size of model
- World-wide collaboration of researchers and SCC sites

Examples from Physics

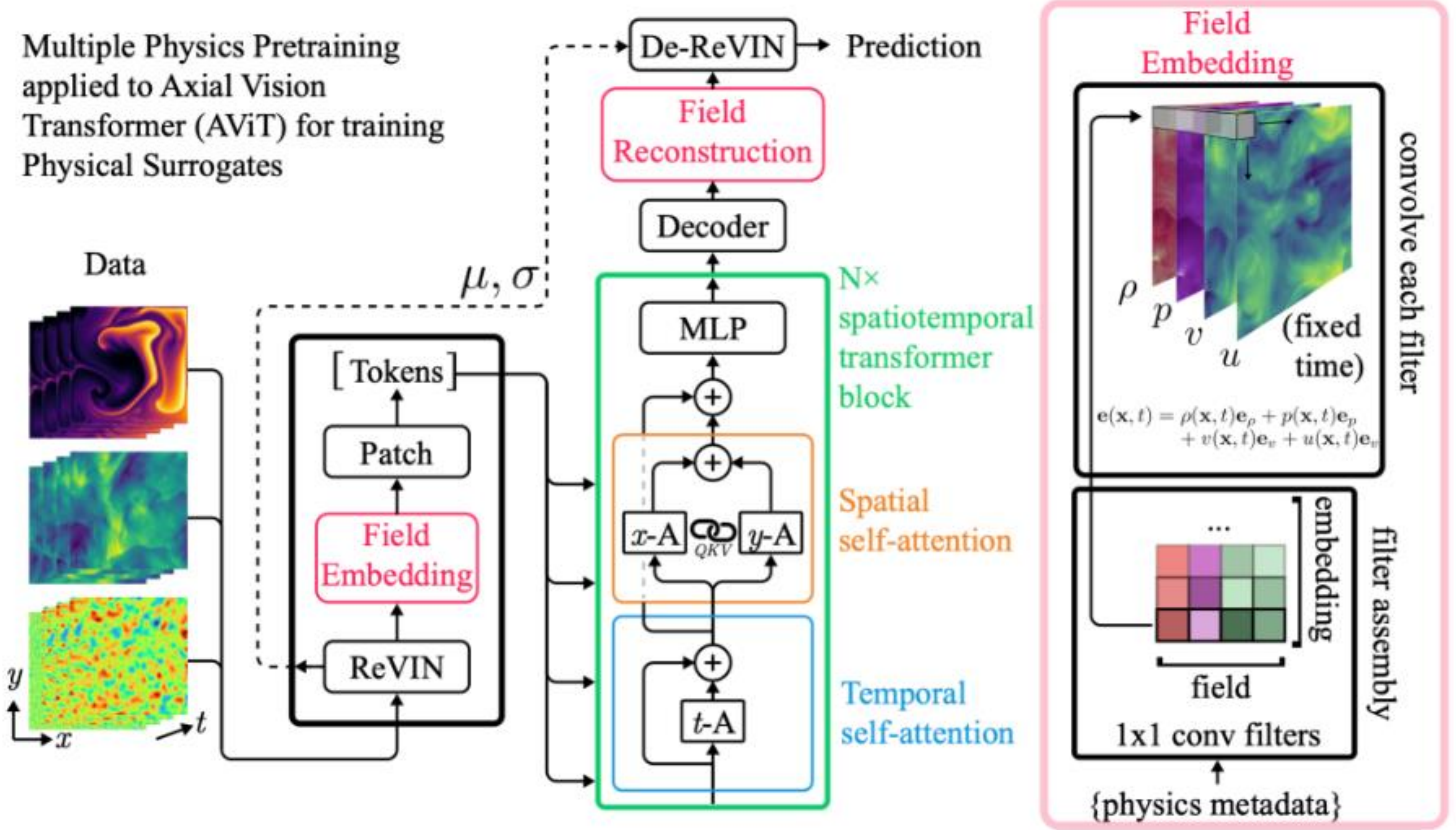
Multiple Physics Pretraining for Physical Surrogate Models

- **Training a transformer-based surrogate model on multiple different physical systems** from PDEBench **outperforms specialized baselines** trained on single physical systems.

- Multiple physics pretraining transfers more effectively to new physics through fine-tuning.



Comparison of test MSE on different physics (lower is better)

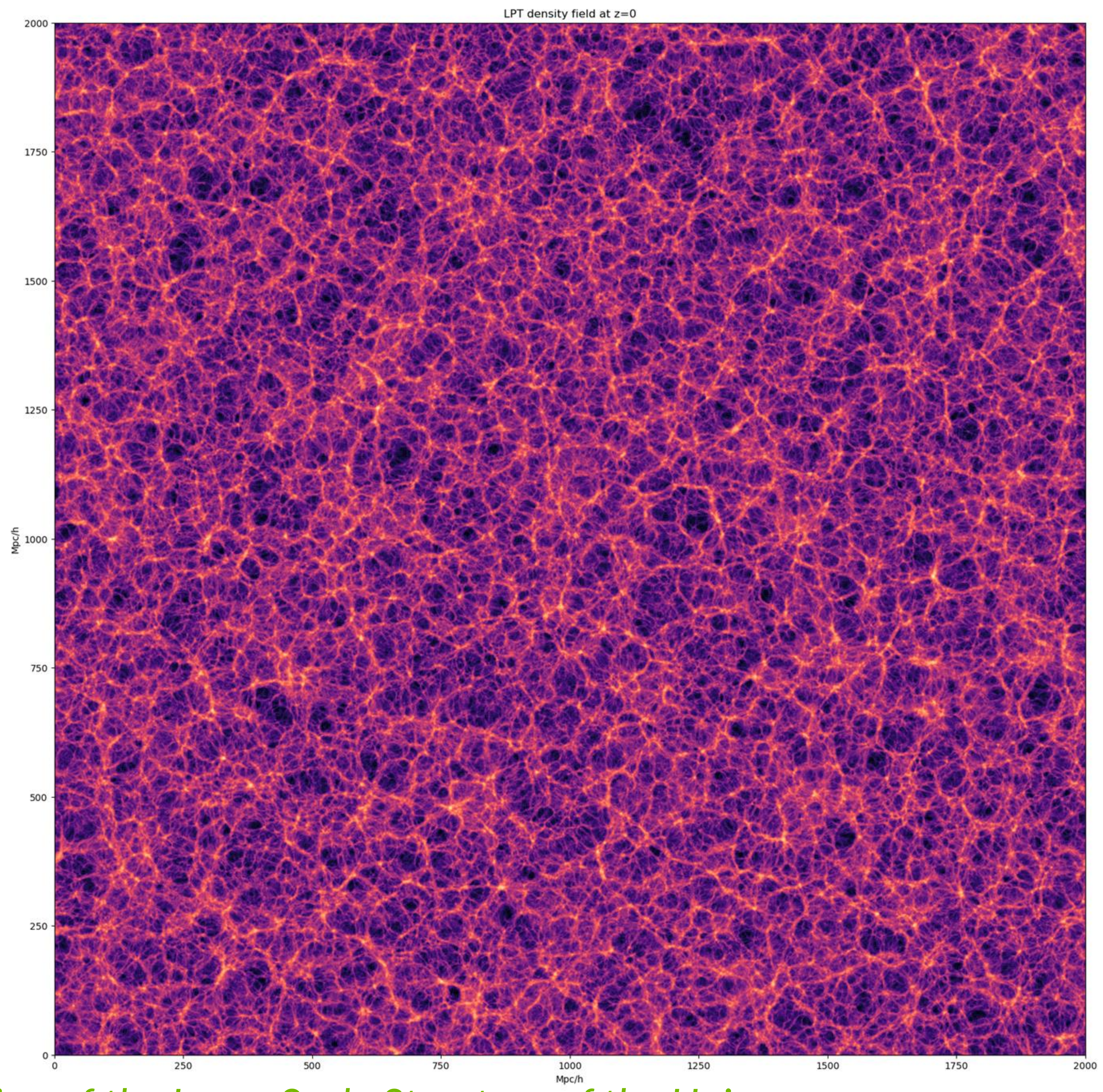


FLATIRON INSTITUTE

Polymathic

Distributed and Differentiable N-body Simulations in JAX powered by cuDecomp

- jaxDecomp: JAX primitive bindings for the NVIDIA cuDecomp **adaptive pencil decomposition library**.
- Enables for the first time the implementation of large-scale and **automatically differentiable N-body simulators** for GPU-based supercomputers.
  - Unlocks the possibility of performing optimization and high-dimensional inference over simulation models, which require backpropagating through these numerical simulations.



*Cosmological simulation of the Large Scale Structure of the Universe on a **2048^3 mesh distributed on 24 A100 GPUs, runs in 4.7s***

FLATIRON INSTITUTE

NVIDIA

AstroCLIP: Cross-Modal Pre-Training for Astronomical Foundation Models

- AstroCLIP is an extension to astrophysics of the CLIP (Contrastive Language Image Pretraining) strategy to **build semantically aligned embeddings** of diverse data modalities (here astronomical images and optical spectra).
- It is the **first multi-modal foundation model for astrophysics**.
- AstroCLIP embeddings extract meaningful physical information, which can be used as very informative features for downstream tasks.

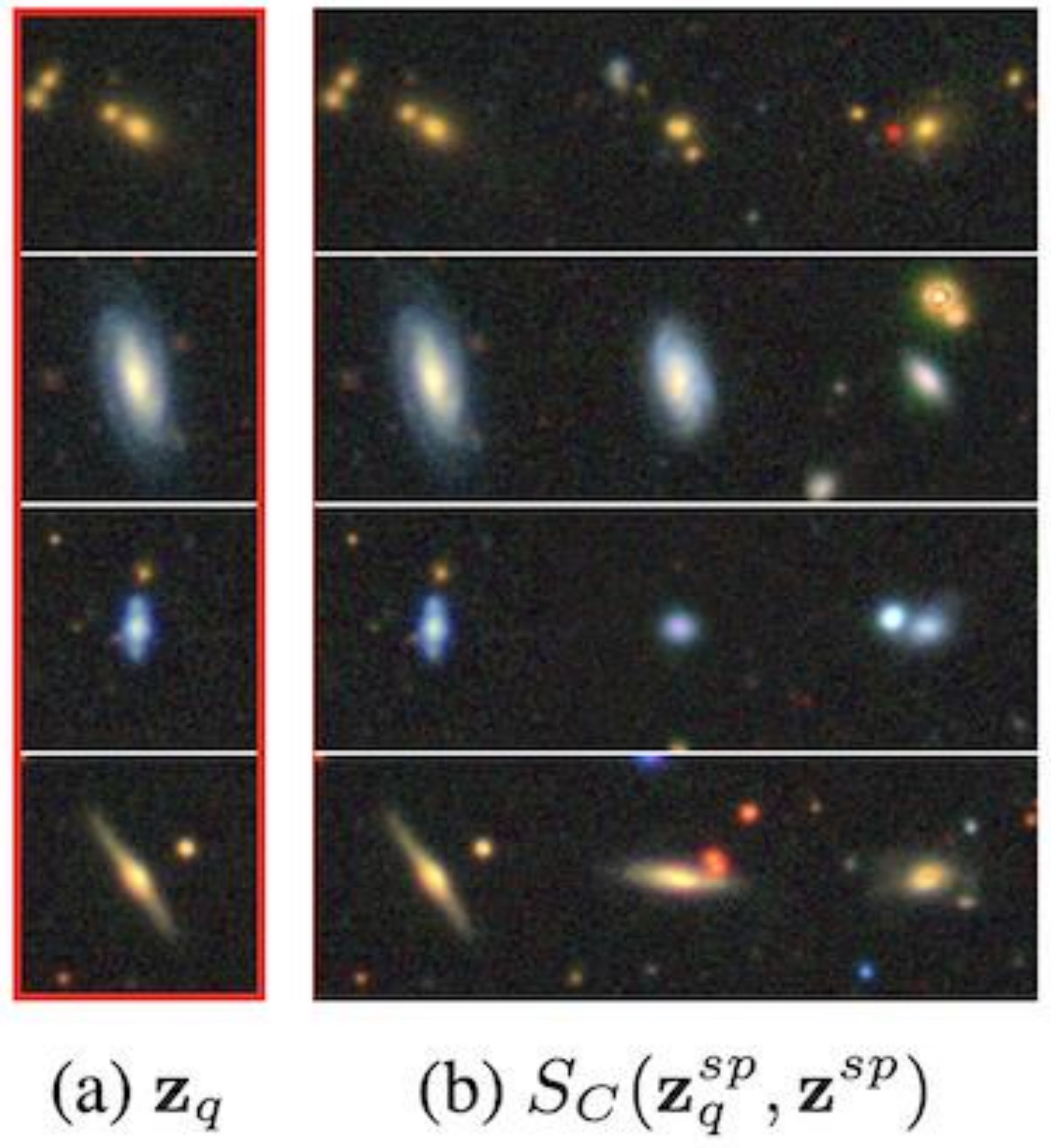
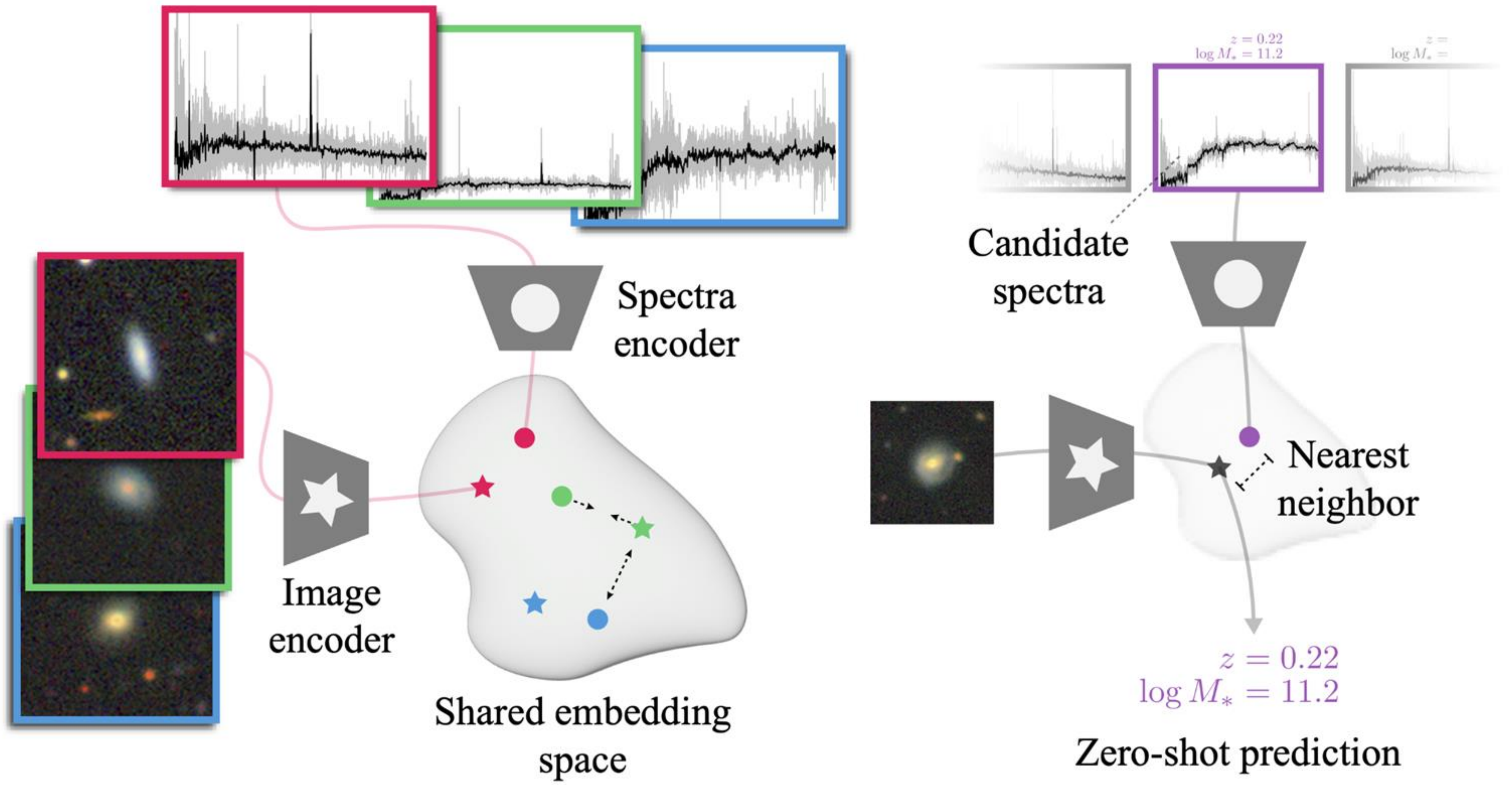


Illustration of retrieval by cosine similarity. Left column shows query objects, right columns shows retrieved objects.



Left: illustration of contrastive training strategy used to train the image and spectra encoders. Right: Once trained, the model can be used to infer physical properties of galaxies simply by nearest neighbour regression.

FLATIRON INSTITUTE

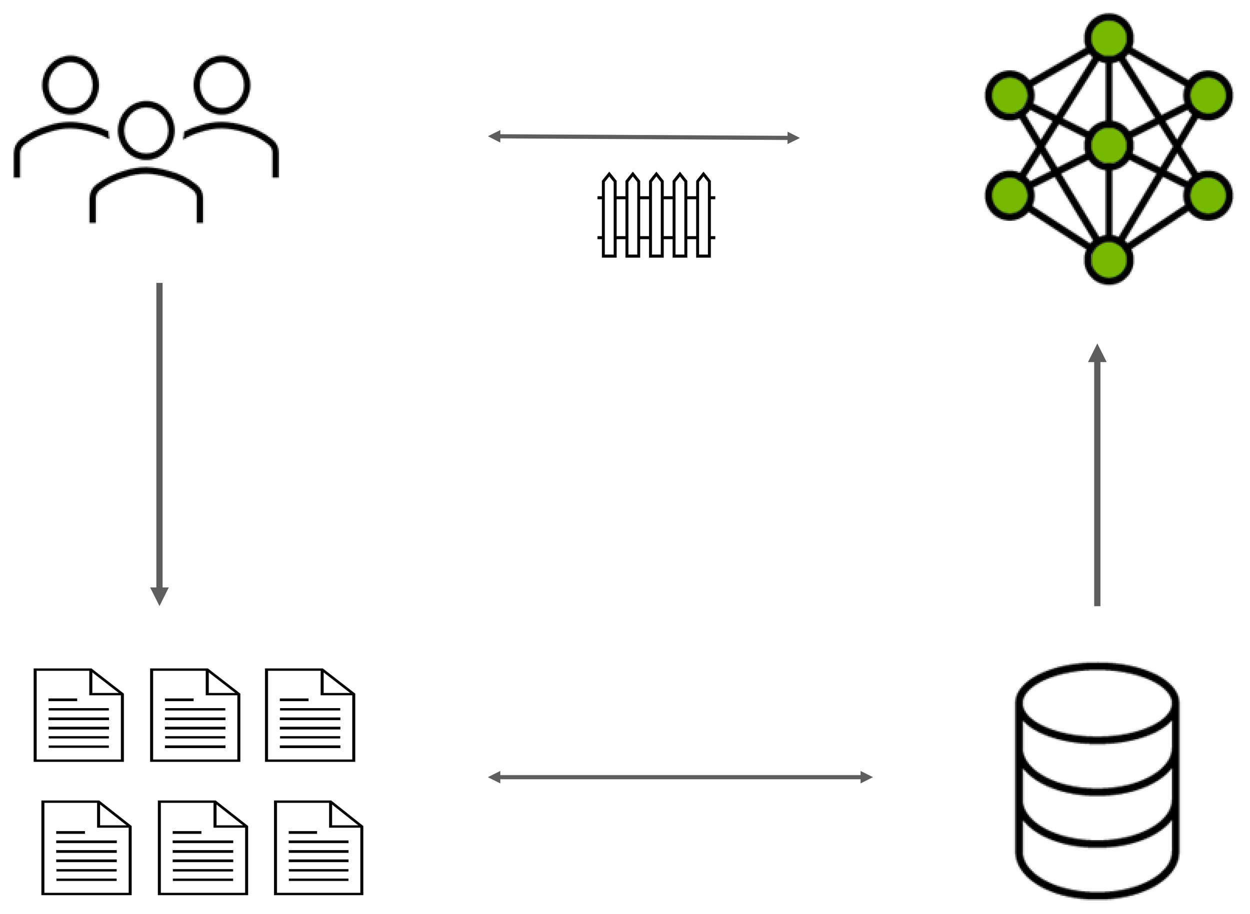
Polymathic

NVIDIA

Intersection of Gen AI and Science

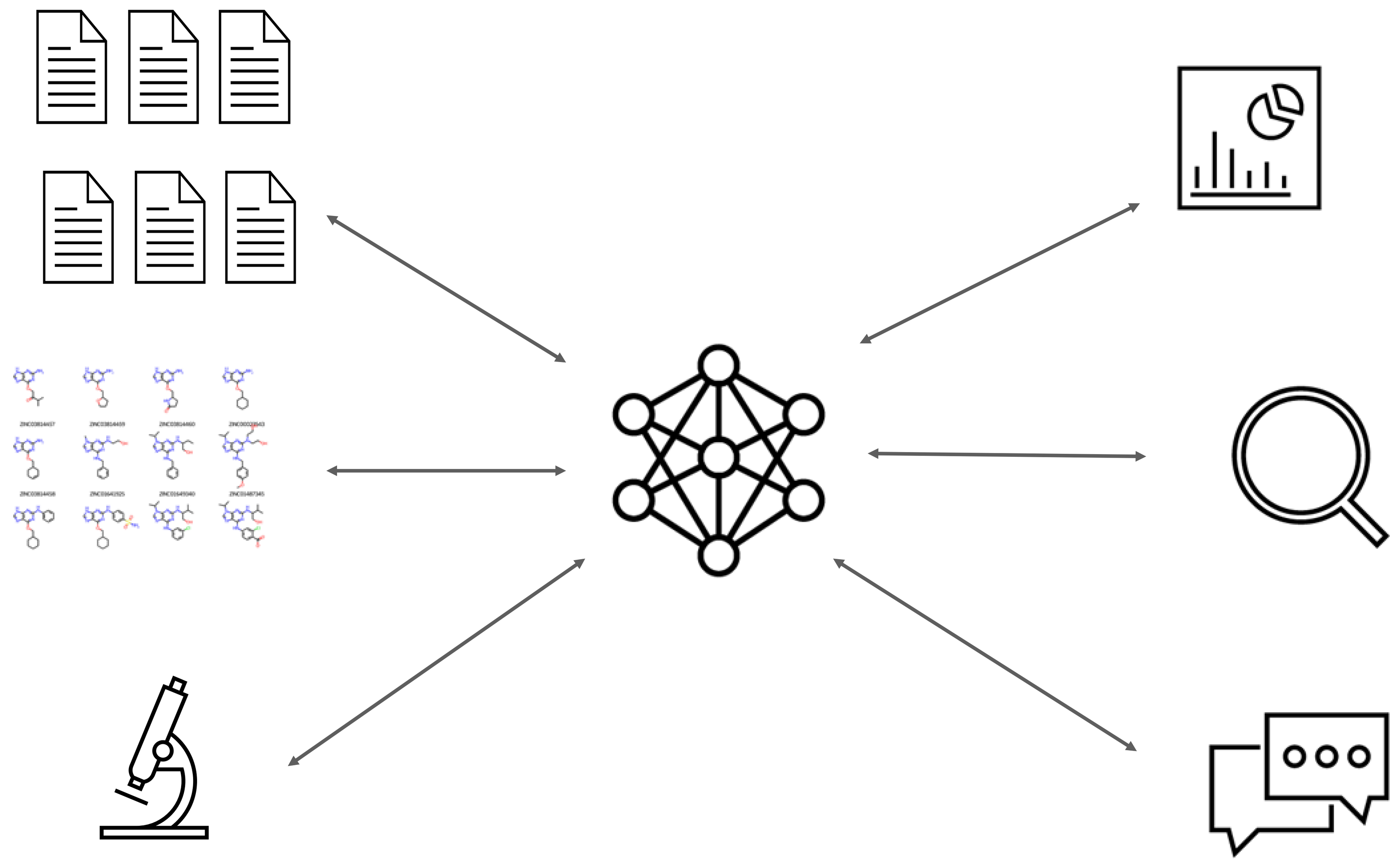
3 Distinct Categories

Summarize



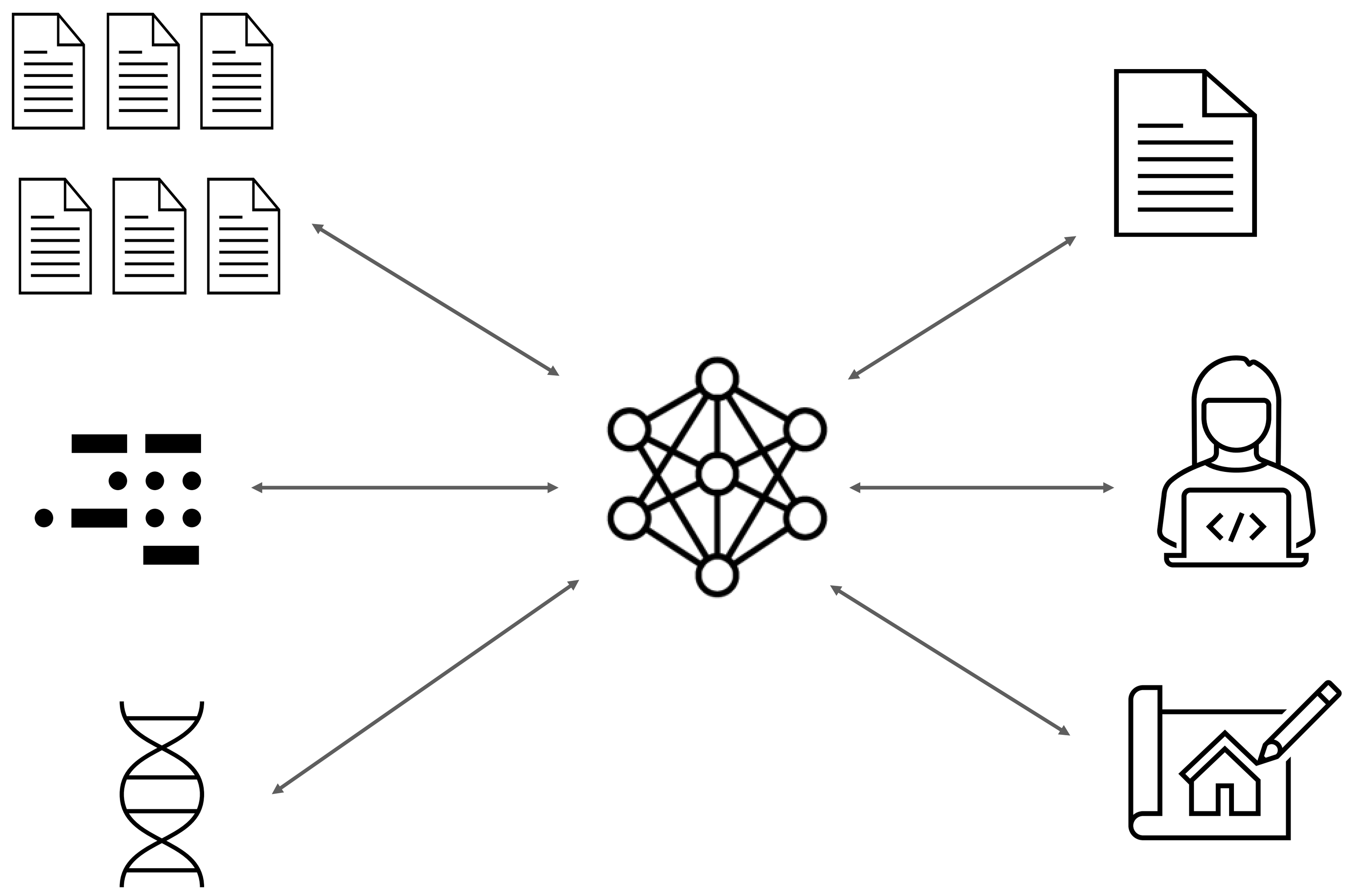
OTS LLMs
RAG
Guardrails

Synthesize



OTS LLM
Multiple Data Sources
Customization/Tuning
Guardrails
RAG

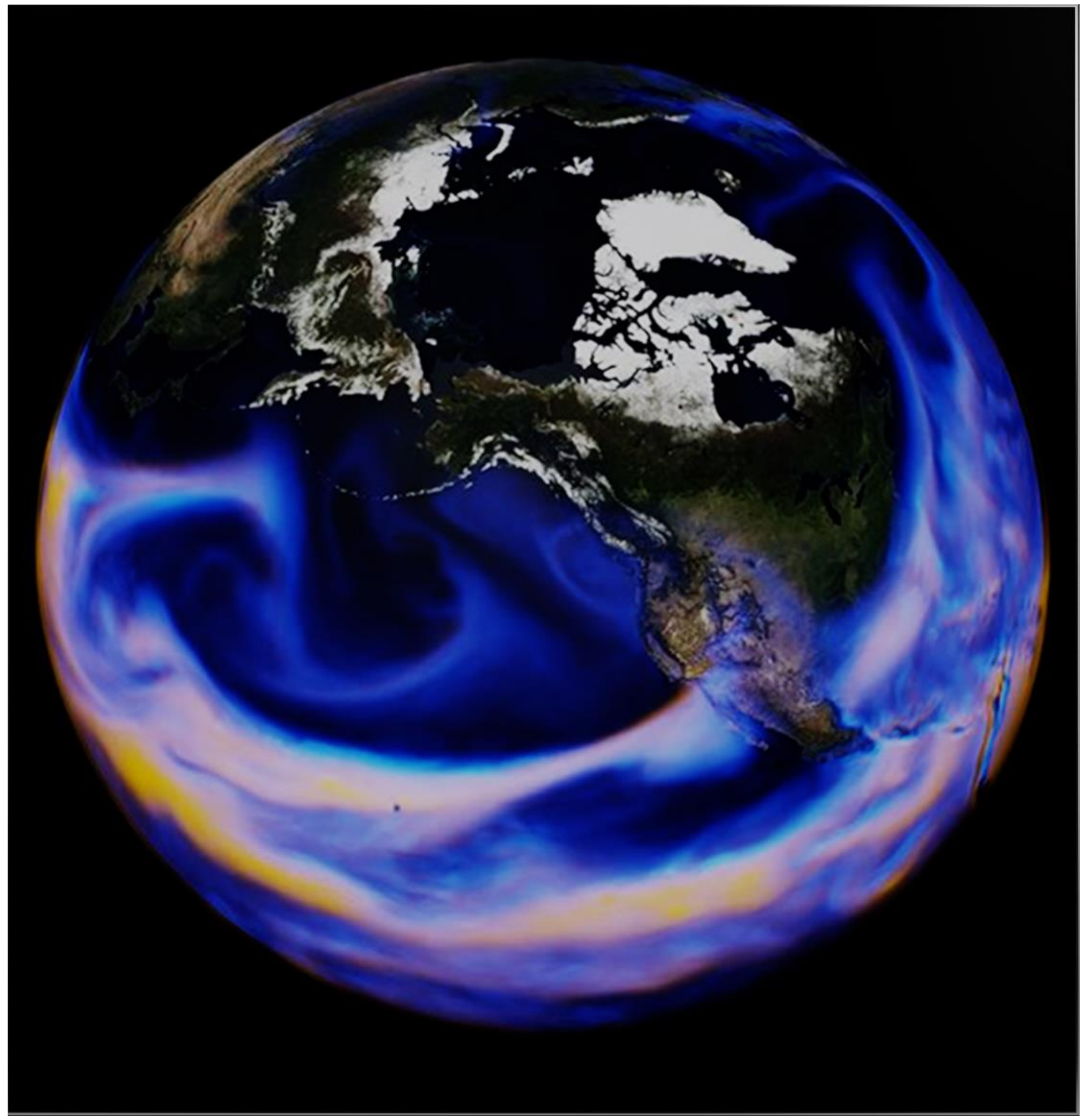
Generate



LLM from Scratch
Multiple Data Sources,
Customization/Tuning
Guardrails
RAG

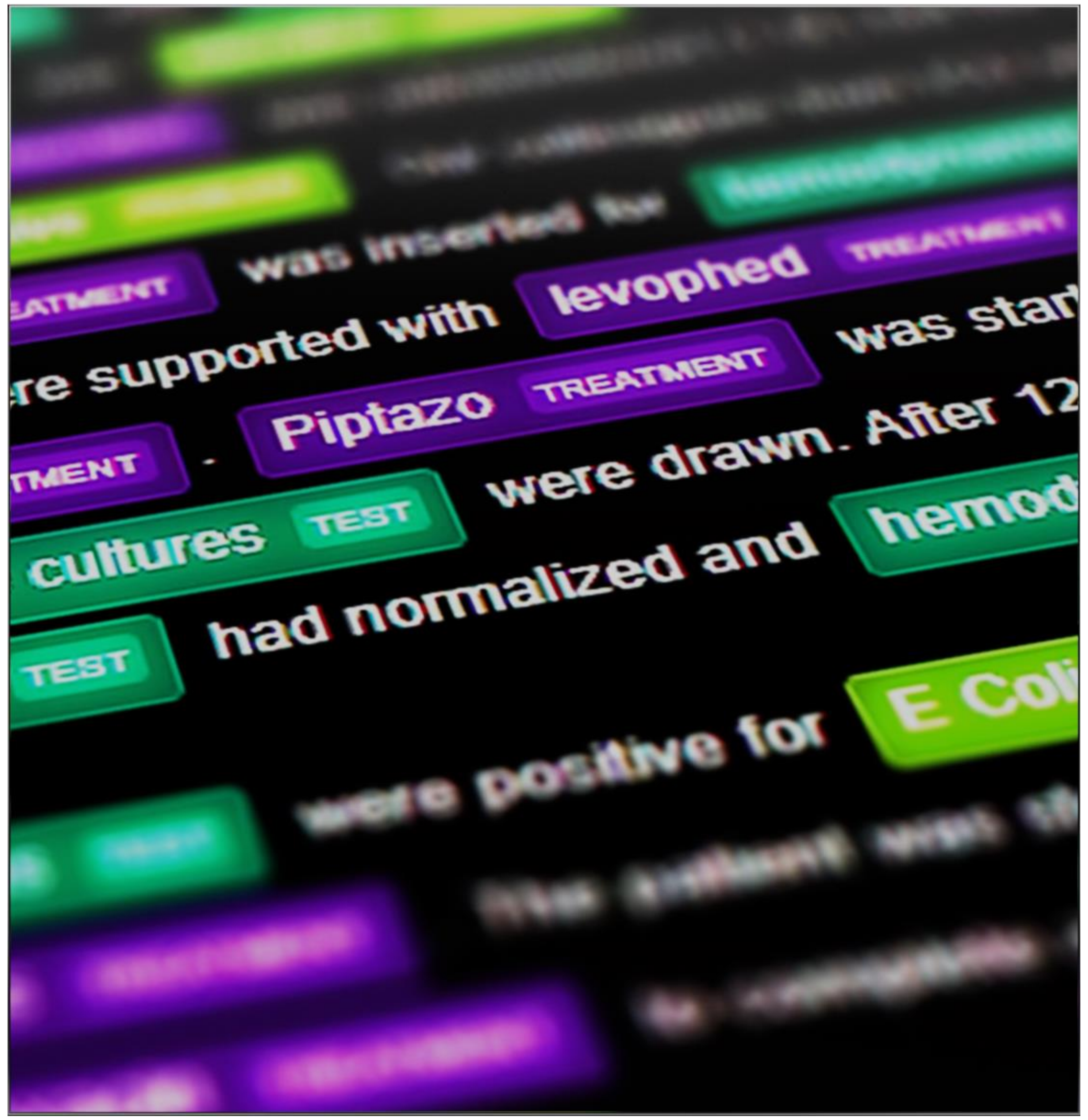
Projects for Science Community to collaborate

MODULUS
Physics-ML Model Training and Inference



https://github.com/NVIDIA/modulus

NEMO FRAMEWORK
Developing Scientific Foundational Models at Scale



https://github.com/NVIDIA/NeMo

Lots more to do…