Argonne NATIONAL LABORATORY

Serving scientific foundation models on leadership computing platforms

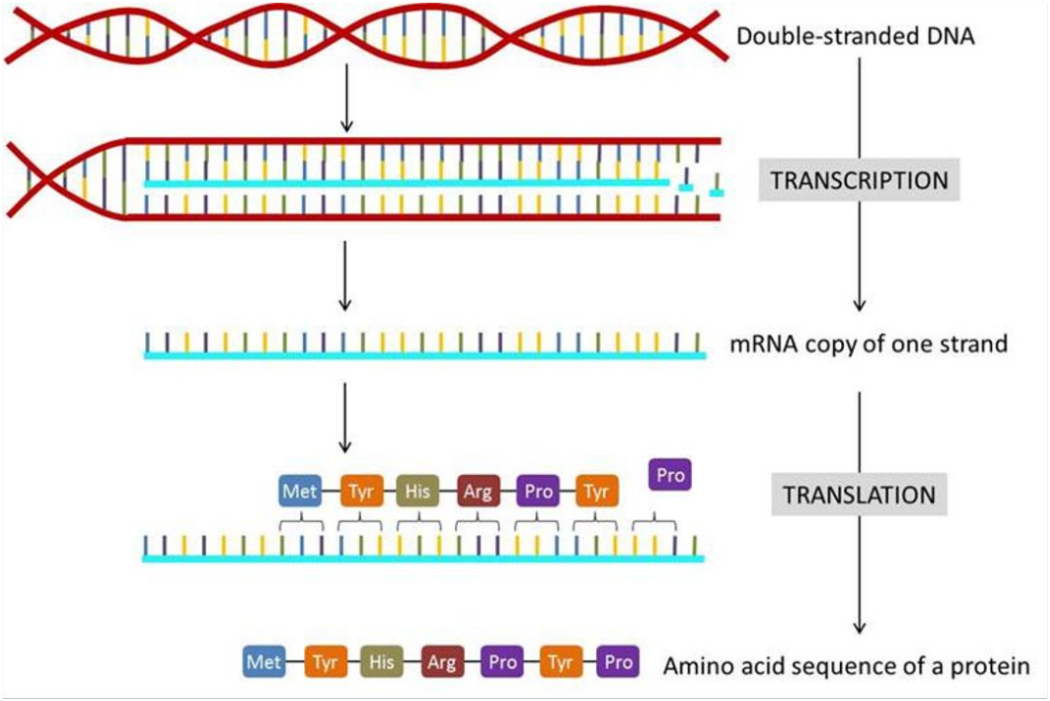
Alexander Brace[1,2*], Arvind Ramanathan[1,2✢], Ian Foster[1,2✢]

[1]Data Science and Learning Division, Argonne National Laboratory, [2]University of Chicago; ✢Senior authors *First authors

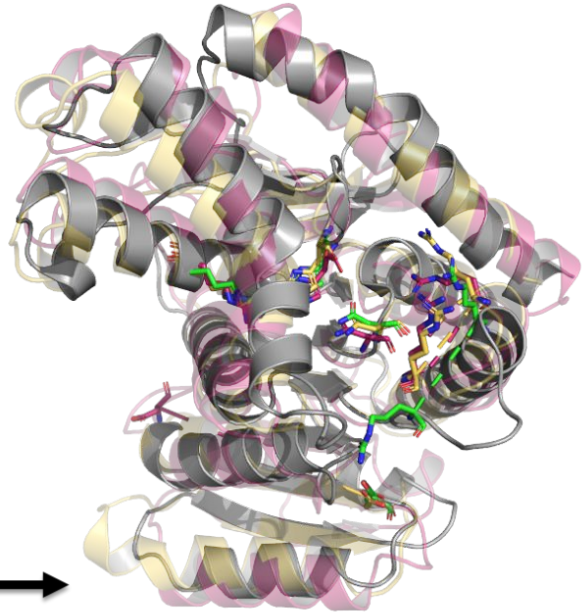
U.S. DEPARTMENT OF ENERGY Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

ParslFest, Oct 19, 2023

Central Dogma of Molecular Biology

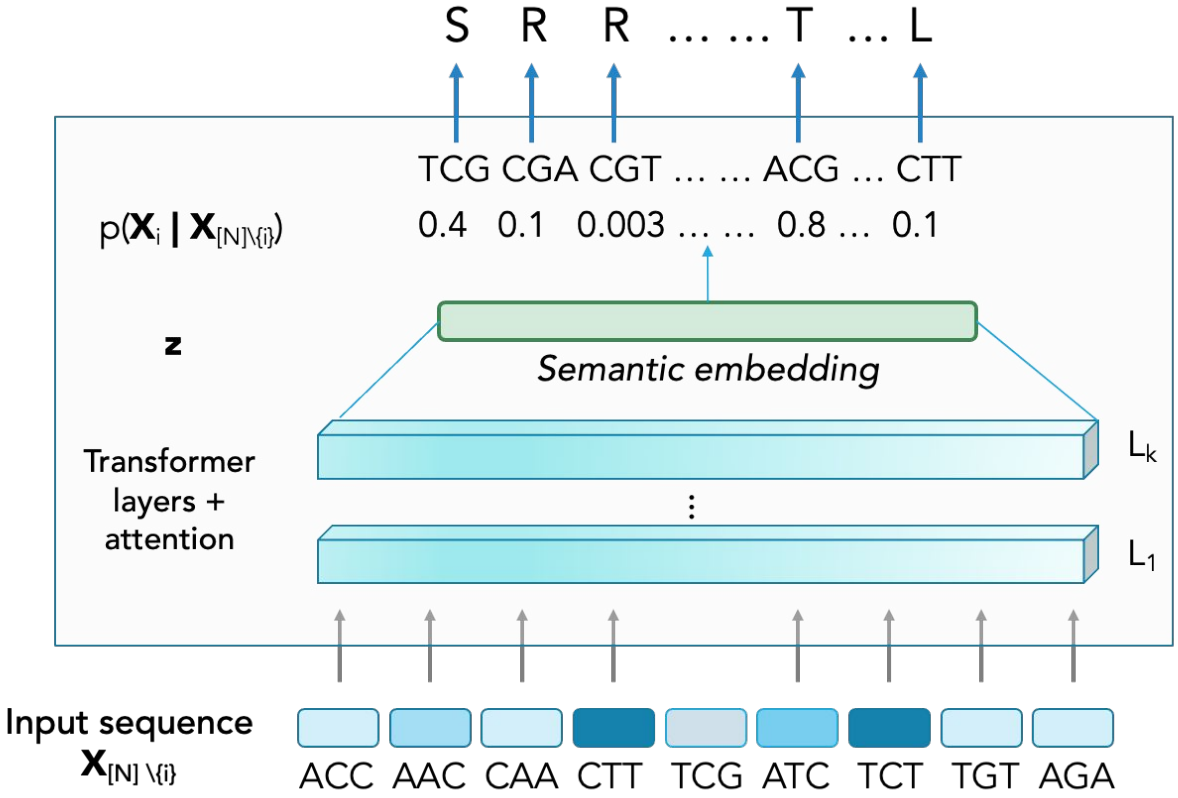


Fold



https://rwu.pressbooks.pub/bio103/chapter/the-central-dogma-genes-to-traits/

Genome-scale Language Models (GenSLMs)

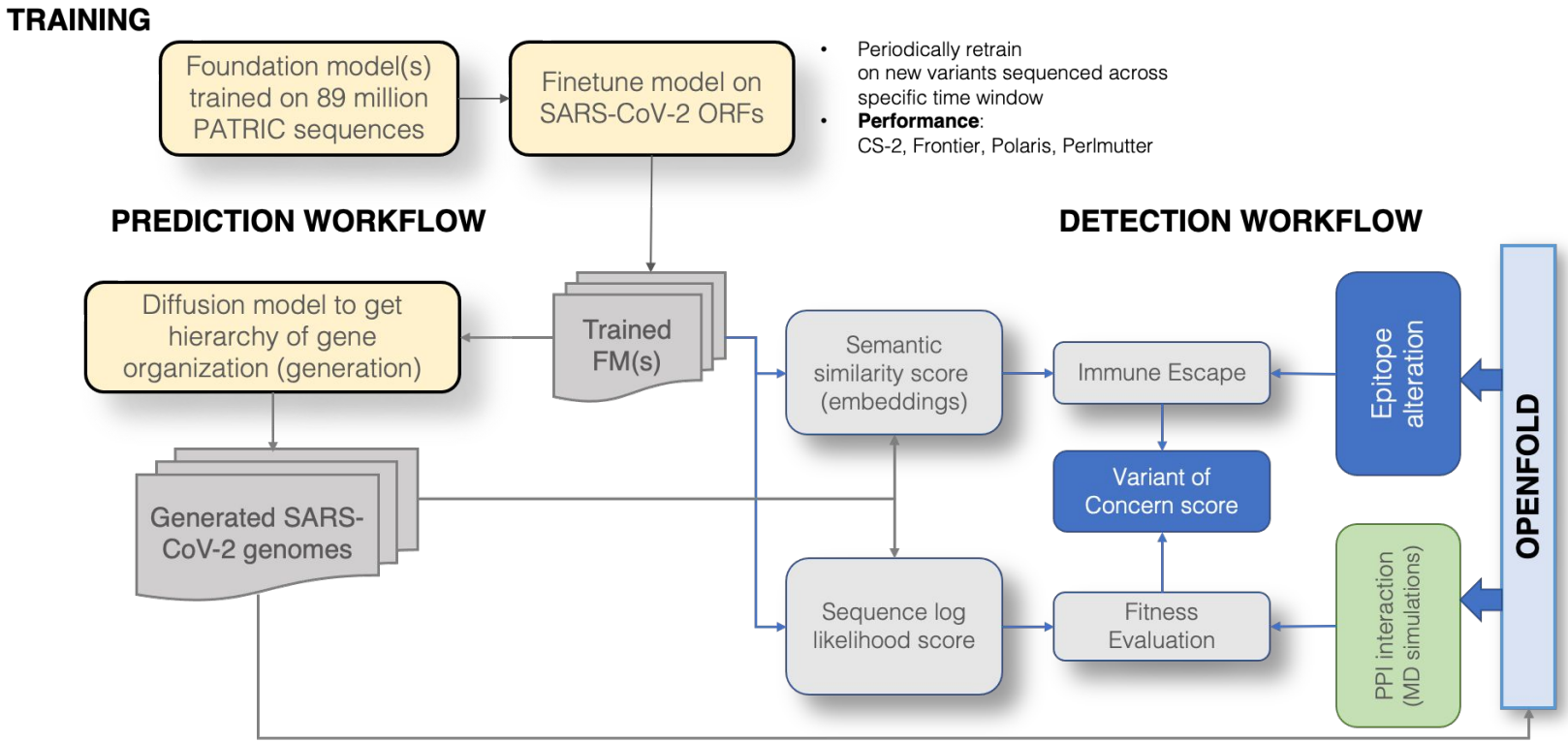


Model	Seq. length	#Parameters	Dataset
GenSLM-Foundation	2048	25M, 250M, 2.5B, **25B**	110M
GenSLM	10240	25M, 250M, 2.5B, **25B**	1.5M
GenSLM-Diffusion	10240	2.5B	1.5M

- Scaling LLMs with 25B parameters:
  - O (L^2) complexity in the attention computation
  - overcome communication overheads, parameters, checkpointing
- Variation within SARS-CoV-2 sequences can be small (< 1% overall variation)
  - Need foundation model to accommodate diversity
- **One of the largest foundation model trained on raw nucleotide sequences**

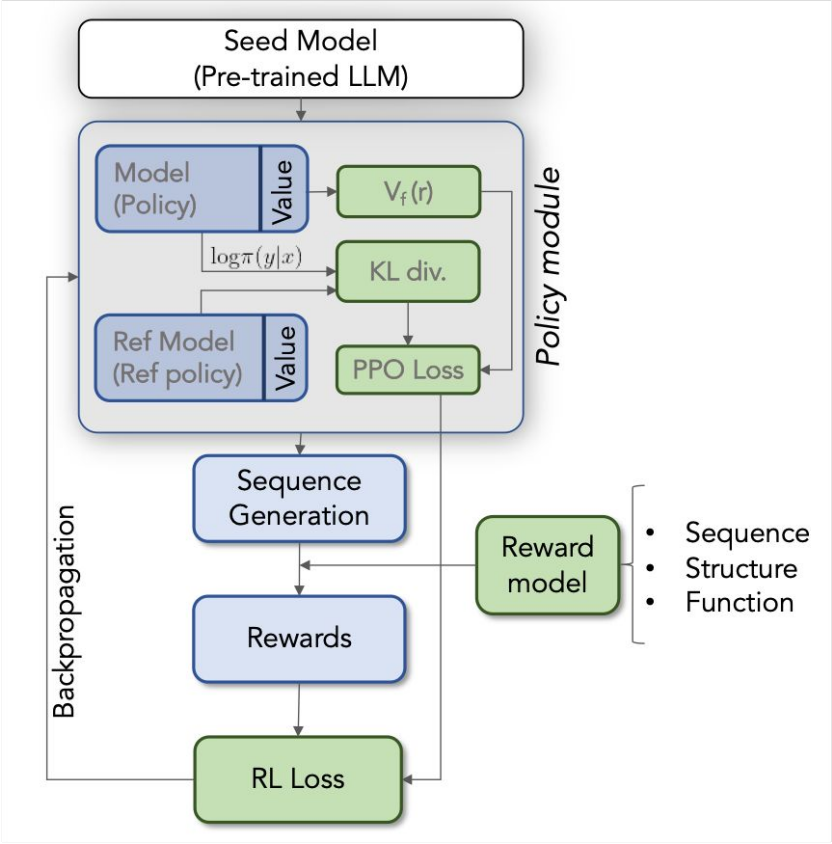
Zvyagin, M. et al., **GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics**., https://www.biorxiv.org/content/10.1101/2022.10.10.511571v2

Detection and prediction of variants of concern



Beguir, K. et al, **Early Computational Detection of Potential High Risk SARS-CoV-2 Variants**, https://doi.org/10.1101/2021.12.24.474095
Hie, B. et al, **Learning the language of viral evolution and escape**, https://www.science.org/doi/10.1126/science.abd7331

Designing enzymes by incorporating experimental feedback (aka ChatGPT for protein design)



- Need general framework that enables generative design of proteins by incorporating experimental feedback
- Approach inspired by Reinforcement Learning through Human Feedback (RLHF)
- Rewards for the model:
  - intrinsic – sequence specific (e.g., GC content for environmental adaptation)
  - extrinsic – functional annotation/ enzyme activity measured via experimentation

Reinforcement Learning with Experimental Feedback for Protein Design, G. Dharuman, H. Ma, L. Ward, P. Setty, O. Gokdemir, K. Hippe, A. Brace, and A. Anandkumar (Patent pending)

LLMs for workflows

☆ **LLMs in workflows**

- Workflows with embedded LLM inference/fine-tuning
- LLM is used as a subroutine to analyze scientific data
- LLM is fine-tuned based on outputs from AI-based and traditional HPC workloads (e.g., simulation)

☆ **LLM-driven workflows**

- LLM chooses the next experiment inputs to advance a scientific campaign
- Workflow components are a mix of traditional and ML/AI based functions

? **LLM-generated workflows**

- Parsl-enabled LangChain for functions with HPC requirements
- Chooses the types of tasks to run (i.e., workflow components) to advance a scientific campaign
- Generates functions from literature and matches compute requirements to existing (or generated) Parsl configs

? **Composable workflows**

- Modular workflow design to enable automatic function composition
- Enables automated LLM-proposed scientific campaigns
- Scientific discovery as an automated search problem over the space of workflows

Example prompt: "You are a workflow developer. Build a workflow which trains an ML-based molecular docking surrogate using high-throughput screening and then runs DeepDriveMD on the top candidates to physically verify the protein-ligand binding site. Please provide example configurations compatible with the Aurora supercomputer to investigate drug candidates from the Enamine database for all known SARS-CoV-2 targets."

vLLM – Serving LLMs on HPC

vLLM is a fast and easy-to-use library for LLM inference and serving.

vLLM is fast with:

- State-of-the-art serving throughput
- Efficient management of attention key and value memory with **PagedAttention**
- Continuous batching of incoming requests
- Optimized CUDA kernels

vLLM is flexible and easy to use with:

- Seamless integration with popular Hugging Face models
- High-throughput serving with various decoding algorithms, including *parallel sampling*, *beam search*, and more
- Tensor parallelism support for distributed inference
- Streaming outputs
- OpenAI-compatible API server

More details: https://github.com/vllm-project/vllm
Globus-Compute / Parsl vLLM tutorial: https://github.com/braceal/vllm-globus-compute/tree/main

Kwon, W. et al., **Efficient Memory Management for Large Language Model Serving with PagedAttention**., https://arxiv.org/abs/2309.06180

Acknowledgements

Rick Stevens (Argonne/ University of Chicago)

Rommie Amaro (University of California San Diego)

John Stone (University of Illinois Urbana-Champaign)

Lillian Chong (University of Pittsburgh)

Tom Gibbs (NVIDIA)

~300 collaborators on COVID-19 research

Funding:

-- DOE ASCR Co-design of AI Approaches for Multi-modal datasets

-- Exascale Computing Project (ECP): Cancer Deep Learning Environment (CANDLE), ECP Co-design for Online Data Analysis and Reduction CODAR, ExaWorks

-- Department of Energy's Advanced Scientific Computing Research Program

-- DOE National Virtual Biotechnology Laboratory

Computing:

-- Argonne/Oak Ridge Leadership Computing Facilities

-- Livermore Computing at the Lawrence Livermore National Laboratory

-- HPC Consortium for COVID-19 research

THANK YOU!!

QUESTIONS/ COMMENTS:

RAMANATHANA@ANL.GOV