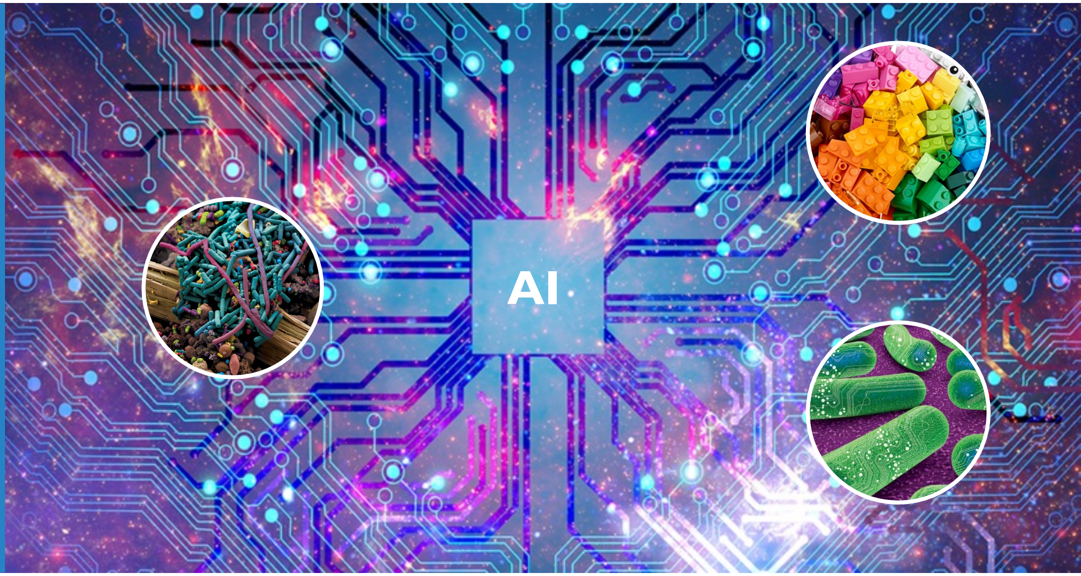
# Foundation models for complex biological systems + design

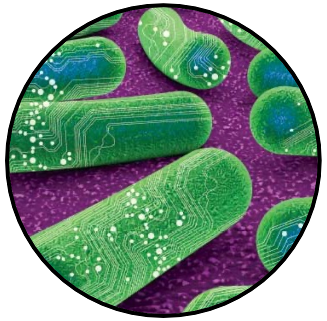


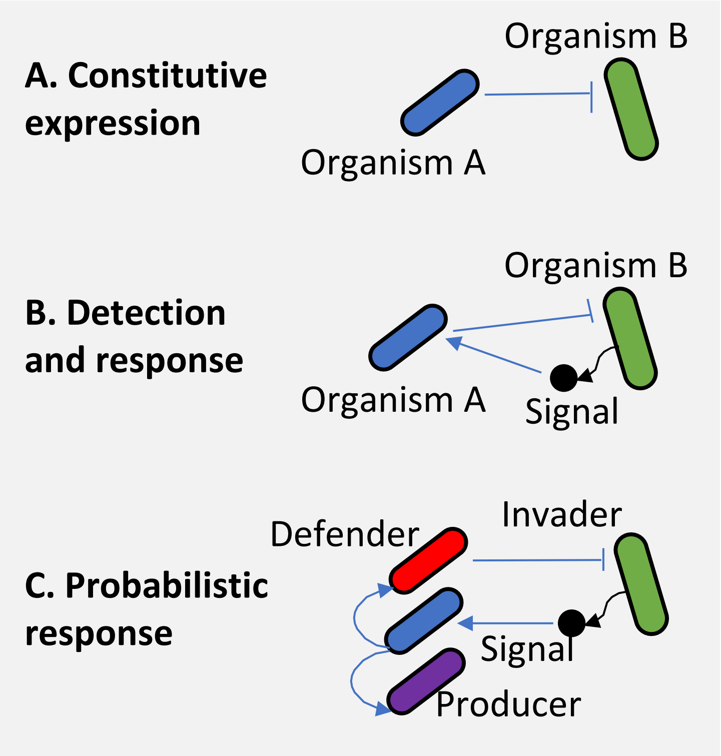
**Arvind Ramanathan/ ramanathana@anl.gov**

Argonne National Laboratory/ University of Chicago Consortium for Advanced Science and Engineering (CASE)

Northwestern-Argonne Institute for Science and Engineering (NAISE)

# Project overview





A. Constitutive expression

Organism B

Organism A

B. Detection and response

Organism B

Organism A

Signal

C. Probabilistic response

Defender

Invader

Signal

Producer

- Building microorganisms
  - System for generating **biological modules** (*in vitro* closed loop system)
  - **Biocontainment** development and implementation (CRISPR)
  - System for **coupling modules with biocontainment** (CRAGE)
  - System for **evaluating integrated systems** (experimental; modeling; ML- and NLP-based modeling)
- Building communities of microorganisms
- Concept of biocontainment

Argonne NATIONAL LABORATORY

BERKELEY LAB

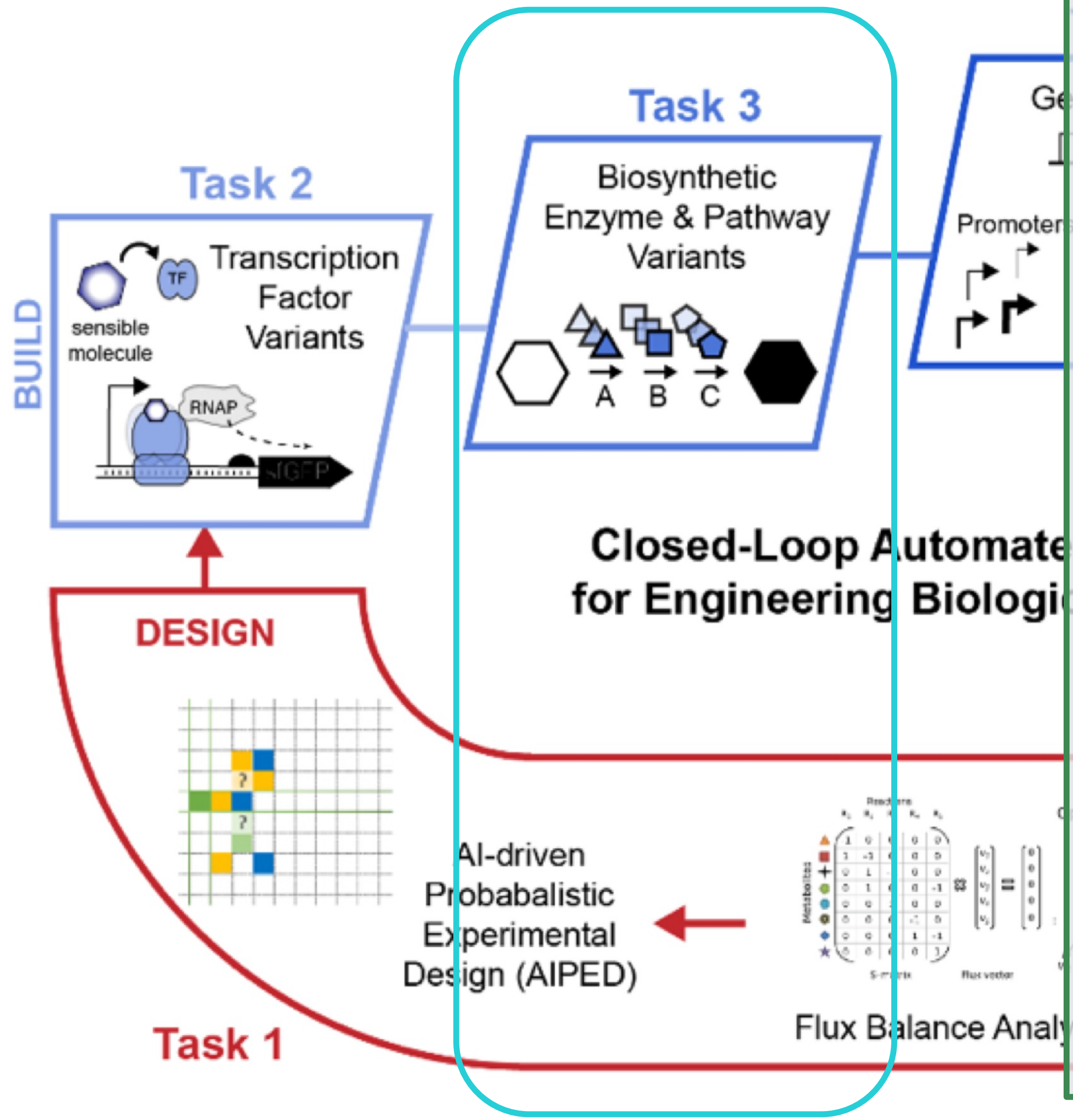
JGI JOINT GENOME INSTITUTE

Northwestern University

THE UNIVERSITY OF CHICAGO

# An integrative platform for rapid engineering of biological parts

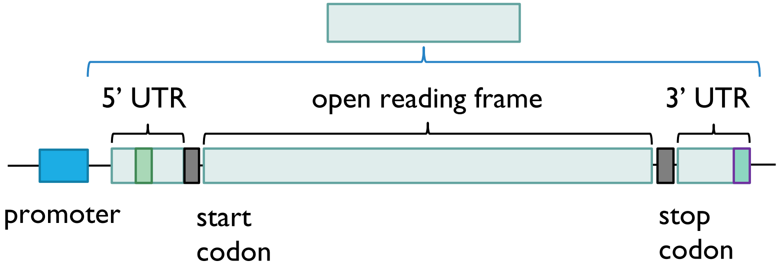
**Poster 2: Priyanka Setty**


Task 3
Biosynthetic Enzyme & Pathway Variants
A
B
C
Task 2
Transcription Factor Variants
TF
sensible molecule
RNAP
sfGFP
BUILD
Genetic Parts
Promoters
Closed-Loop Automated Systems for Engineering Biological Modules
DESIGN
AI-driven Probabalistic Experimental Design (AIPED)
Flux Balance Analysis (FBA)
Task 1

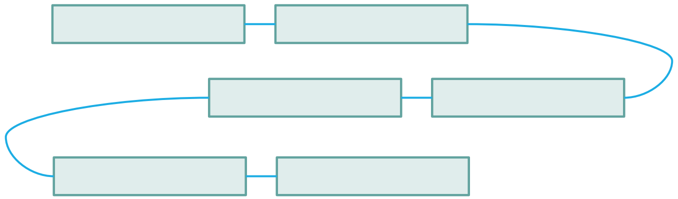
- How to think about biological hierarchy and information?
  - implicit and explicit representation learning
  - genome-scale language models
- How to understand hierarchy of biological information?
  - individual gene/protein
  - pathways
  - genome-scale
- How do we generate new examples?
  - designing new genes (MDH as an example)
  - understanding how SARS-CoV-2 is evolving
  - integrating with biophysics (experiments + simulations)
- Future work/perspectives

# Biological information and hierarchy

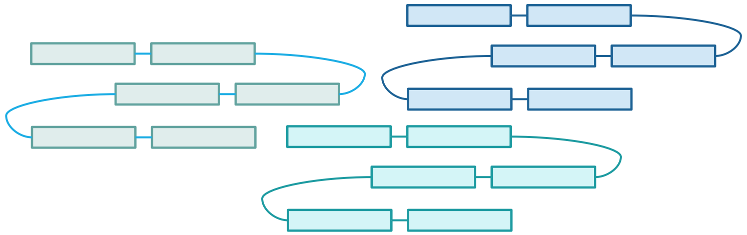
# Hierarchical information within ‘-omics’ data



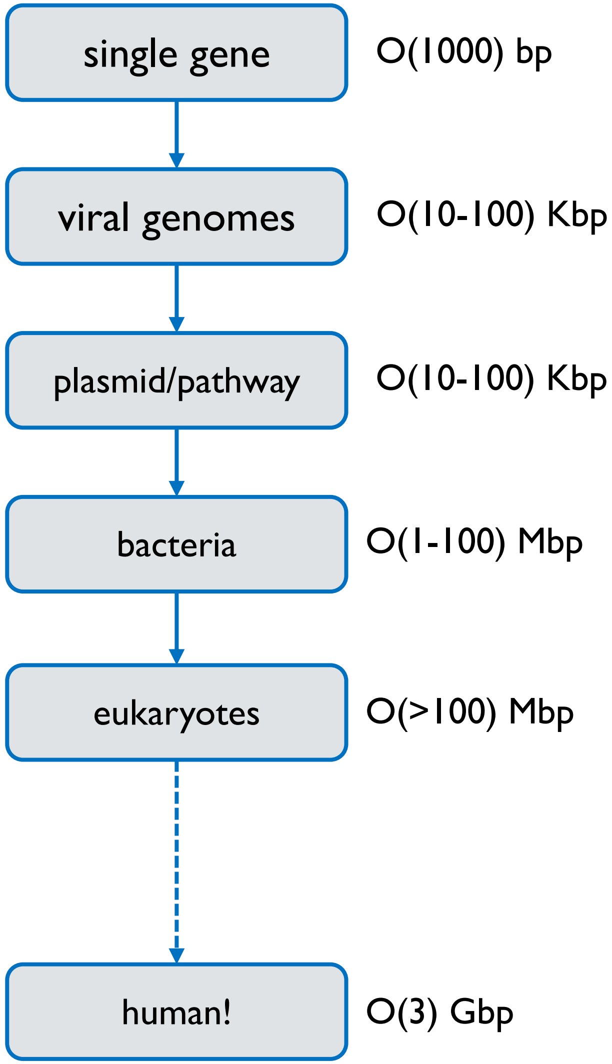
*gene/ gene product (aka protein)*



*collection of genes (either as ”contigs” or ORFs)*



*entire genomes*



single gene O(1000) bp

viral genomes O(10-100) Kbp

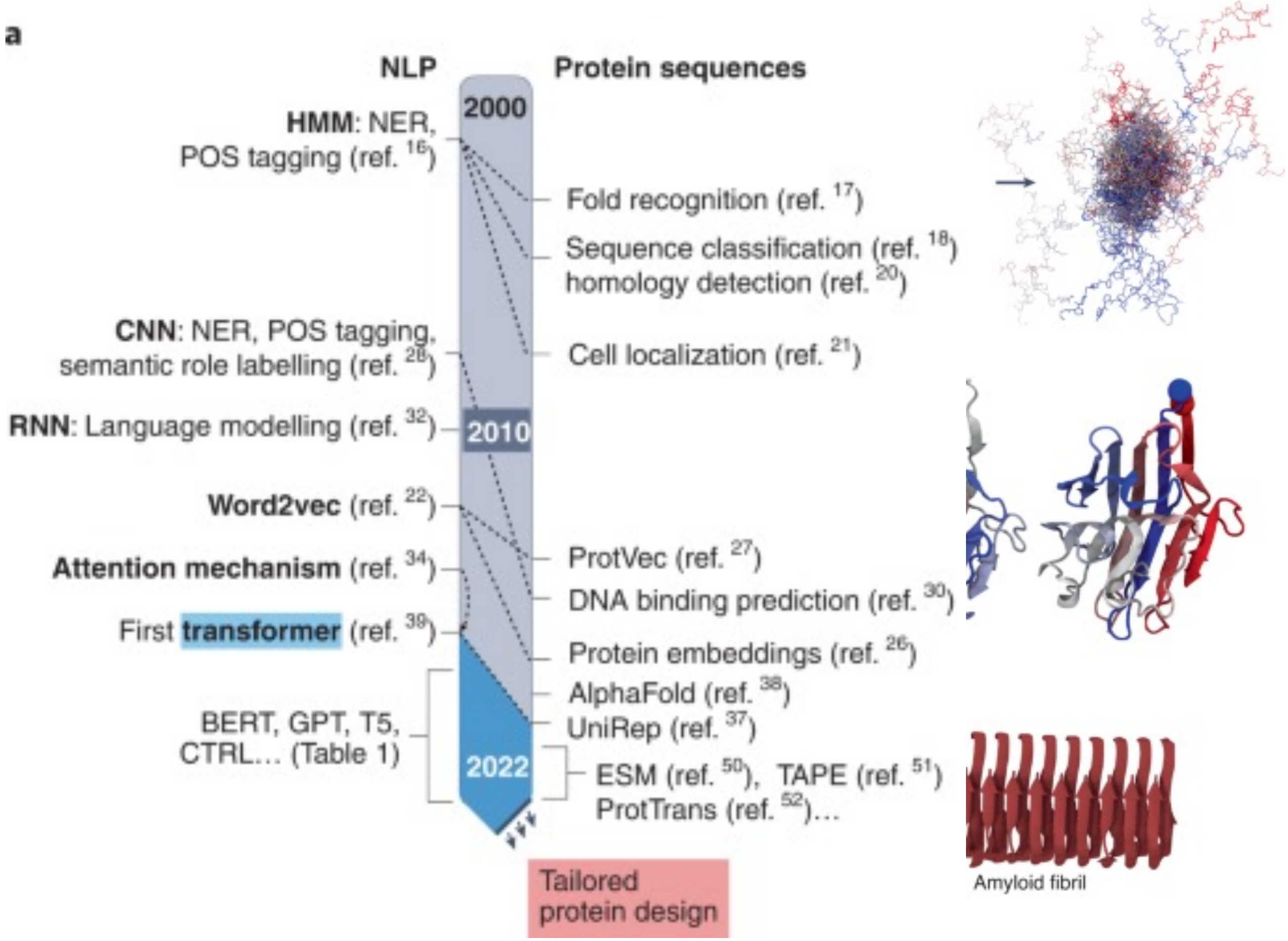
plasmid/pathway O(10-100) Kbp

bacteria O(1-100) Mbp

eukaryotes O(>100) Mbp

human! O(3) Gbp

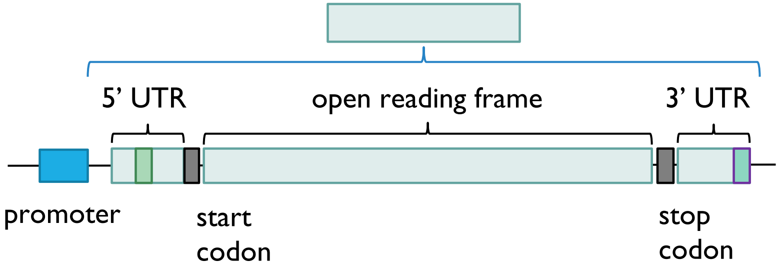
# Traditional approach: use protein language models…


a
NLP
Protein sequences
2000
HMM: NER, POS tagging (ref. 16)
Fold recognition (ref. 17)
Sequence classification (ref. 18) homology detection (ref. 20)
CNN: NER, POS tagging, semantic role labelling (ref. 28)
Cell localization (ref. 21)
RNN: Language modelling (ref. 32)
2010
Word2vec (ref. 22)
ProtVec (ref. 27)
Attention mechanism (ref. 34)
DNA binding prediction (ref. 30)
First transformer (ref. 39)
Protein embeddings (ref. 26)
AlphaFold (ref. 38)
UniRep (ref. 37)
BERT, GPT, T5, CTRL… (Table 1)
2022
ESM (ref. 50), TAPE (ref. 51) ProtTrans (ref. 52)…
Tailored protein design
Amyloid fibril

Controllable protein design with language models, N. Ferruz, B. Hocker, Nature Machine Intelligence (https://www.nature.com/articles/s42256-022-00499-z)

- Transformers have evolved to be good constructs for capturing protein "language"
  - consists of 20 natural amino acids chained together
  - attention mechanism capture interactions across amino-acids
- Considerable challenges in translating NLP based approaches to protein language
  - sequence length
  - attention mechanism

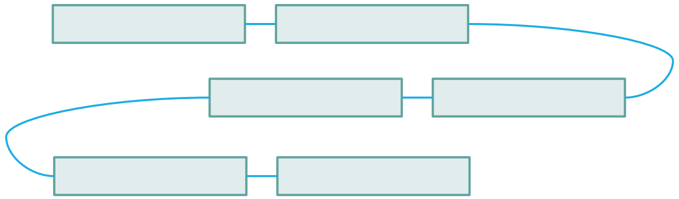
# Genome-scale language models (GenSLM)



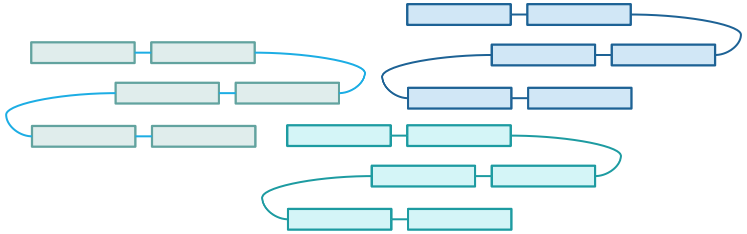
*gene/ gene product (aka protein)*

Enzyme/transcription factor sequences - codon level tokenization, GPT models

Open reading frames – codon level tokenization, Reformer model



*collection of genes (either as ”contigs” or ORFs)*

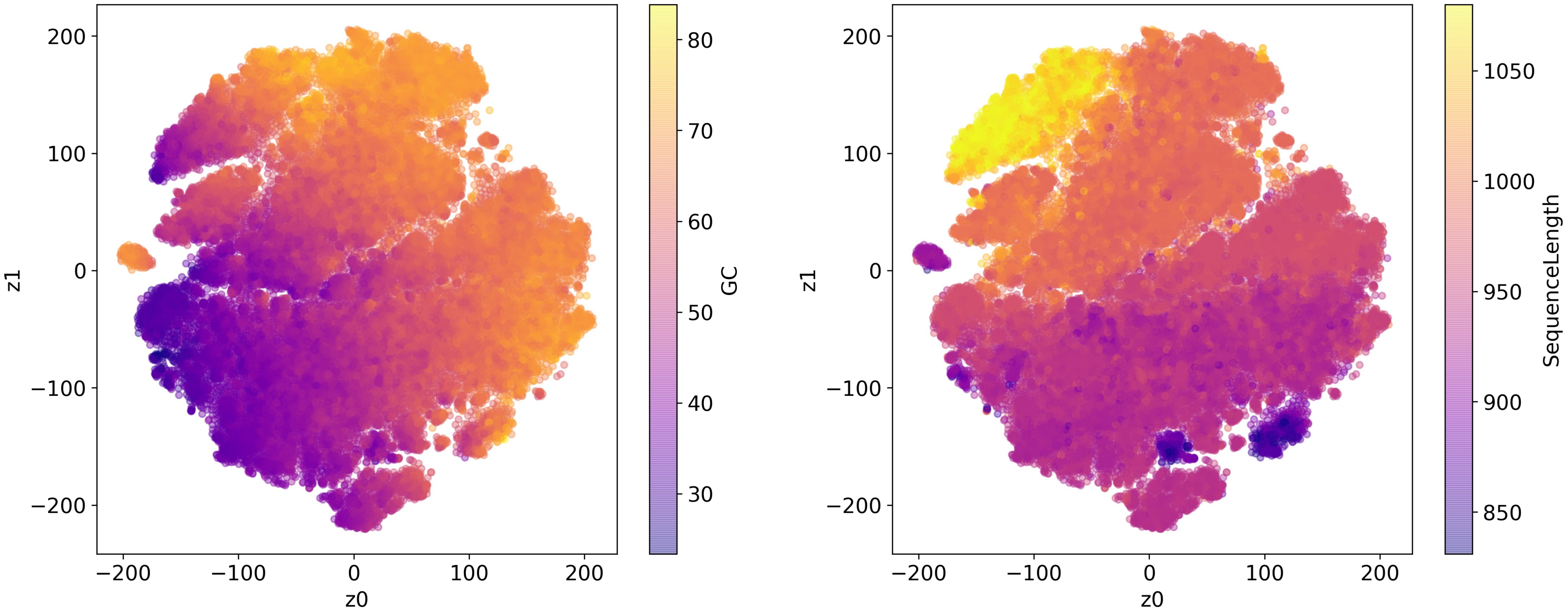


*entire genomes*

Full genome sequences - BPE Encoding, cannot currently generate full genomes

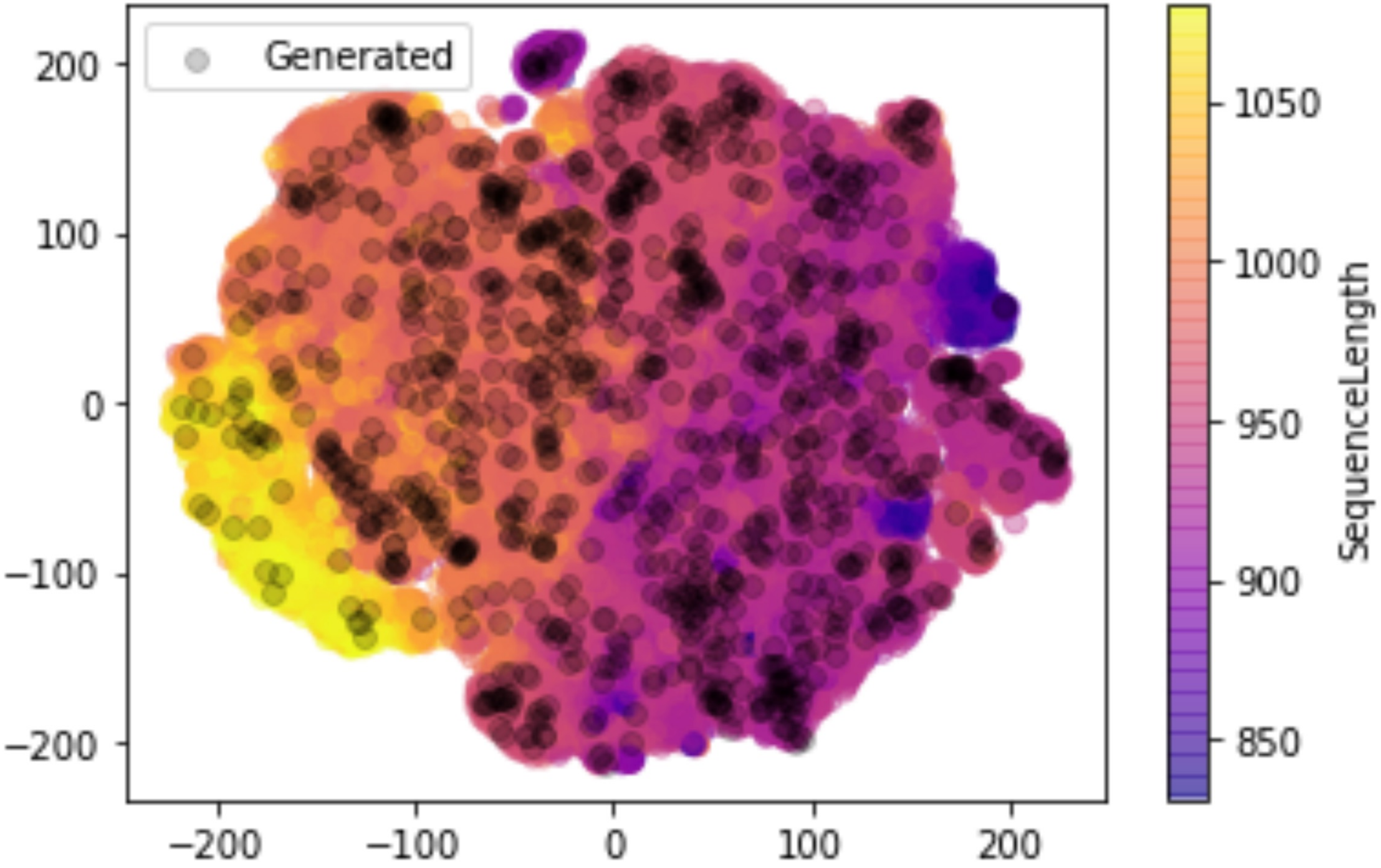
- Go beyond traditional k-mer models:
  - variable length issues
- At each level of hierarchy maintain information learned at the lower levels (gene → collection/cluster → full genomes)
- Scale at each level but “tie” it together with stable diffusion models

# GenSLMs capture individual gene-level information

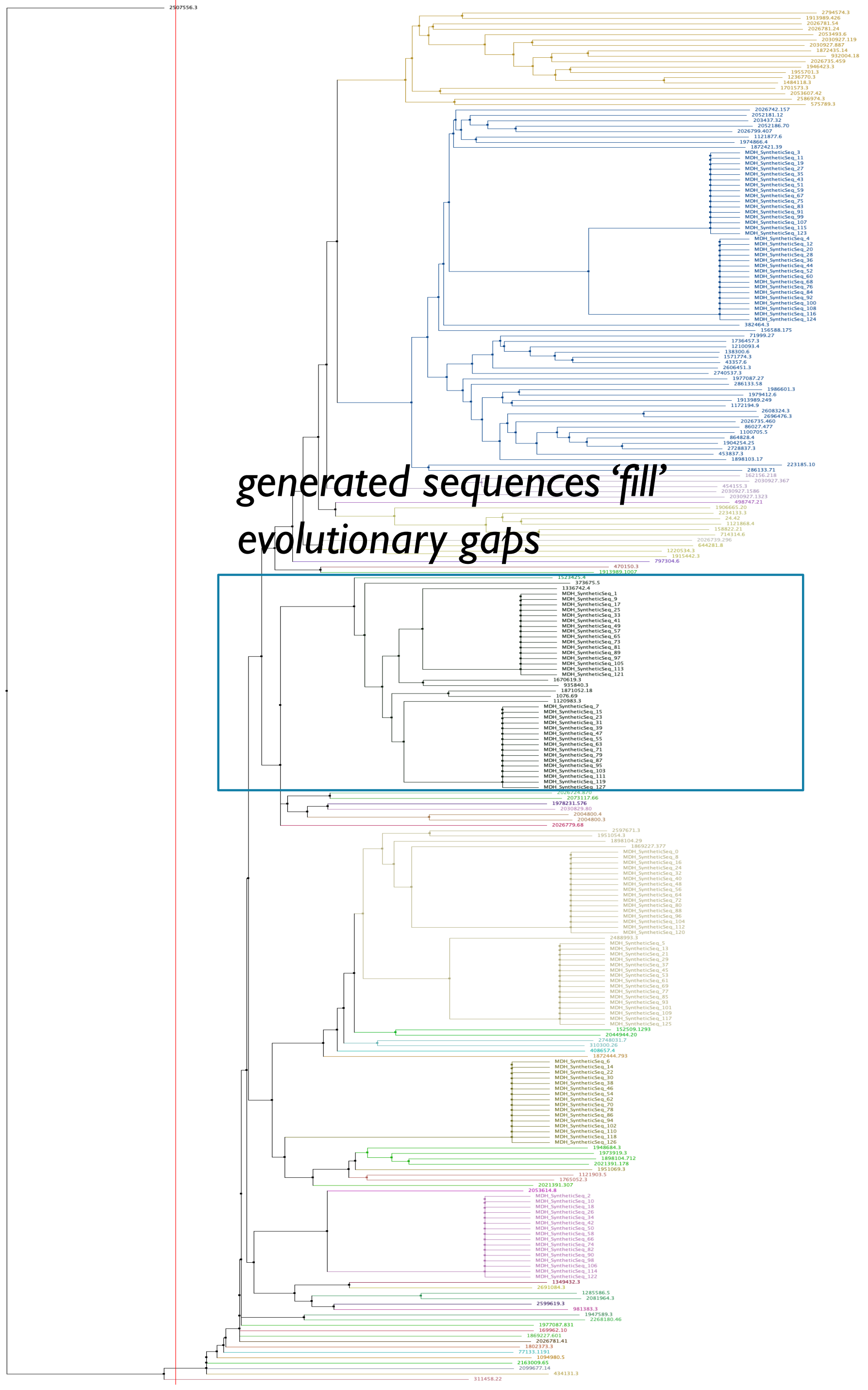


- Unsupervised learning on 36K malate dehydrogenase (MDH) sequences
- Embedding of the latent space using t-SNE (or even UMAP) reveals characteristic features:
  - GC content, sequence length variations, molecular weight…
- These individual gene-level models can be used to interpolate and sample

# GenSLMs are accurate enough to generate gene sequences…

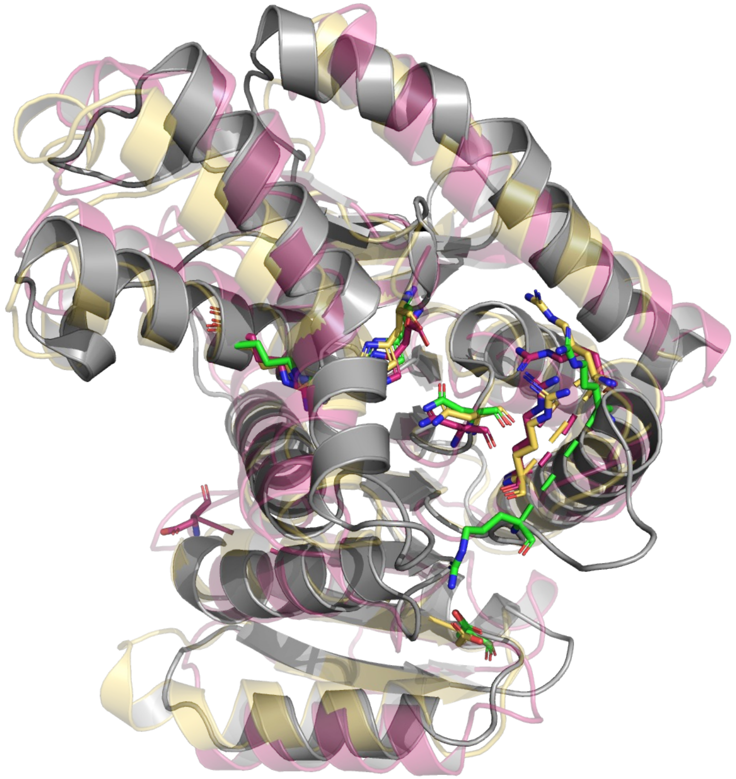


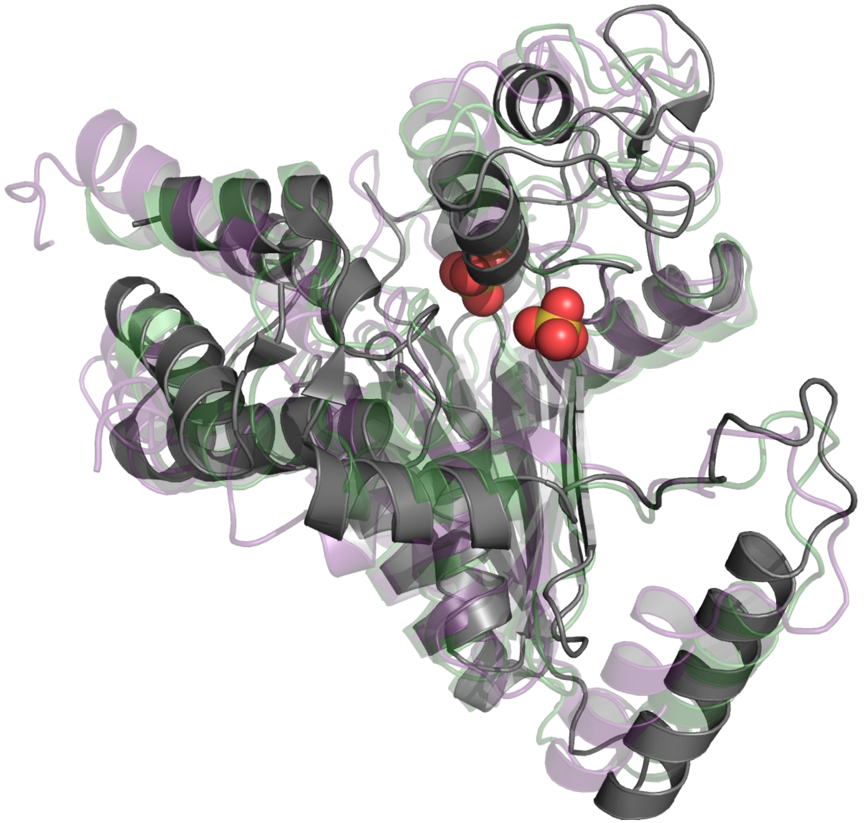
UMAP embeddings of generated sequences agree with learned embeddings using GPT-2



*generated sequences 'fill' evolutionary gaps*

# … genes → proteins share MDH similarity at key sites as predicted via OpenFold

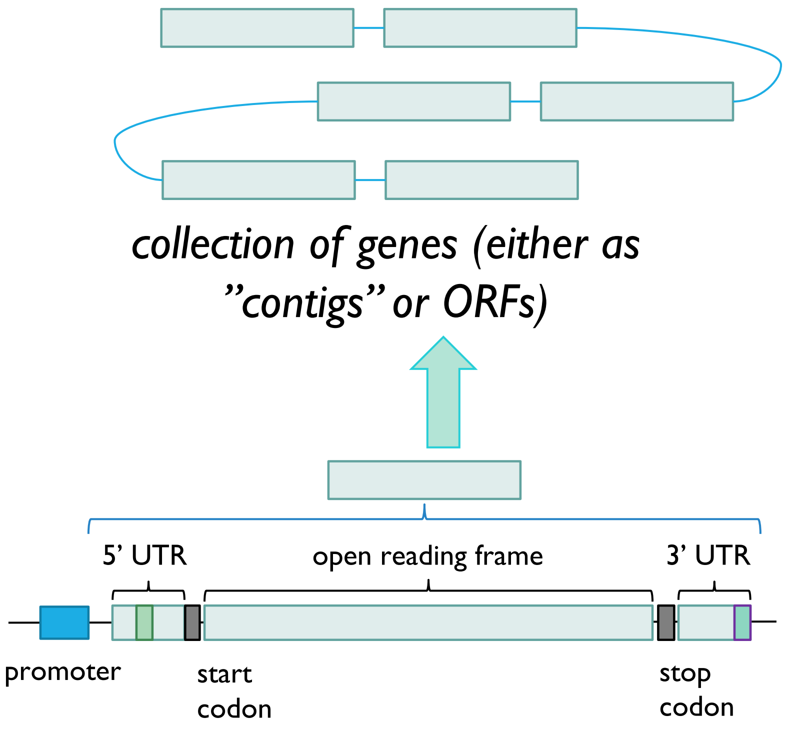


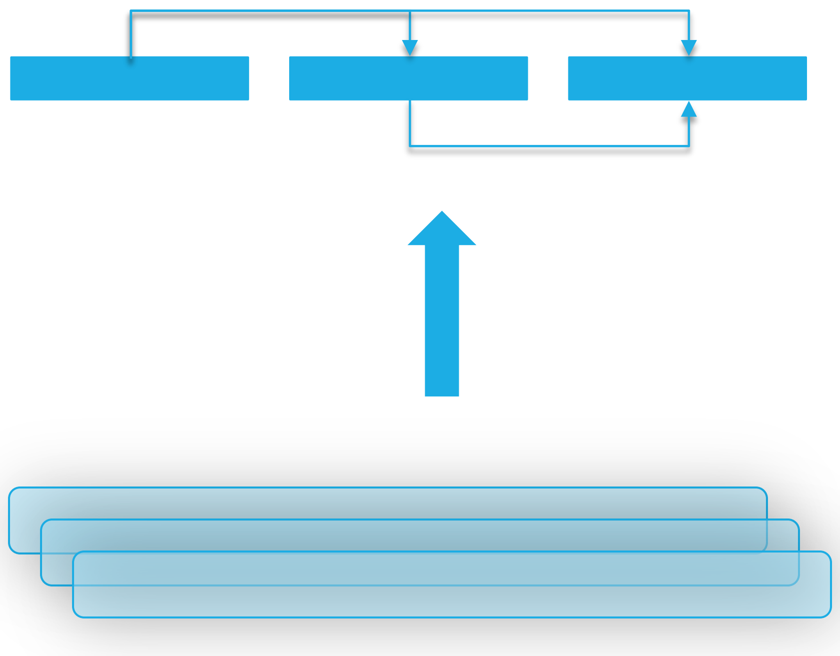


GenSLMs learn the two distinct isoforms for MDH and within each isoform we find conservation of key residues and placement of binding sites

# Extending GenSLMs to model SARS-CoV-2 evolutionary dynamics

# Overcoming length limitations of GenSLMs





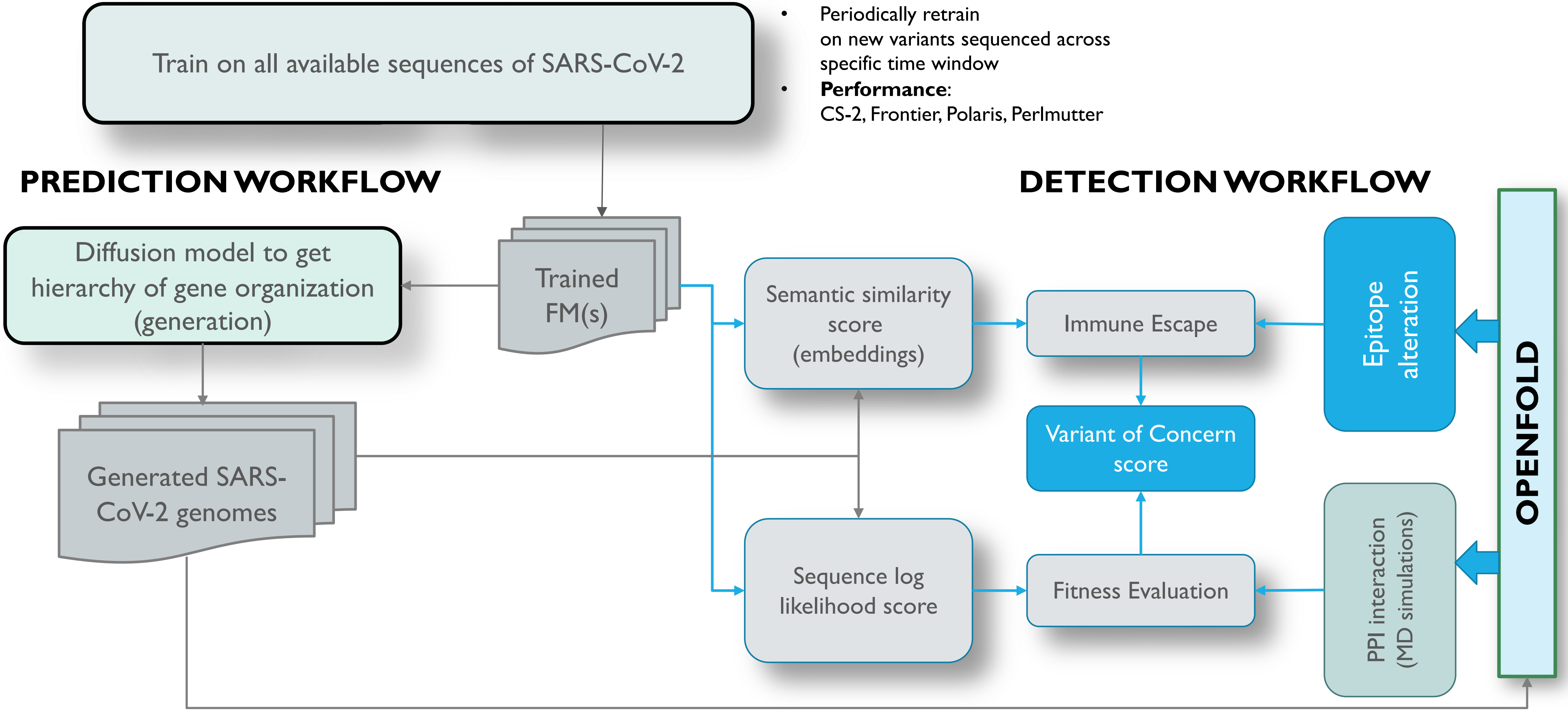
Stable diffusion models (read and learn context amongst ORFs)
GPT-2 like (8 layers + 8 attention heads + 10240 x 10240 positional encoding (no ordering in ORFs)

- Implicit representation of hierarchy by integrating LLMs with stable diffusion models

# A Foundation Model approach for SARS-CoV-2 genomes …

## TRAINING

- Periodically retrain on new variants sequenced across specific time window
- **Performance**: CS-2, Frontier, Polaris, Perlmutter



Train on all available sequences of SARS-CoV-2

## PREDICTION WORKFLOW

Diffusion model to get hierarchy of gene organization (generation)

Trained FM(s)

Generated SARS-CoV-2 genomes

## DETECTION WORKFLOW

Semantic similarity score (embeddings)

Sequence log likelihood score

Immune Escape

Variant of Concern score

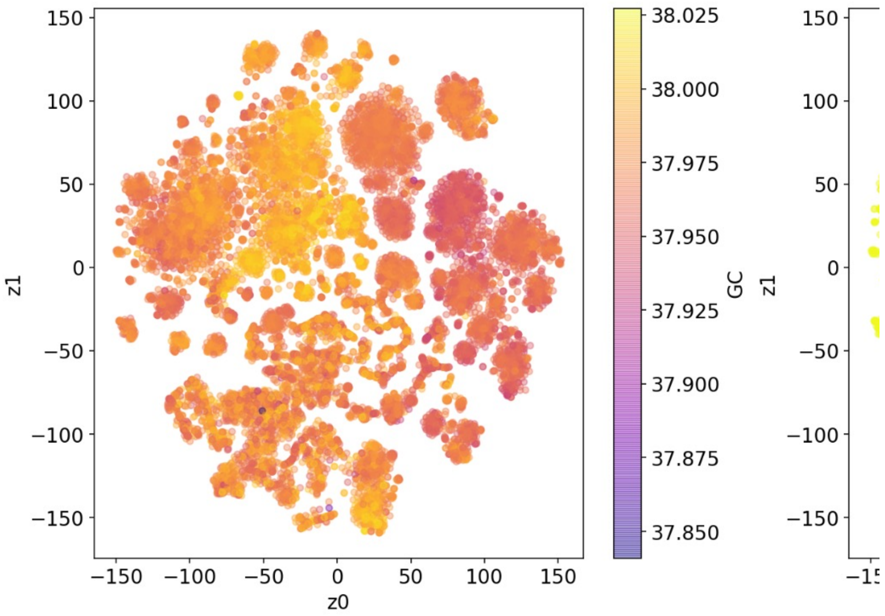
Fitness Evaluation

Epitope alteration

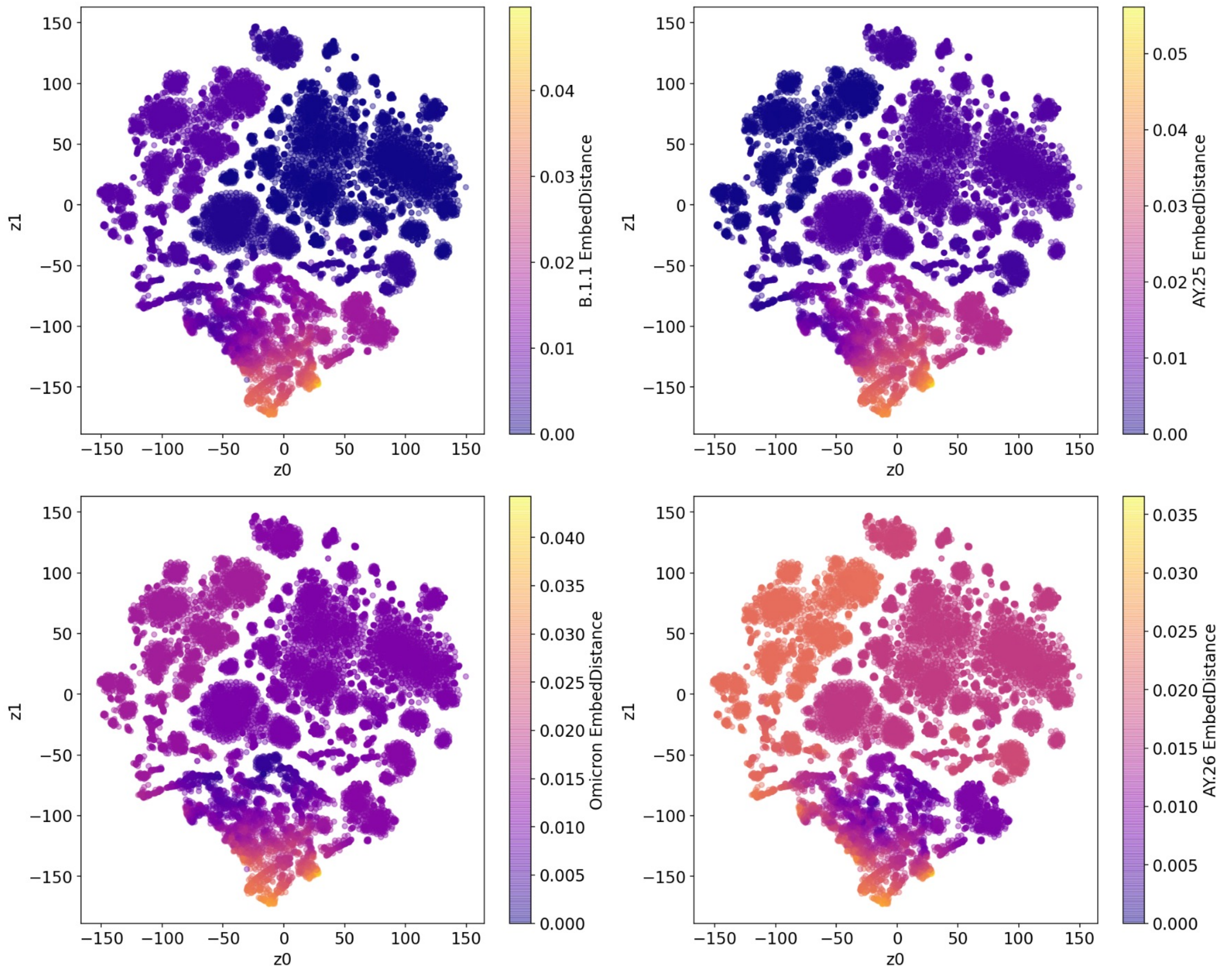
PPI interaction (MD simulations)

OPENFOLD

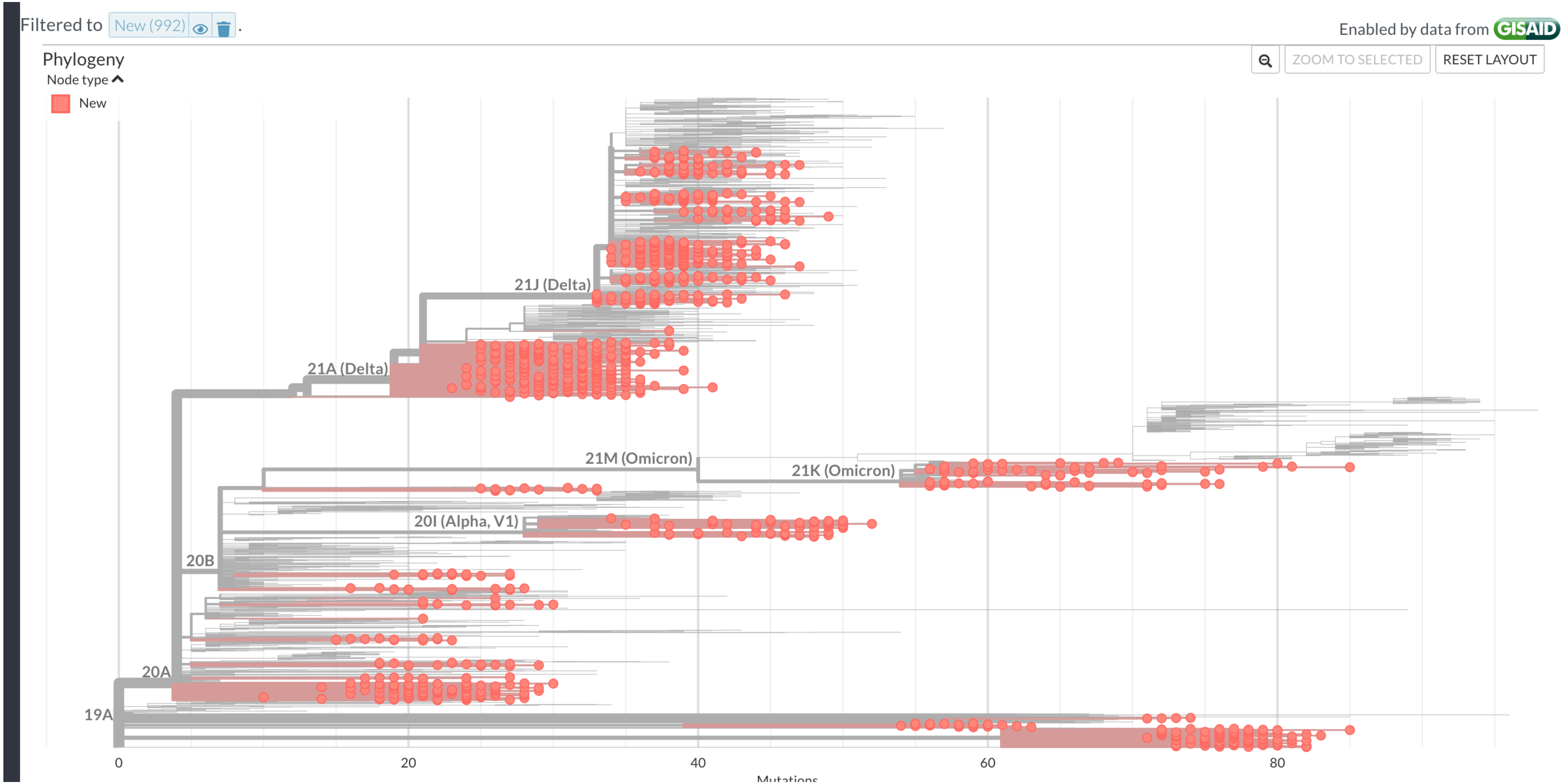
# reveals intrinsic evolutionary patterns of SARS-CoV-2



- Variant clustering
- Semantic distance in the latent space can classify variants of interst

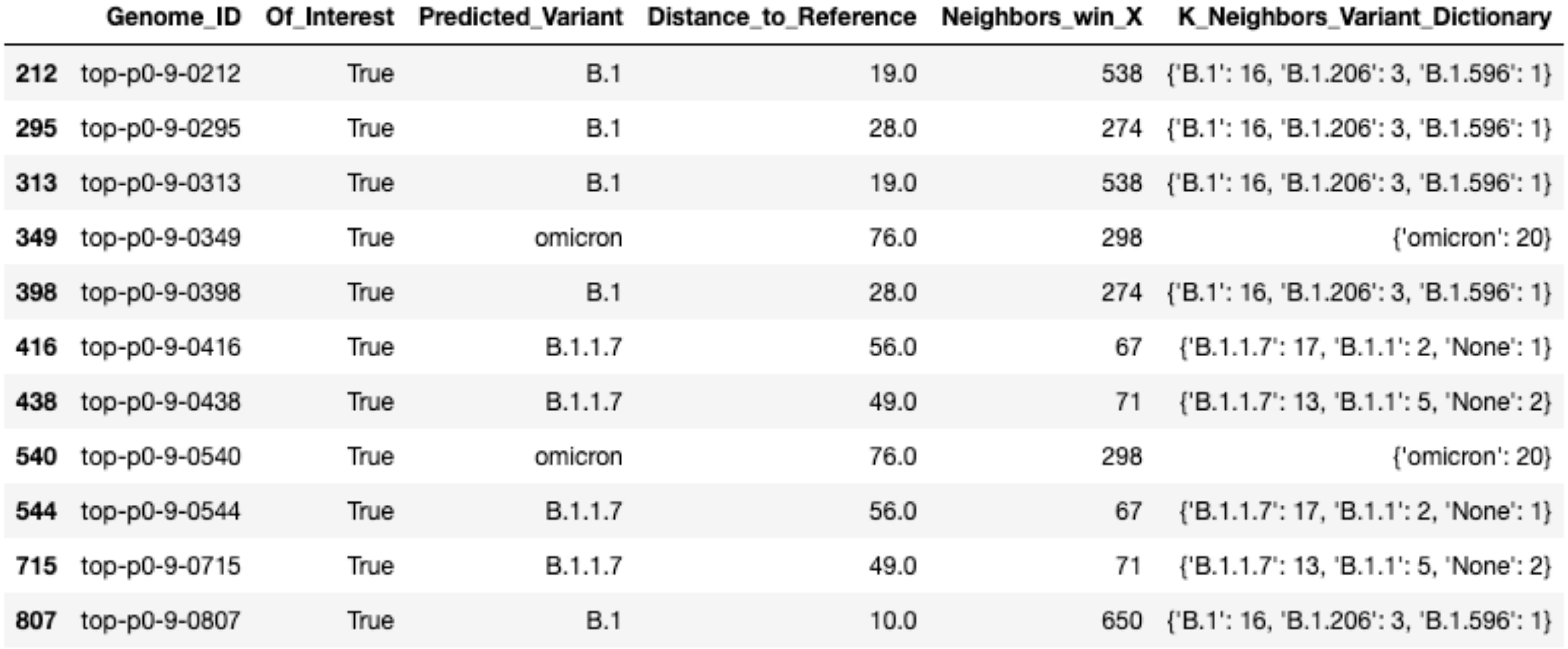


# We can generate new sequences that look like SARS-CoV-2



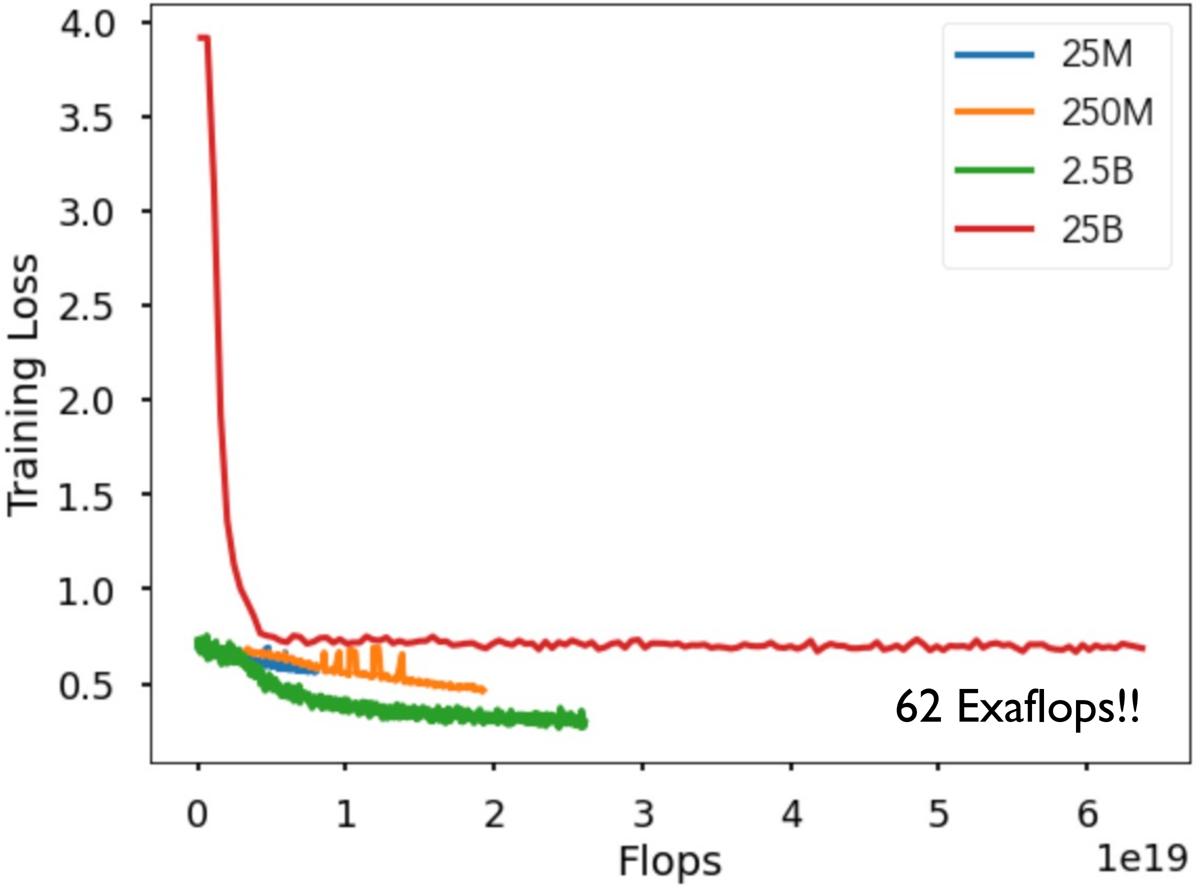
Filtered to New (992).

# We can prioritize sequences that are novel and perceived VOCs



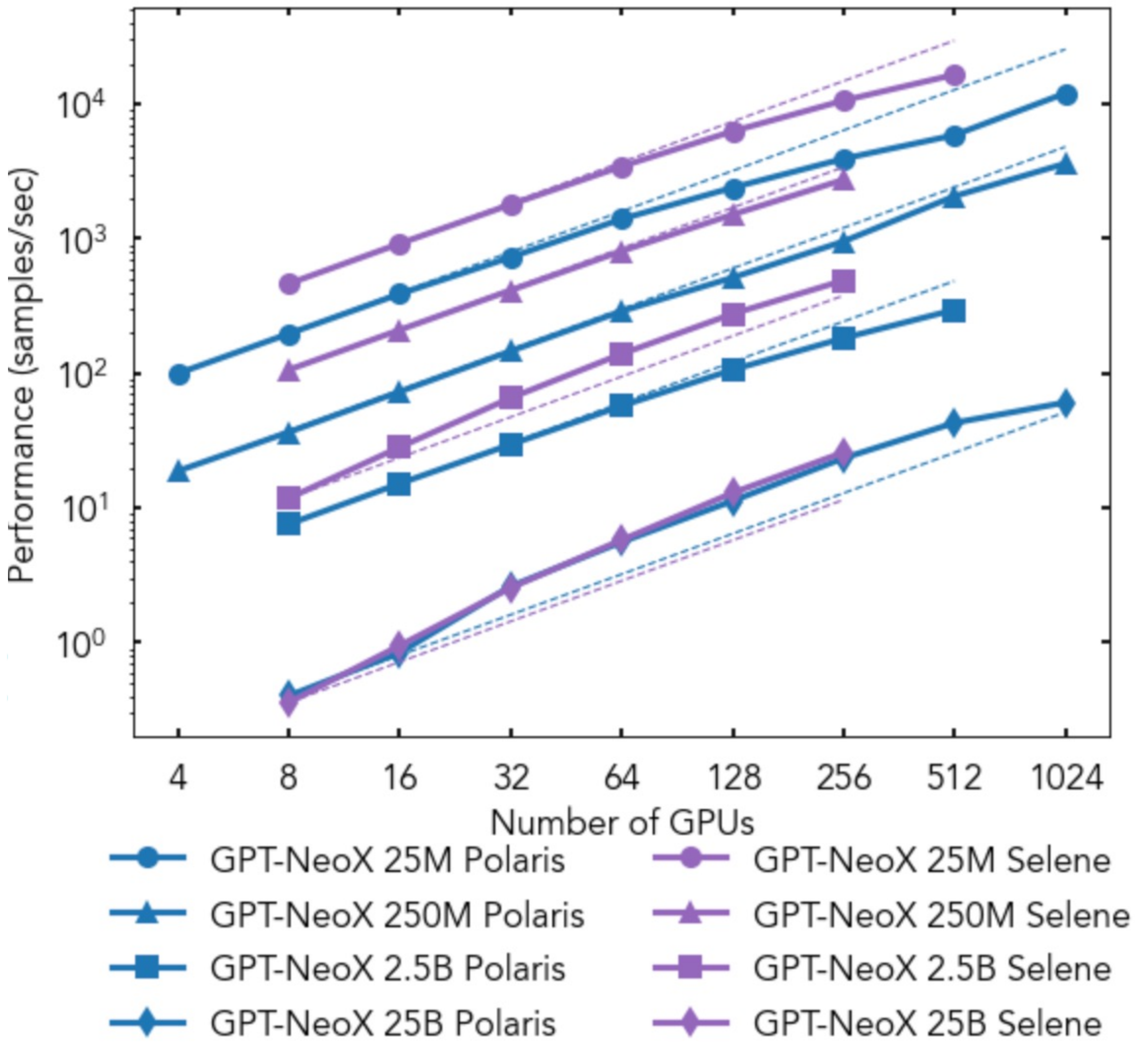
| | Genome_ID | Of_Interest | Predicted_Variant | Distance_to_Reference | Neighbors_win_X | K_Neighbors_Variant_Dictionary |
|---|---|---|---|---|---|---|
| **212** | top-p0-9-0212 | True | B.1 | 19.0 | 538 | {'B.1': 16, 'B.1.206': 3, 'B.1.596': 1} |
| **295** | top-p0-9-0295 | True | B.1 | 28.0 | 274 | {'B.1': 16, 'B.1.206': 3, 'B.1.596': 1} |
| **313** | top-p0-9-0313 | True | B.1 | 19.0 | 538 | {'B.1': 16, 'B.1.206': 3, 'B.1.596': 1} |
| **349** | top-p0-9-0349 | True | omicron | 76.0 | 298 | {'omicron': 20} |
| **398** | top-p0-9-0398 | True | B.1 | 28.0 | 274 | {'B.1': 16, 'B.1.206': 3, 'B.1.596': 1} |
| **416** | top-p0-9-0416 | True | B.1.1.7 | 56.0 | 67 | {'B.1.1.7': 17, 'B.1.1': 2, 'None': 1} |
| **438** | top-p0-9-0438 | True | B.1.1.7 | 49.0 | 71 | {'B.1.1.7': 13, 'B.1.1': 5, 'None': 2} |
| **540** | top-p0-9-0540 | True | omicron | 76.0 | 298 | {'omicron': 20} |
| **544** | top-p0-9-0544 | True | B.1.1.7 | 56.0 | 67 | {'B.1.1.7': 17, 'B.1.1': 2, 'None': 1} |
| **715** | top-p0-9-0715 | True | B.1.1.7 | 49.0 | 71 | {'B.1.1.7': 13, 'B.1.1': 5, 'None': 2} |
| **807** | top-p0-9-0807 | True | B.1 | 10.0 | 650 | {'B.1': 16, 'B.1.206': 3, 'B.1.596': 1} |

# Scaling laws for GenSLMs…



62 Exaflops!!

Throughput



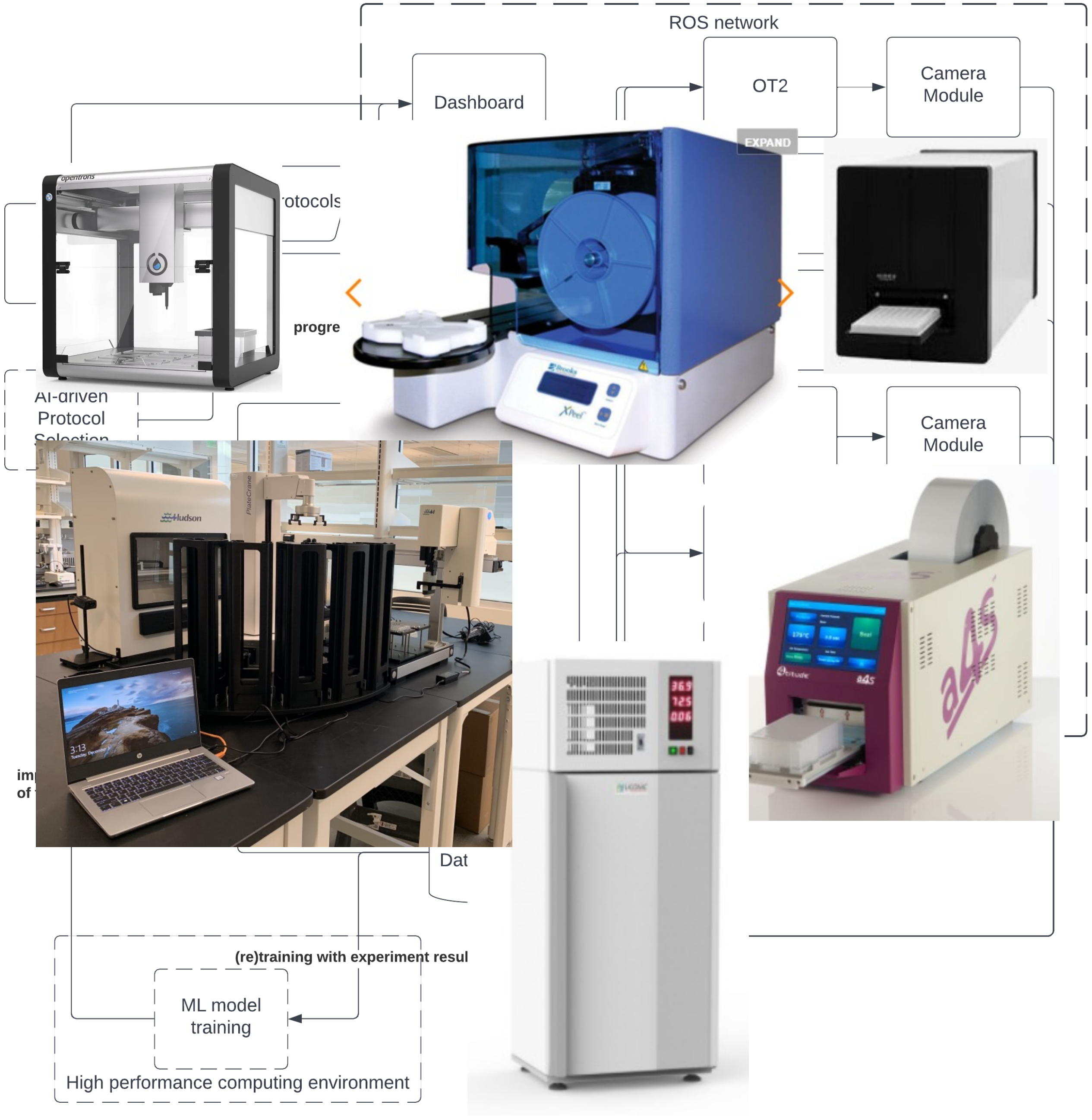
gth

ngth

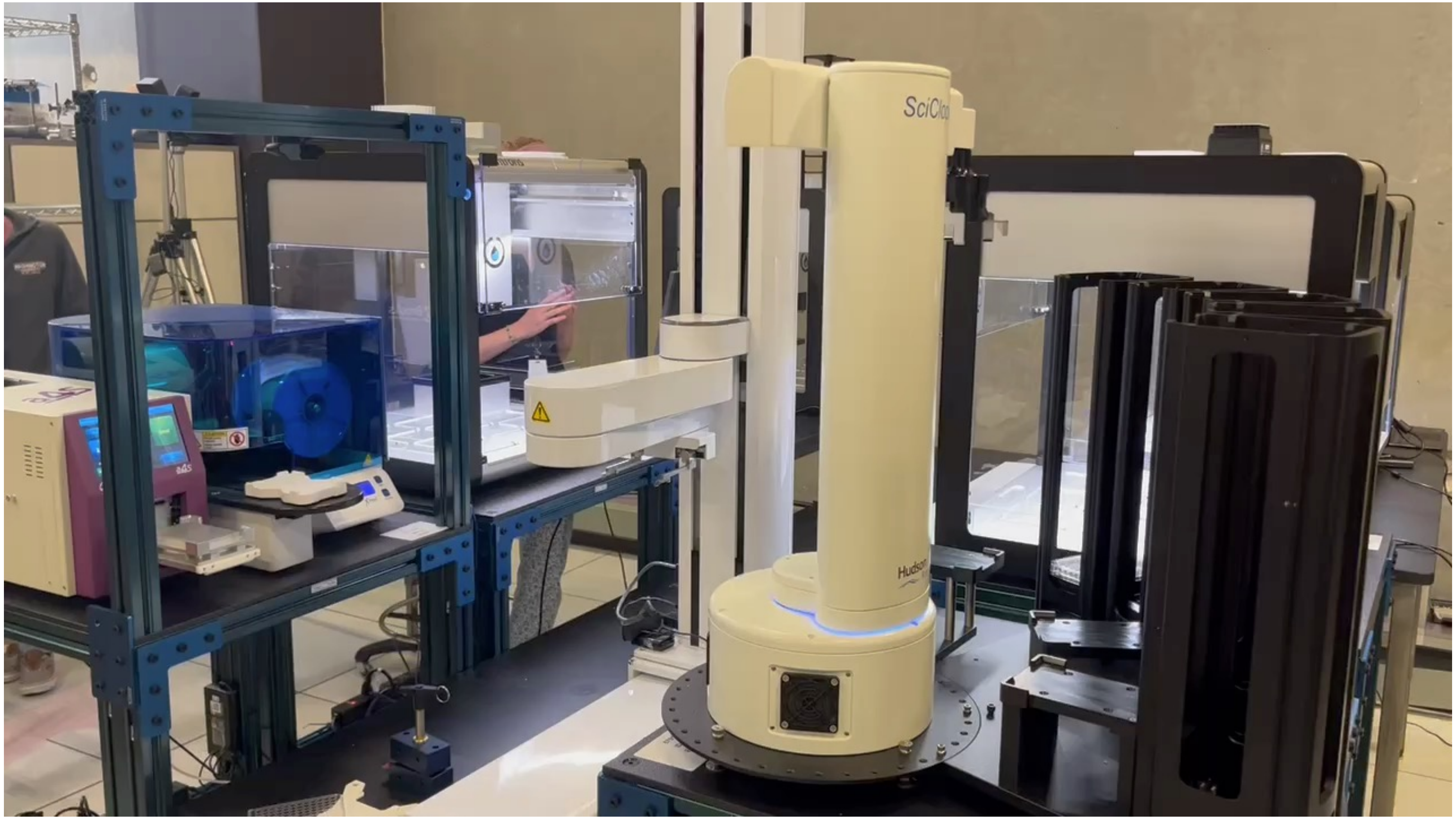
# Designing a robust automated protein engineering platform…

# System architecture of a protein engineering platform

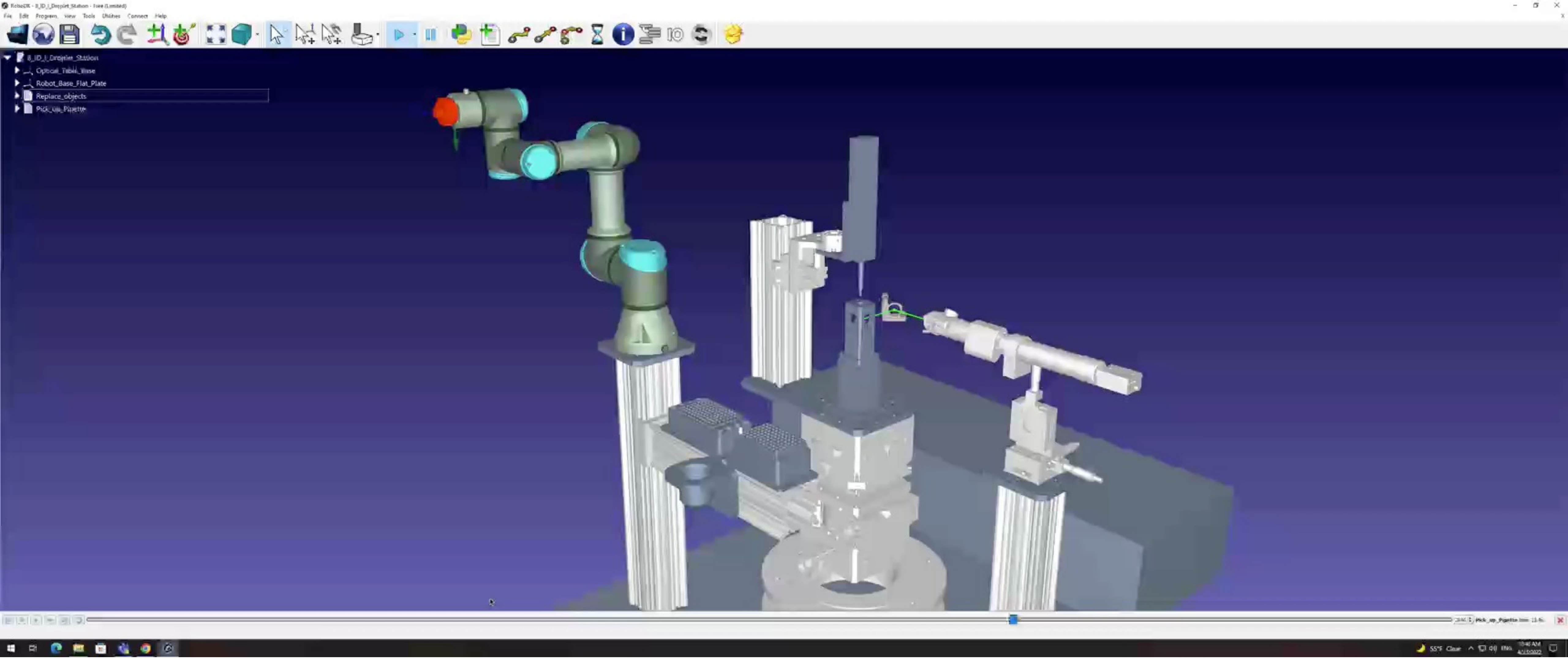
- Diverse robotic platforms that don’t necessarily talk to each other:
  - Liquid handling, sealer, peeler
  - no standardized interface
  - sample movement
- Lack of a programming environment for even relatively simple protocols
- An engineering platform that can inter-operate across diverse robots
- Scalable, open Python API: https://github.com/AD-SDL
- Coordinates many different systems (UR-3 arms, Hudson platform, etc.)


ROS network
Dashboard
OT2
Camera Module
otocols
EXPAND
progre
AI-driven Protocol Selection
Camera Module
im of
Dat
(re)training with experiment resul
ML model training
High performance computing environment

# Making it all work together



# Digital twins enabled beamline experiments can enable reproducibility



# Towards the automation of scientific experiments…

HPC application

Methods

Extract Protocol

LLM

Step 1:

Step 2:

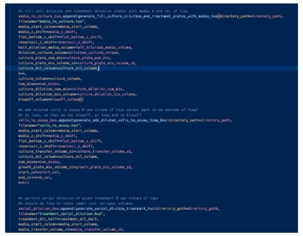
Step 3:

Step 4:

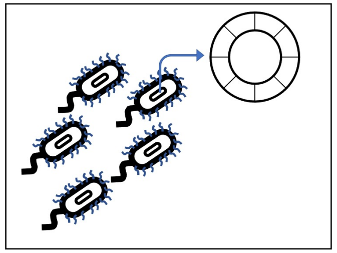
Robot Compatible

LLM

Python Script

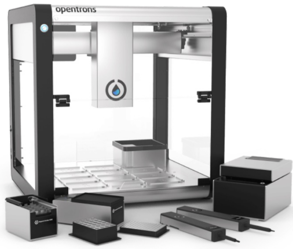


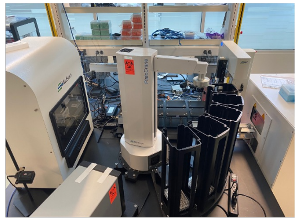
```
#+ fill cell dilution and treatment dilution plates with media # one col of tips
media_to_culture_hso.append(generate_fill_culture_dilution_and_treatment_plates_with_media_hso(directory_path=directory_path,
filename="media_to_culture.hso",
media_start_column=media_start_column,
media_z_shift=media_z_shift,
flat_bottom_z_shift=flat_bottom_z_shift,
reservoir_z_shift=reservoir_z_shift,
half_dilution_media_volume=half_dilution_media_volume,
dilution_culture_volume=dilution_culture_volume,
culture_plate_num_mix=culture_plate_num_mix,
culture_plate_mix_volume_s1=culture_plate_mix_volume_s1,
culture_dil_column=culture_dil_column,
k=k,
culture_column=culture_column,
num_mixes=num_mixes,
culture_dilution_num_mix=culture_dilution_num_mix,
culture_dilution_mix_volume=culture_dilution_mix_volume,
blowoff_volume=blowoff_volume))

#+ add diluted cells to assay # one column of tips unless want to do outside of loop?
#+ no loop, so that we can blowoff, or loop and no blowoff
cells_to_assay_hso.append(generate_add_diluted_cells_to_assay_loop_hso(directory_path=directory_path,
filename="cells_to_assay.hso",
media_start_column=media_start_column,
media_z_shift=media_z_shift,
flat_bottom_z_shift=flat_bottom_z_shift,
reservoir_z_shift=reservoir_z_shift,
culture_transfer_volume_s1=culture_transfer_volume_s1,
culture_dil_column=culture_dil_column,
num_mixes=num_mixes,
growth_plate_mix_volume_s1=growth_plate_mix_volume_s1,
start_col=start_col,
end_col=end_col,
k=k))


#+ perform serial dilution of given treatment # one column of tips
#+ should be fine to leave same? just increase volumes
serial_dilution_hso.append(generate_serial_dlution_treatment_hso(directory_path=directory_path,
filename="treatment_serial_dilution.hso",
treatment_dil_half=treatment_dil_half,
media_start_column=media_start_column,
media_transfer_volume_s2=media_transfer_volume_s2,
```



Identify Genes of Interest

GSLM



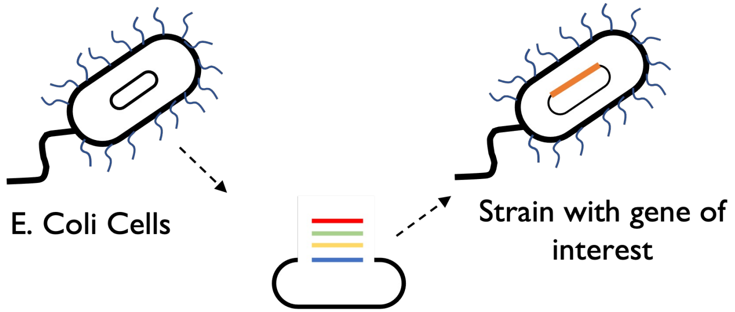


Autonomous Lab Set-Up

Factors for maximum Threonine Production

Final Strain

**ML guides next set of experiments to perform**



DNA Transformation

# Summary

- Foundation models (GenSLMs) can learn from genome-scale datasets:
  - applications for function annotation workflows, completing metagenome data, and many more
  - generative models provide suggestions for experiments → can be integrated with fitness optimization tools for design workflows
  - scaling on whole genomes faces many challenges
- AI-driven simulations will be important for advances in integrating experiments and theoretic understanding of complex bio-systems:
  - Hierarchical AI models enable multi-scale simulations
  - Acceleration with emerging AI hardware
- Automated platforms for experimental design and workflows

# Acknowledgements

Tom Gibbs
NVIDIA

Alexander Brace
Argonne/ University of Chicago

Venkatesh Mysore
NVIDIA

Maxim Zvyagin
Argonne

Anima Anandkumar *
Caltech/ NVIDIA

Hyunseung Yoo
Argonne

Zongyi Li
Caltech

i Yin
L

Michael Salim
Argonne

Aristedis Tsaris
NL

Austin Clyde
Argonne/ University of Chicago

Sriniva
Chi

Uni
Pittsburgh

Ian Foster
Argonne/ University of Chicago

hn McCalpin
ACC

Lei Huang
TACC

Visualization Aids

Joseph Insley
Argonne

ssica Liu
Cerebras Inc

Silvio Rizzi
Argonne

bbiah
ras Inc.

rdy

Victor Mateevisti
Argonne

Sarah Harris*
University of Leed

Heng Ma
Argonne

Anda Trifan
UIUC/ Argonne

Geoffr
Univer

Venkatram Vishwanath*
Argonne

D e Gorgun
UIUC/ Argonne

Rick Stevens
Argonne/ University of Chicago

ne*
UIUC

Tom Burnley
Scientific Facilities and Technology Council

Anda Trifan
UIUC/ Argonne

Noah Trebesch

n

Murali Emani
Argonne/ University of Chicago

Emad Tajkhorshid*
UIUC

Jim Philipps
UIUC

Janet Knowles
Argonne

TACC

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

NERSC

Argonne NATIONAL LABORATORY | 75 1946–2021

# Acknowledgements

## Funding

- DOE- National Virtual Biotechnology Laboratory (NVBL)
- Exascale Computing Project Cancer Deep Learning Environment (CANDLE)
- Exascale Workflows Project (ExaWorks)
- Codesign for Online Data Reduction and Analytics (CODAR)
- DOE Codesign for multimodal AI approaches
- NSF MRI: Multi-modal imaging

## Computing Time

- Argonne Leadership Computing (Theta/ Theta-GPU/ AI-testbed)
- Oak Ridge Leadership Computing (Summit)
- Texas Advanced Computing Center (Frontera/Longhorn)
- National Energy Research Computing Center (Perlmutter)

## Colleagues

- Dan Stanzione
- Paul Lim
- Jack Dillespie
- Dwight Nissley
- Shantenu Jha
- Mike Papka
- Katherine Riley
- Bronson Messer
- Harry Petty
- Rommie Amaro

## Questions/Comments

ramanathana@anl.gov