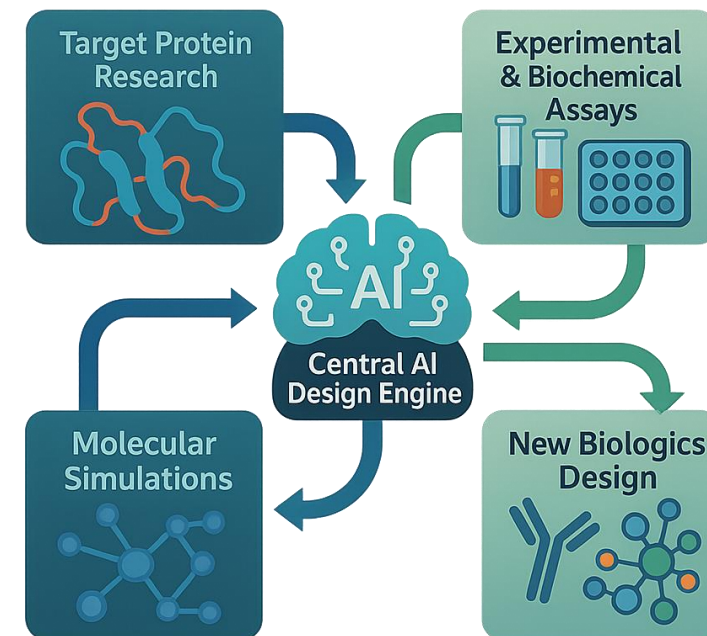


Generative AI, multi-scale modeling and simulations with experiments-in-the-loop for biotherapeutic design



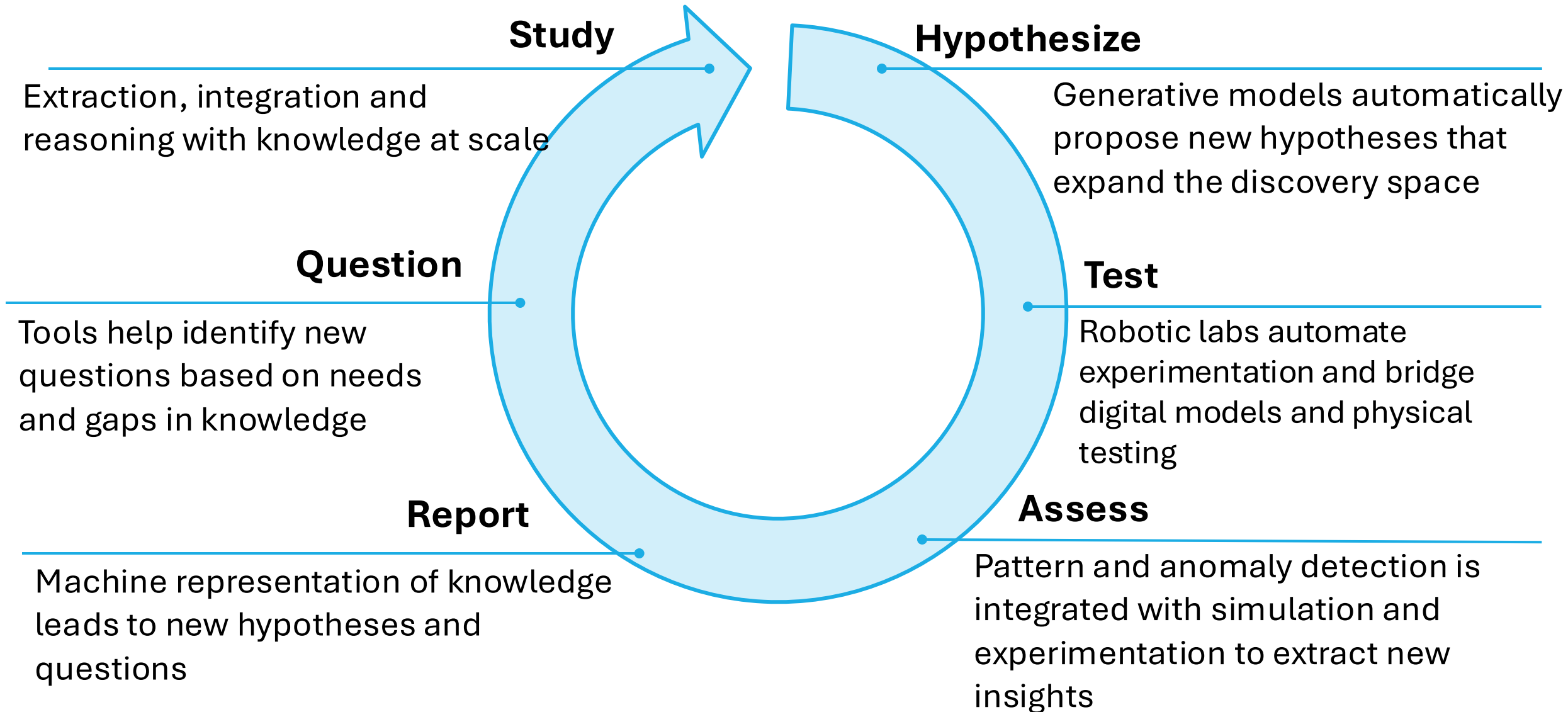
Arvind Ramanathan/ ramanathana@anl.gov

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

<https://ramanathanlab.org/>

<https://github.com/ramanathanlab>

Accelerating discovery using AI assistants



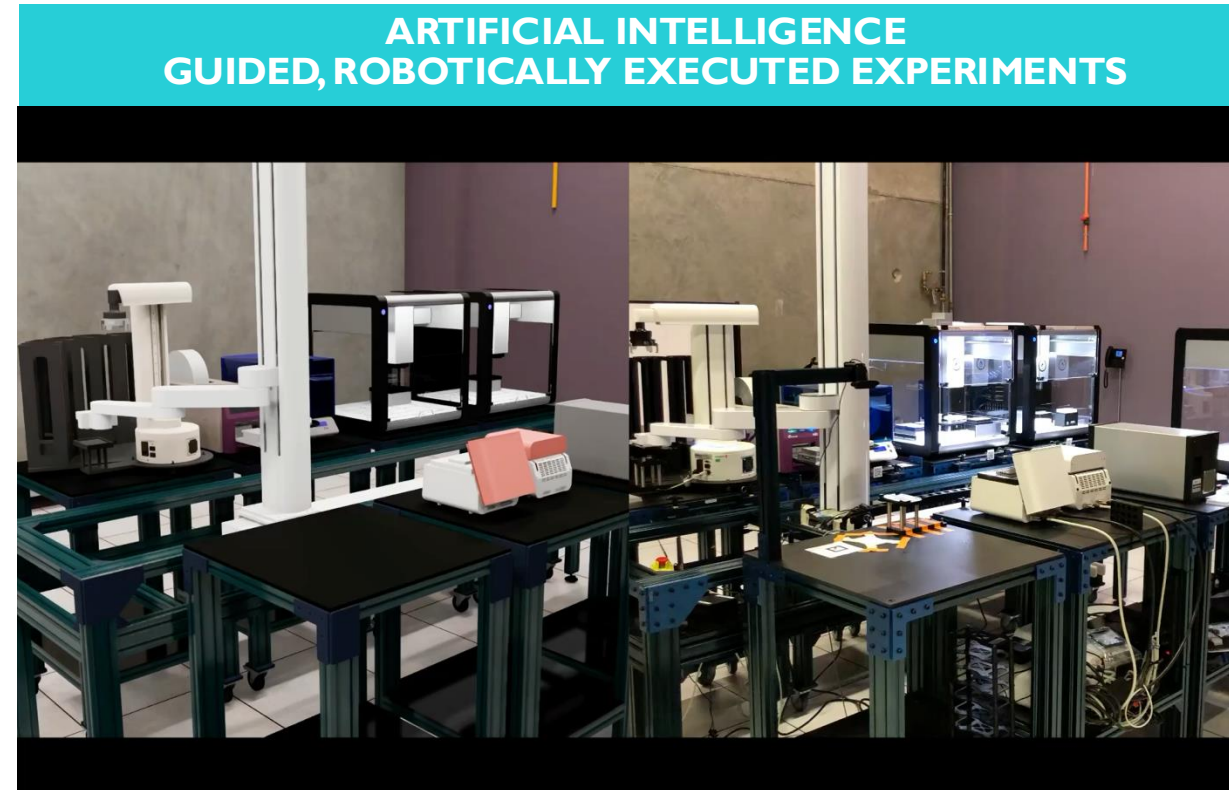
Autonomous Discovery @Argonne

- **The vision**

- A system that starts with a high-level description of a hypothesis and autonomously carries out computational and experimental workflows to confirm or reject that hypothesis
- **Use of AI in robotics and simulations to close the loop** on planning, execution, and analysis of experiments

- **Builds on**

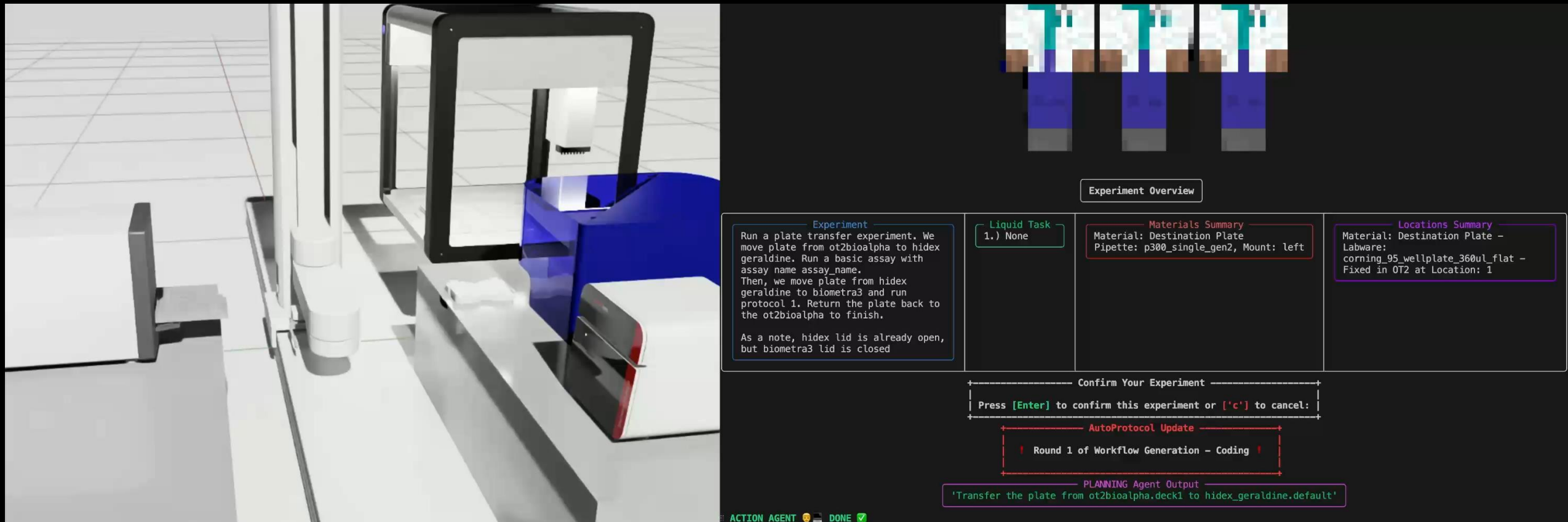
- **AI approaches to planning** (multiple steps), and integration of results, causality, etc.
- **Machine learning/simulation** to design and predict properties and outcomes
- **Automation of experimental protocols** (robotic steps and workflows)
- **Active Learning or RL** for selection of next experimental targets, etc.



<https://github.com/anl-sdl/>

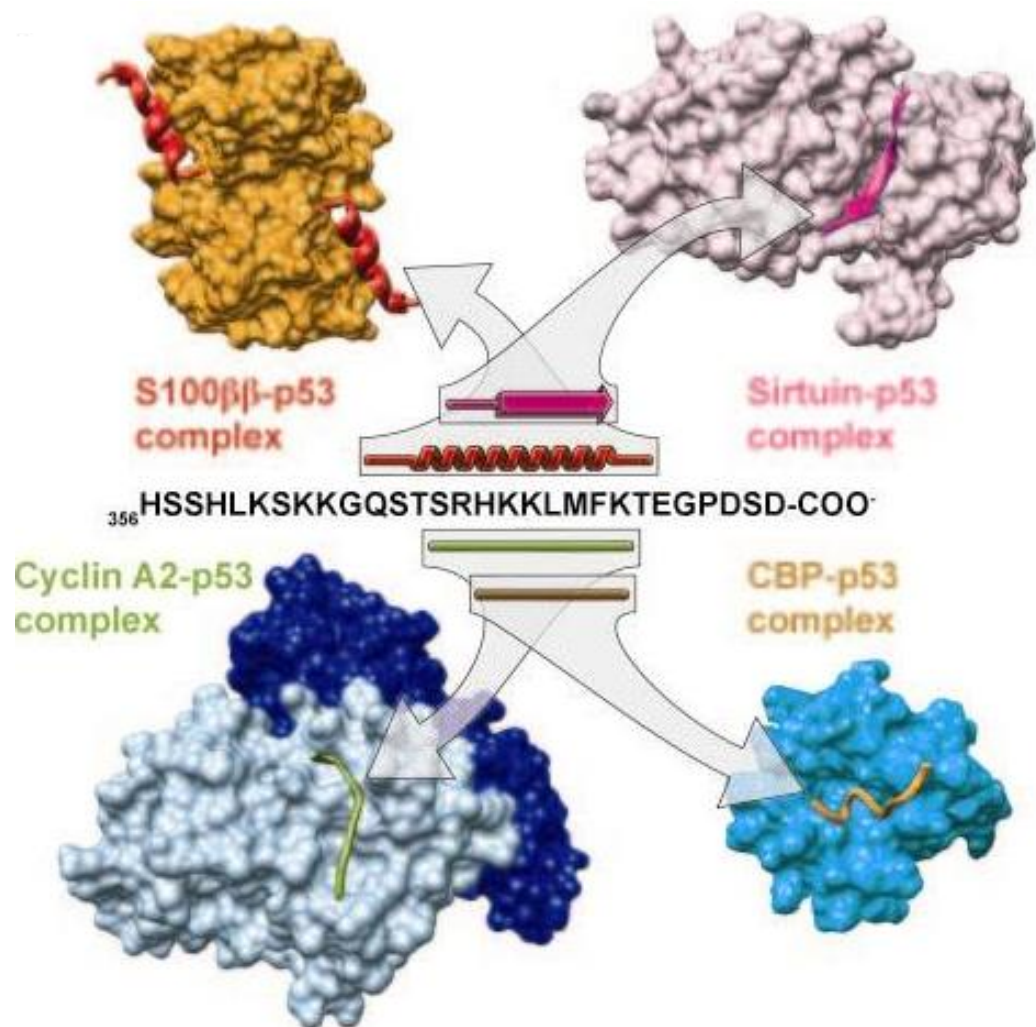
<https://www.cs.uchicago.edu/~rorymb/>

Agentic implementation of laboratory workflows

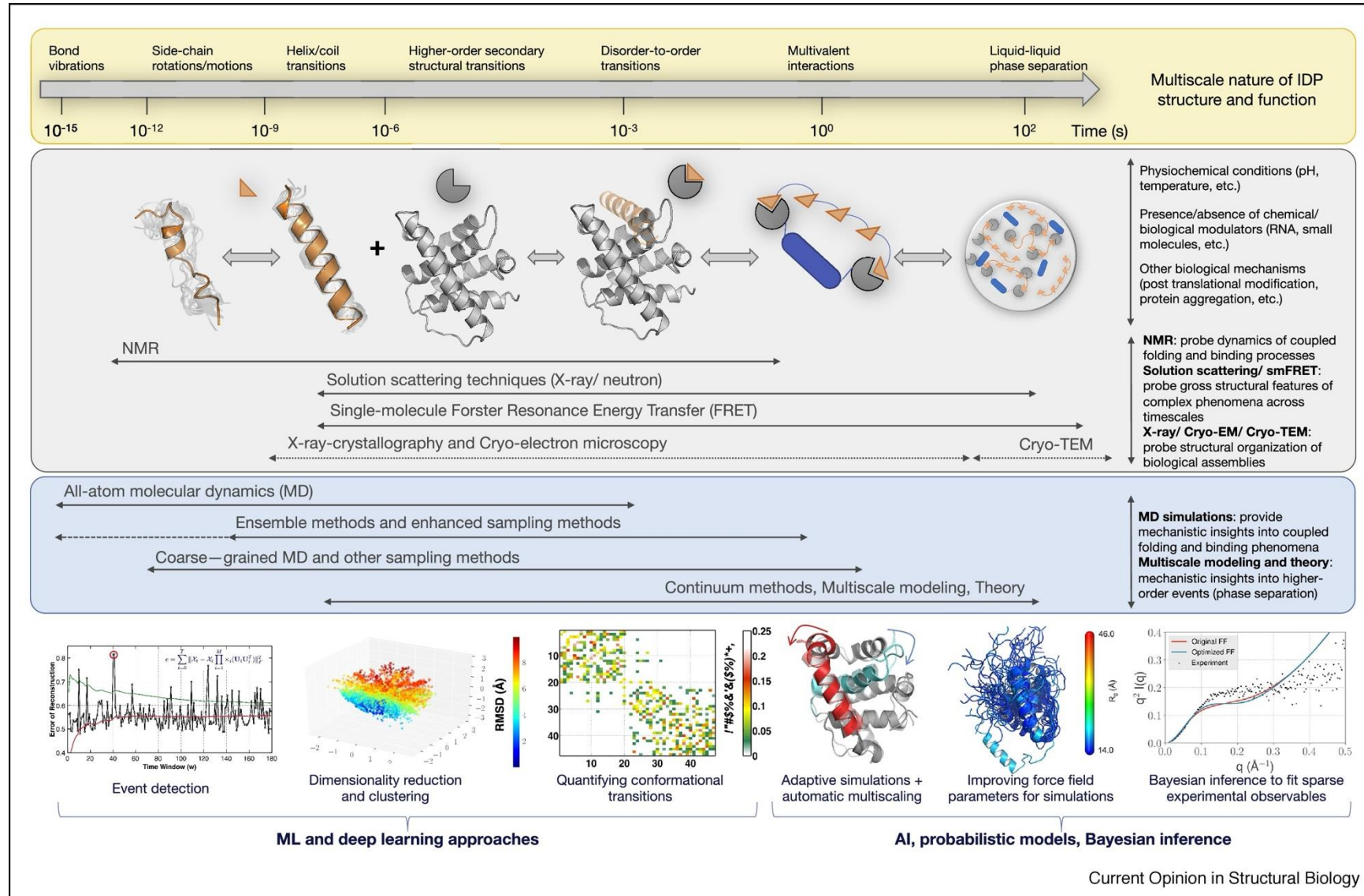


Disordered proteins span over 30% of the human proteome and are important drug targets

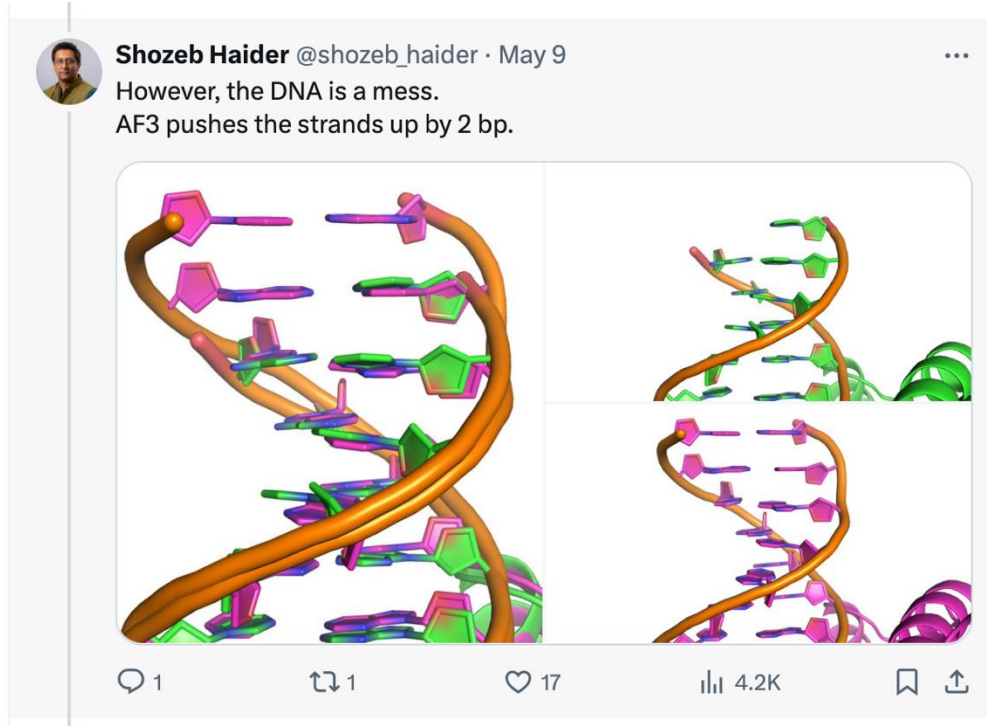
- Proteins without a stable tertiary structure:
 - High flexibility
 - Adaptable binding interfaces
- 65% of these proteins are involved in diseases:
 - Cancer
 - Neurodegenerative
 - Cardio-vascular
 - Diabetes
- We want to largely target the “undruggable” genome as part of this project
- This is not restricted to just human genomes; we are looking at viral, bacterial, fungal pathogens (for infectious diseases)



Background: Biomolecular dynamics spans multiple length- and time-scales...

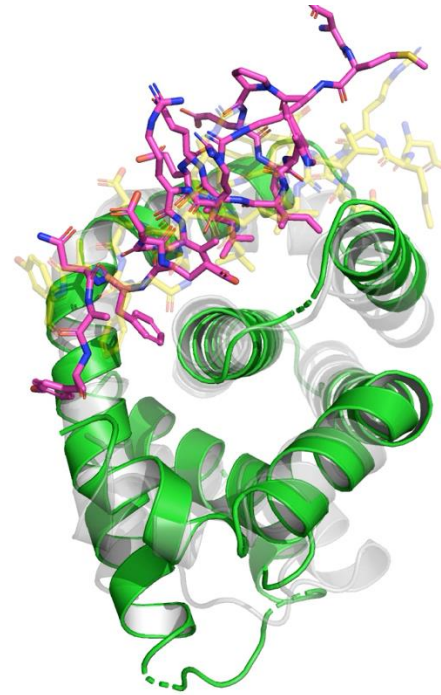


... current AI ecosystem of tools are not biophysically aware



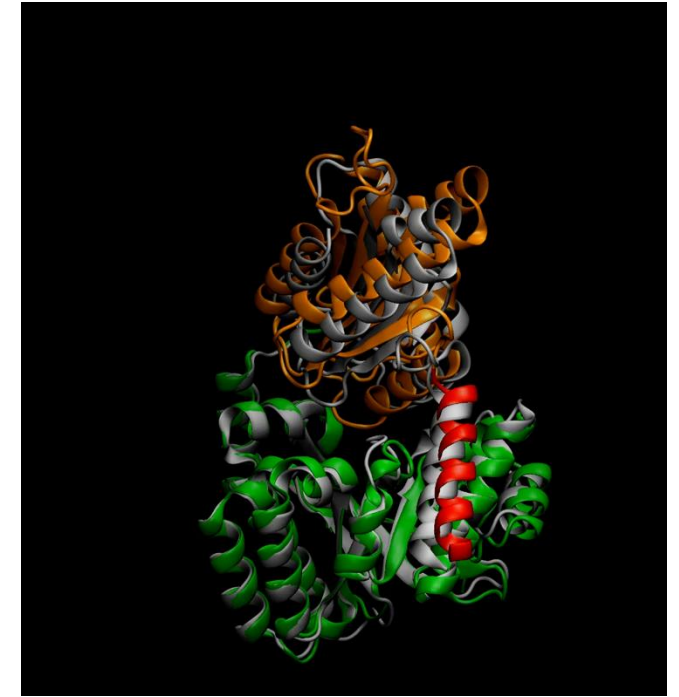
AlphaFold3: Sometimes remarkable results have interesting failure modes...

AlphaFold3: <https://www.nature.com/articles/s41586-024-07487-w>



DiffDock/NeuralPlexer: none of them able to associate protein-protein interactions or protein-ligand interactions accurately

DiffDock: <https://arxiv.org/html/2402.18396v1>
NeuralPlexer: <https://www.nature.com/articles/s42256-024-00792-z>



BioEmu: ~200 ms of simulation training time, but applications are still limited

Scalable emulation of protein equilibrium ensembles with generative deep learning

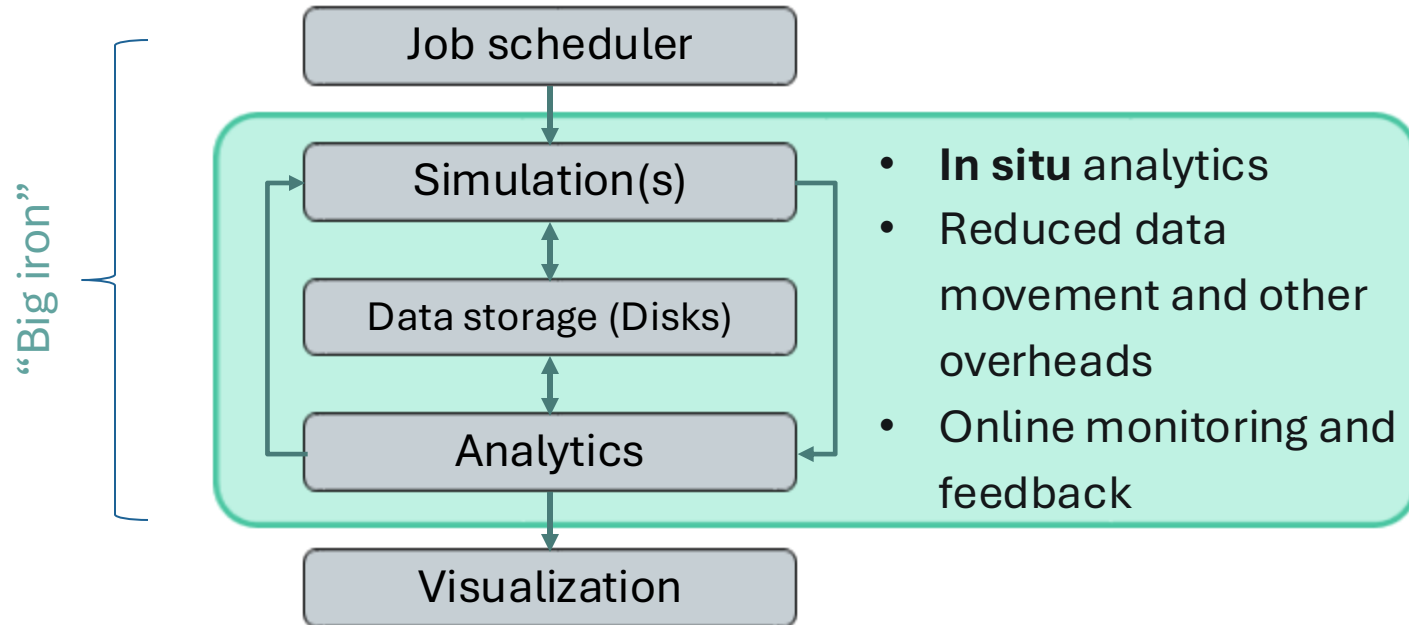
Sarah Lewis, et al., Cecilia Clementi, Frank Noé
<https://doi.org/10.1101/2024.12.05.626885>

Outline

- Can we use AI to effectively learn biophysically relevant features automatically?
 - DeepDriveMD: Accelerating biomolecular simulations with surrogate models
 - StreamAI-MD: Heterogeneous computing to accelerate simulation workflows
- Can we use AI to effectively accelerate length- and time-scales accessible to MD simulations to bridge multi-modal experimental techniques?
 - Intelligent Resolution: Integrating cryo-EM with X-ray crystallography using AI-driven simulation workflows
 - AA2CG2AA: Agentic AI to allow all-atom to coarse-grained to all-atom simulation campaigns
- Can we use AI to guide experimental campaigns for enabling biological systems design?
 - Experiments in loop for designing better enzymes by integrating multi-modal data
 - Using AI to simulate actual labs

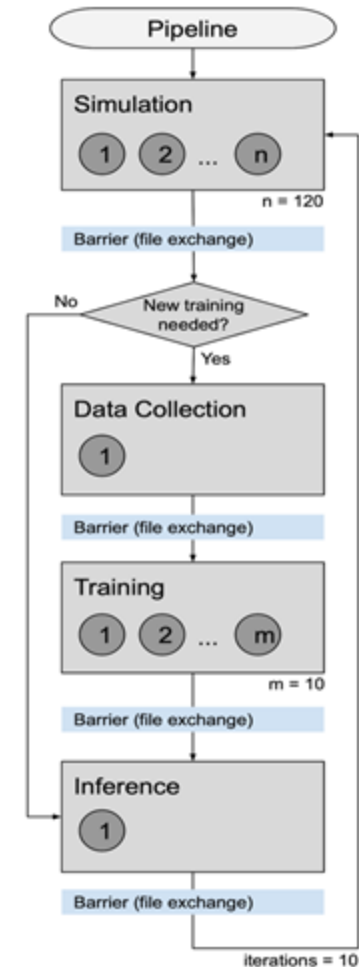
Standard simulation approaches face significant data movement and parallel analytics challenges

Need for interleaving analytics (AI/ML) + Simulations (HPC)



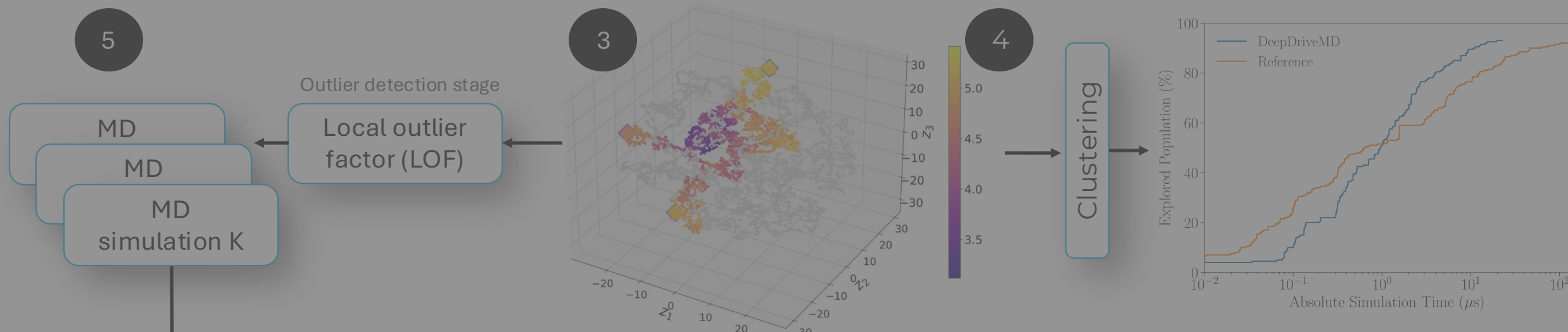
- Large simulations generate > **O(100 TB)** of data
- Humanly impossible to peek into “biologically” interesting events!
- <http://deepdrivemd.github.io>
- Ma, Lee, et al. PARCO (2019)
- Lee, Ma, et al. Workshop on Deep Learning on Supercomputers, Supercomputing (2019)

Ensemble Toolkit Workflow



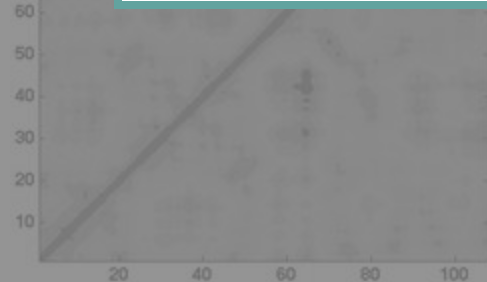
Interesting conformational states sampled

Tracking conformational states sampled

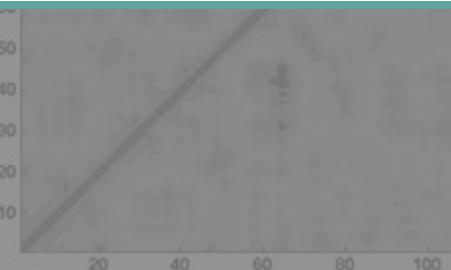
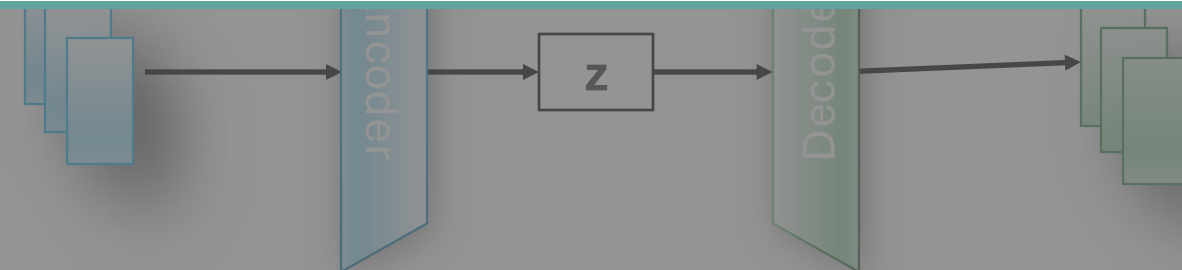


Computational challenges

- Representation of contact maps as sparse matrices
- Parameters for training – $O(10^{12}) \rightarrow$ harder to train



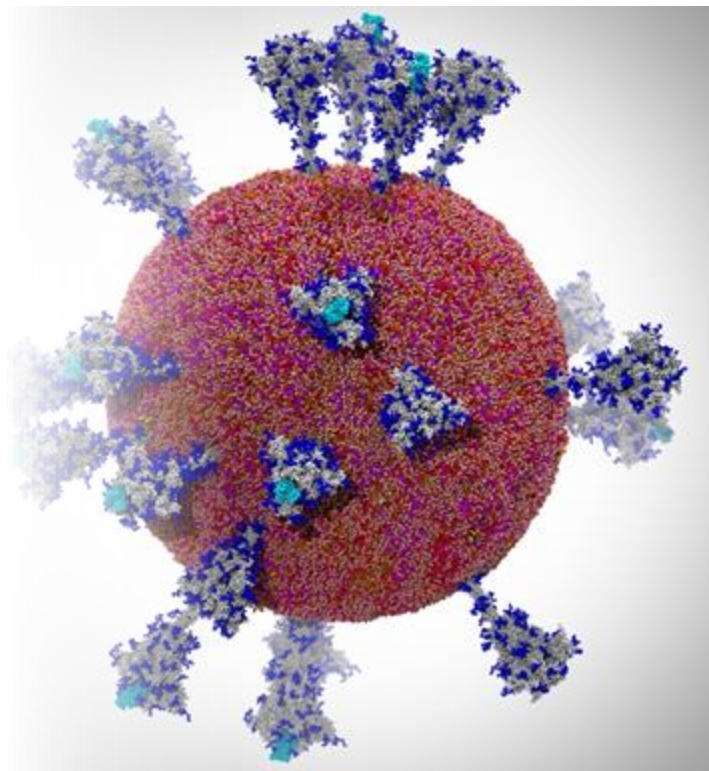
Input Contact Matrix



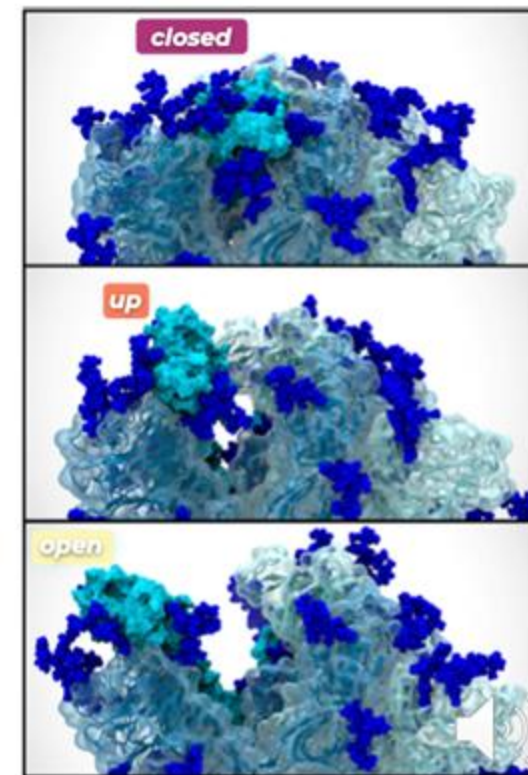
Reconstructed Contact Matrix

- Bhowmik, Gao, et al. BMC Bioinformatics (2018)
- Romero, Ramanathan, et al. Proc. Natl. Acad. Sci. USA (2019)

Deep learning can identify reaction coordinates for complex conformational transitions



How does the spike protein open to fuse with human cells?



Low-dimensional latent representation learned by convolutional variational autoencoder

T Sztain*, SH Ahn*,...**LTC**, R Amaro. *Nat. Chem.* (2021).

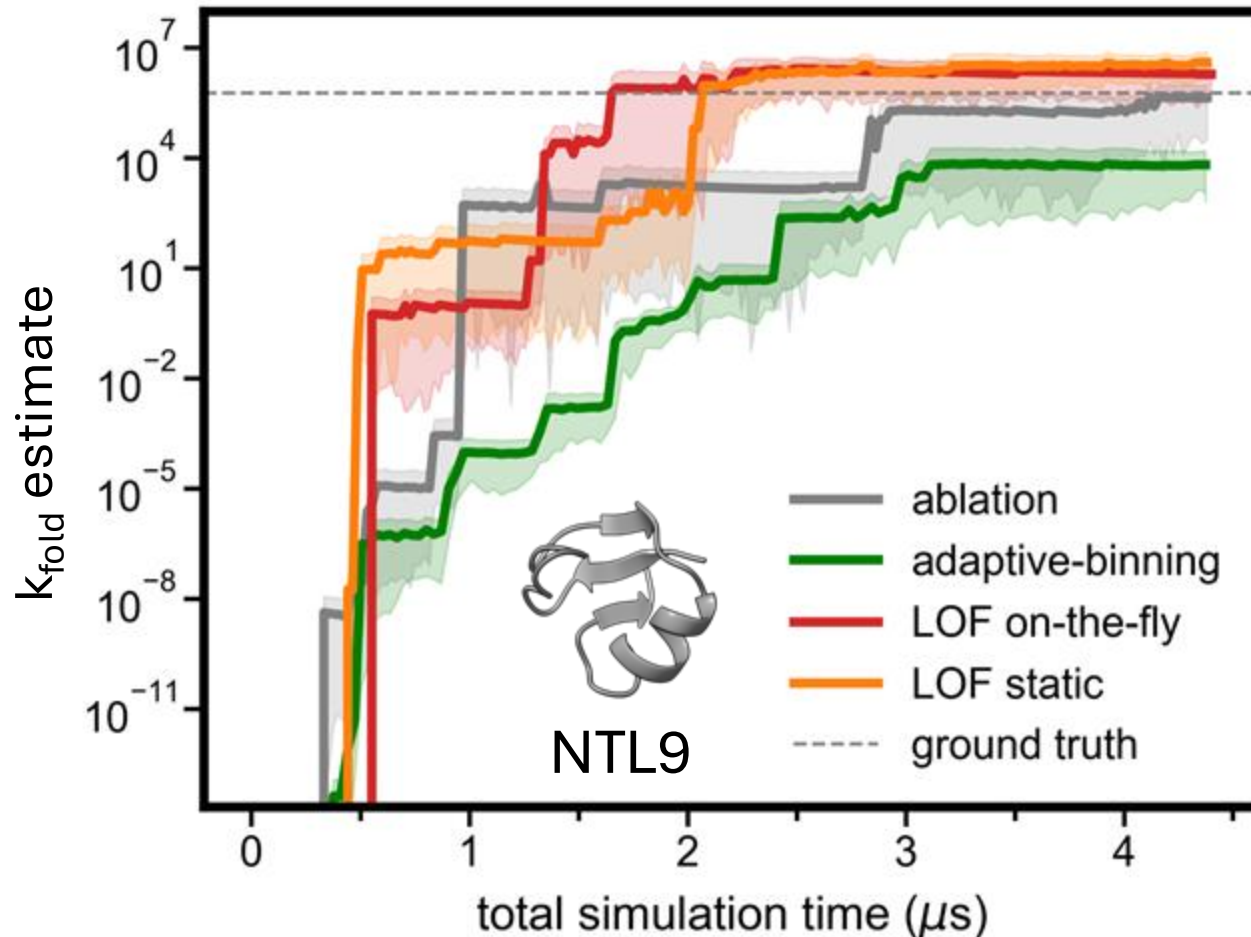
A Trifan, T Sztain, et al, LTC, A. Ramanathan, R. Amaro *IJHPCA* (2020, 2021)

A. Ramanathan, et al *Current Opinion in Structural Biology* (2019)

D. Bhowmik, et al, *BMC Bioinformatics* (2017)

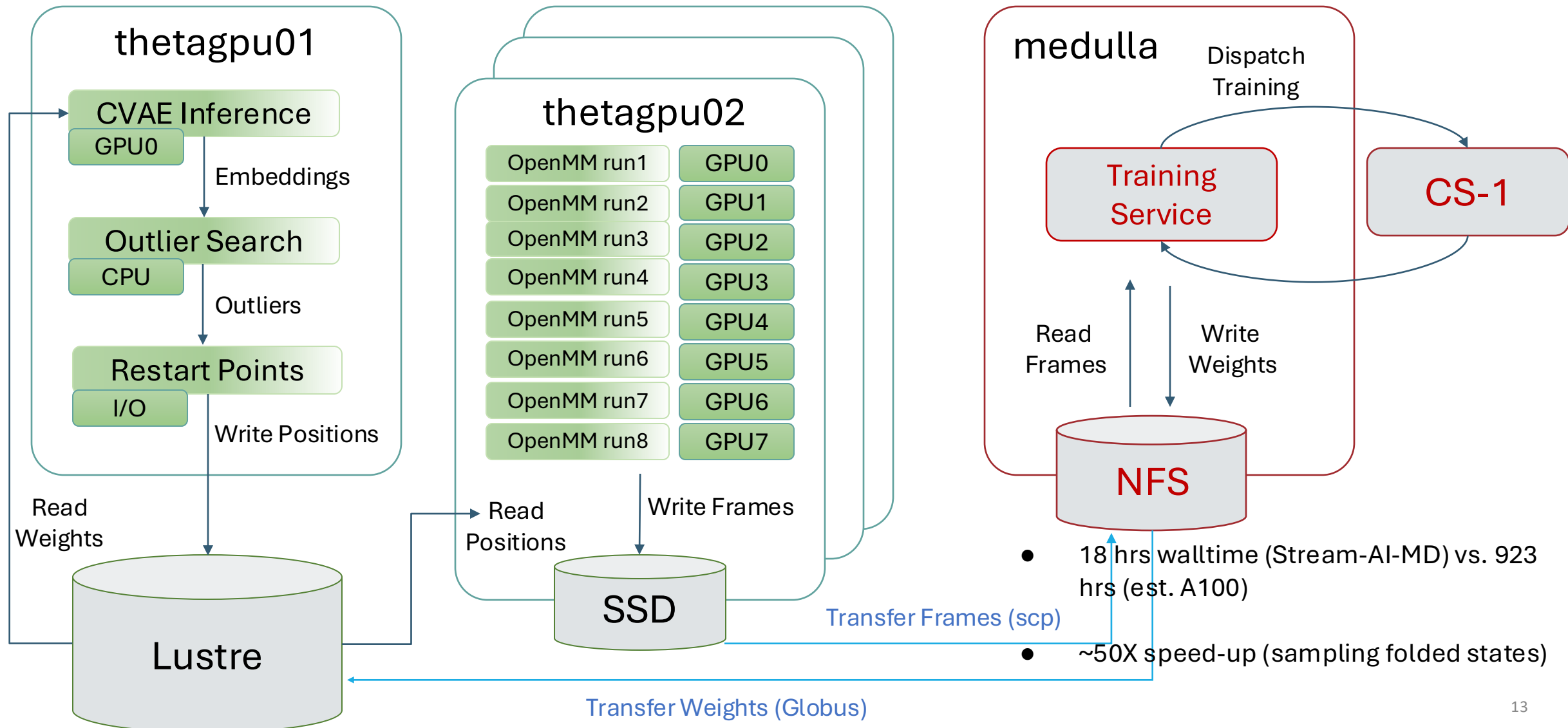
software/notebooks: <https://github.com/ramanathanlab/deepdrivemd>

DeepDriveMD enables 4-fold acceleration of sampling effectiveness for FSD-EY ($\beta\beta\alpha$) folding




- Embedding states into the VAE latent space and clustering with k-means keeps a constant definition of the number of states sampled enabling fair comparison between simulations
- The ML + RMSD strategy reaches 80% sampling at least 4 fold faster than Anton-1 simulations
- Integrating with weighted ensemble techniques we can get access to kinetics

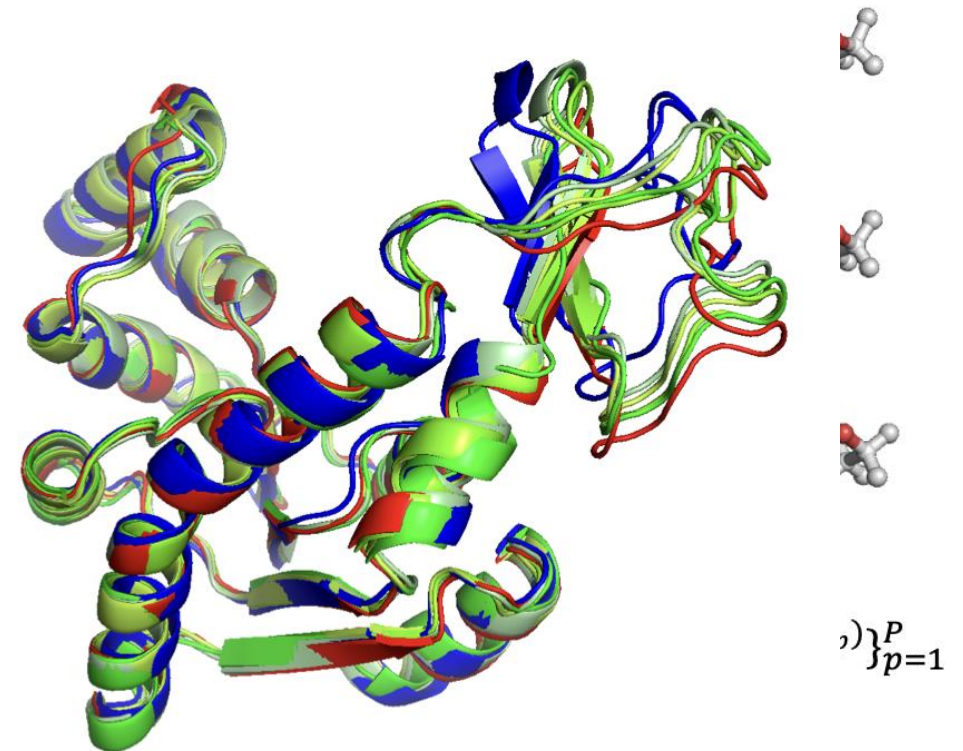
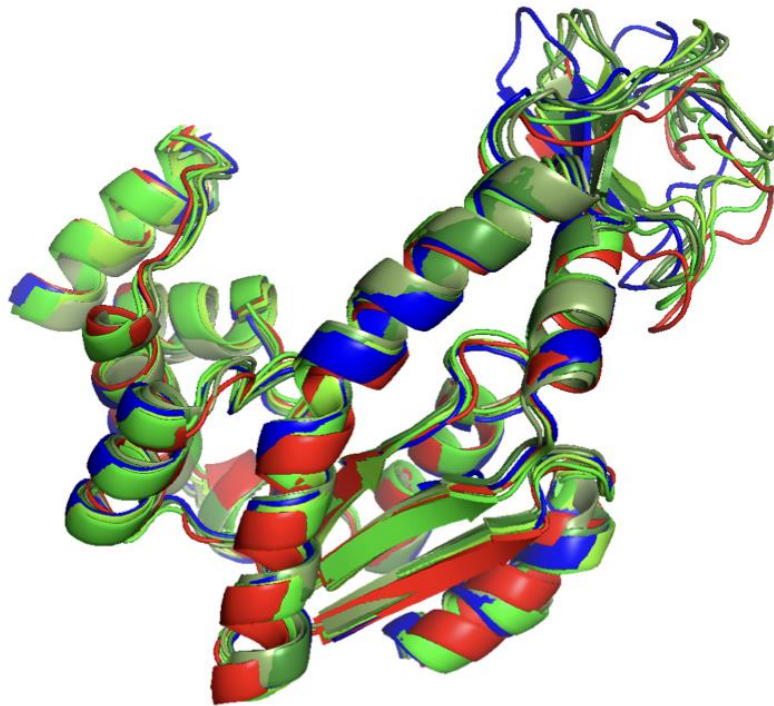
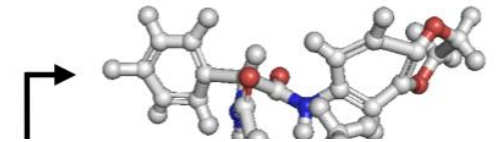
Bridging ThetaGPU + CS-1 with Stream-AI-MD



Accelerating MD simulations with surrogate models provides $\sim 10^{3-5}$ speedup for protein systems

Table 6. F-MSE on AdK equilibrium trajectory dataset.

Linear	RF	MPNN	EGNN	EGHN	EGNO	EGHNO	Blocks
2.890	2.846	2.322	2.735	2.034	2.231	1.801	

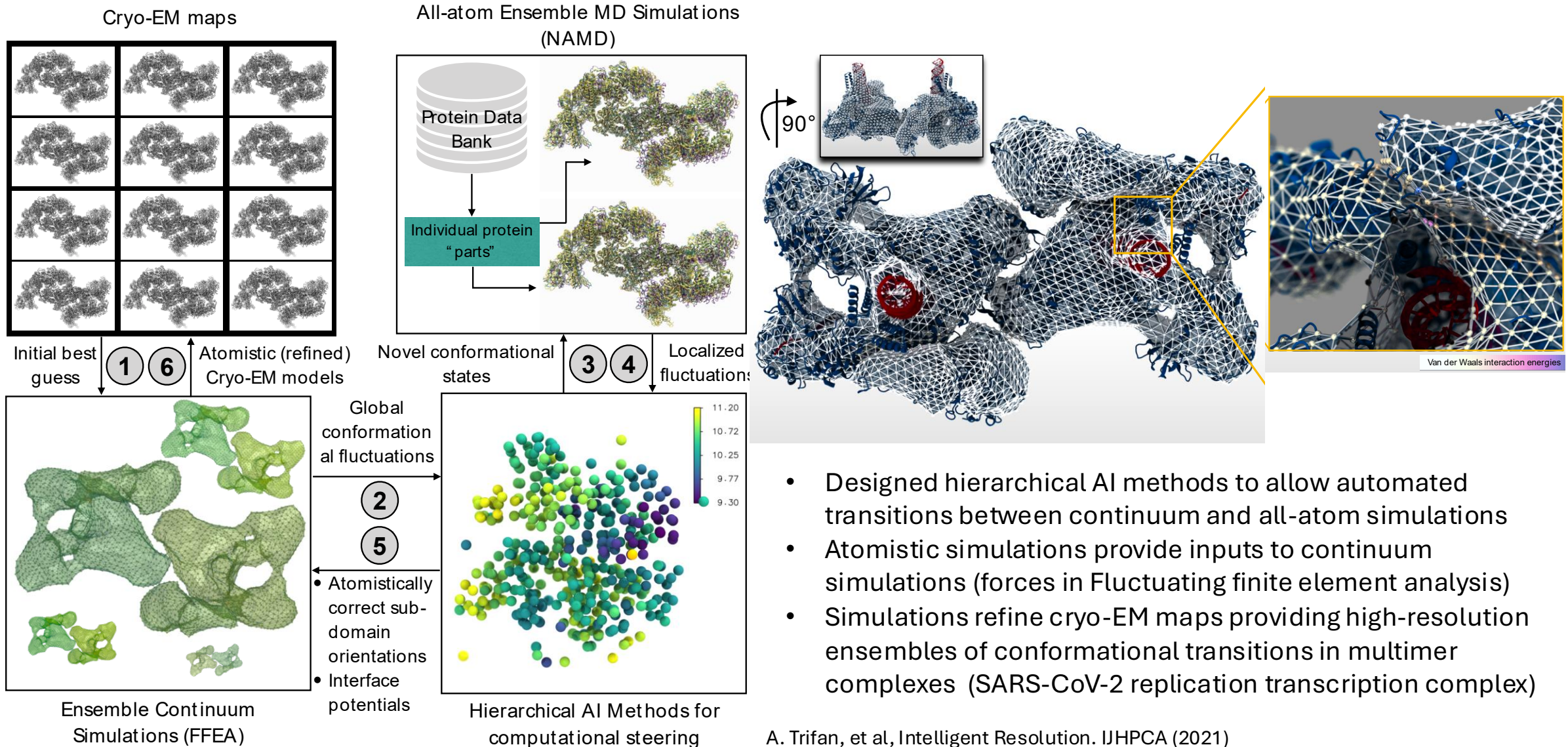


Minkai Xu, Jiaqi Han, Aaron Lou, Jean Kossaifi, Arvind Ramanathan, Kamyar Azizzadenesheli, Jure Leskovec, Stefano Ermon, Anima Anandkumar

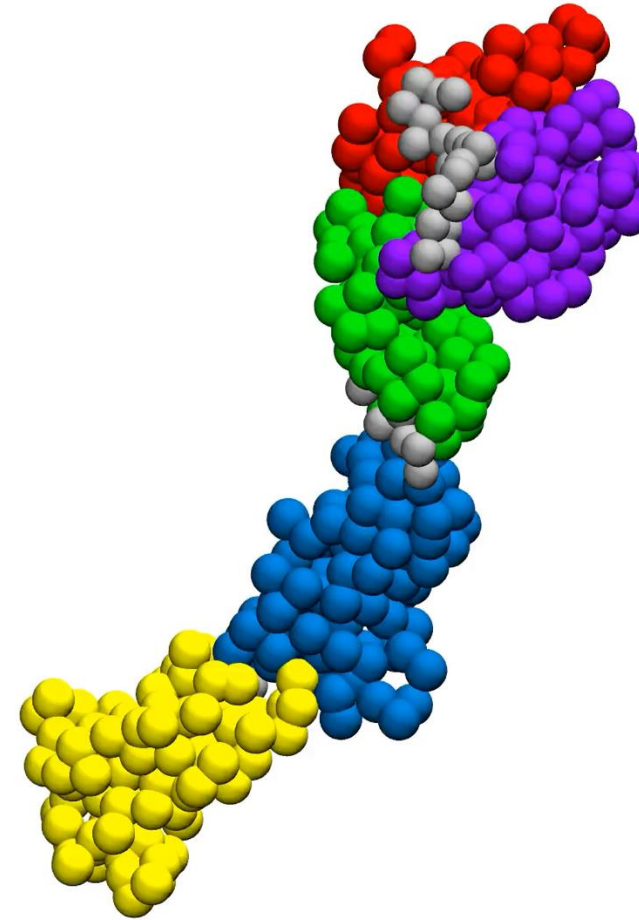
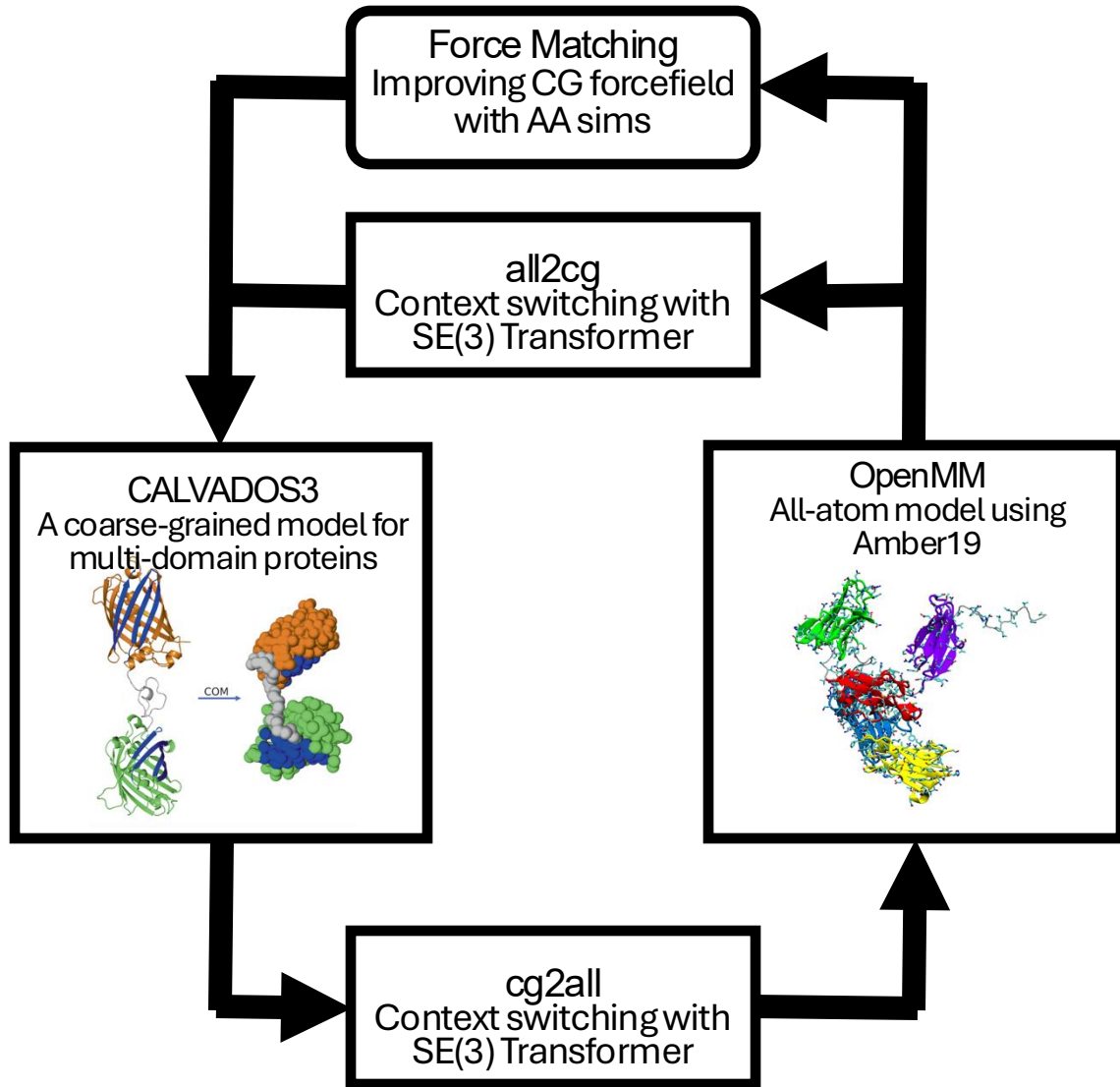
Outline

- Can we use AI to effectively learn biophysically relevant features automatically?
 - DeepDriveMD: Accelerating biomolecular simulations with surrogate models
 - StreamAI-MD: Heterogeneous computing to accelerate simulation workflows
- Can we use AI to effectively accelerate length- and time-scales accessible to MD simulations to bridge multi-modal experimental techniques?
 - Intelligent Resolution: Integrating cryo-EM with X-ray crystallography using AI-driven simulation workflows
 - AA2CG2AA: Agentic AI to allow all-atom to coarse-grained to all-atom simulation campaigns
- Can we use AI to guide experimental campaigns for enabling biological systems design?
 - Experiments in loop for designing better enzymes by integrating multi-modal data
 - Using AI to simulate actual labs

Continuum \Leftrightarrow all-atom simulations: using AI to guide and refine Cryo-electron microscopy data

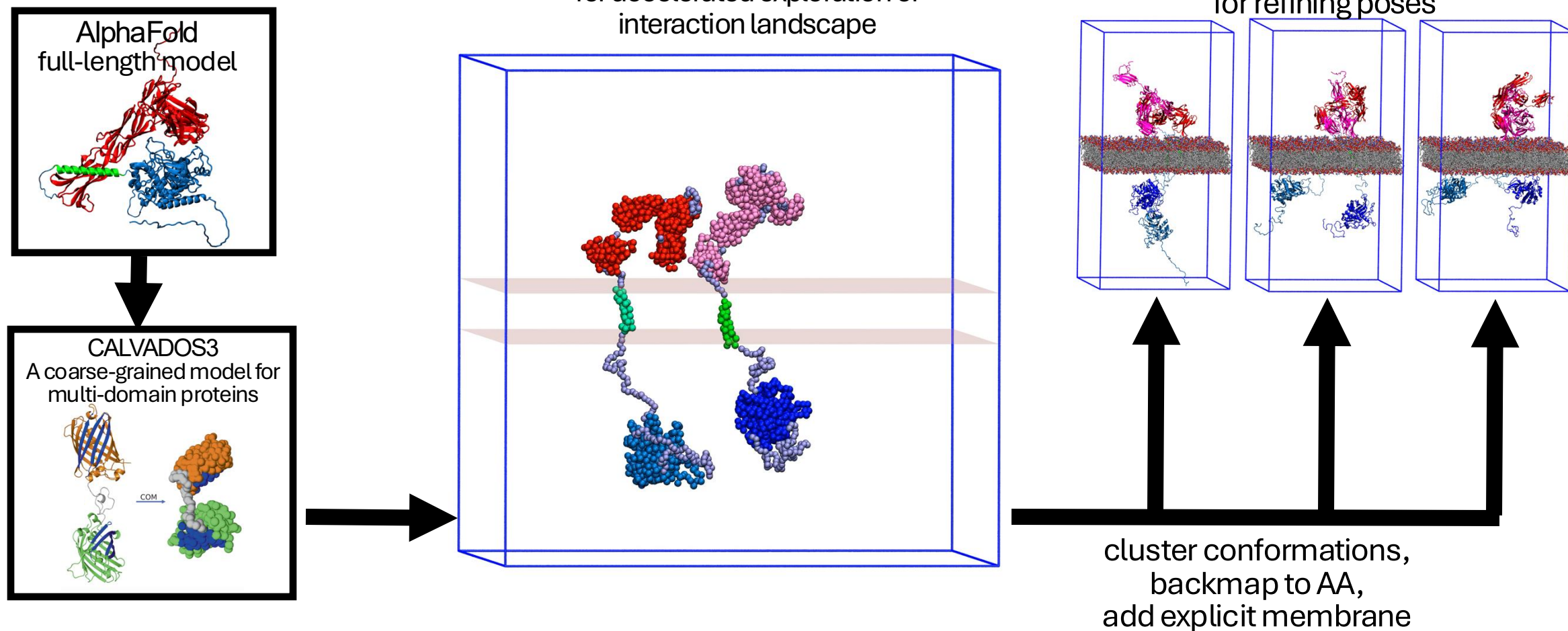


Multiscale modeling workflows switching between all-atom and coarse-grained representations for CSF1R model - extracellular domain



Full-length membrane-bound CSF1R model

Capturing protein-membrane interactions and complexation



Outline

- Can we use AI to effectively learn biophysically relevant features automatically?
 - DeepDriveMD: Accelerating biomolecular simulations with surrogate models
 - StreamAI-MD: Heterogeneous computing to accelerate simulation workflows
- Can we use AI to effectively accelerate length- and time-scales accessible to MD simulations to bridge multi-modal experimental techniques?
 - Intelligent Resolution: Integrating cryo-EM with X-ray crystallography using AI-driven simulation workflows
 - AA2CG2AA: Agentic AI to allow all-atom to coarse-grained to all-atom simulation campaigns
- Can we use AI to guide experimental campaigns for enabling biological systems design?
 - Experiments in loop for designing better enzymes by integrating multi-modal data
 - Using AI to simulate actual labs

Using the agentic co-scientist to summarize results...

Research cycle complete!
Generated 50 total hypotheses
Completed 30 tournament matches

Top 3 drug repurposing hypotheses:

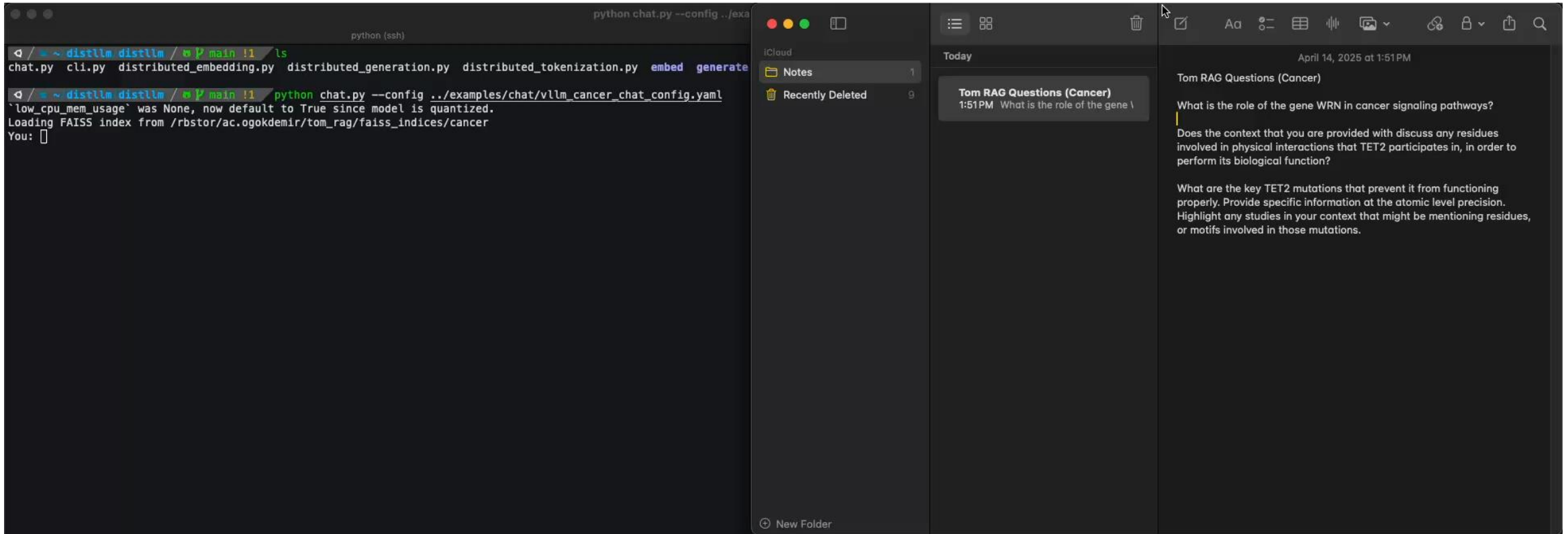
1. A biologic therapy using a cell-penetrating antibody-peptide conjugate is proposed to target and inhibit the SET domain of WHSC1, leveraging unique structural features of WHSC1 for cancer therapy. (Elo rating: 1258.4)
Content snippet: This hypothesis proposes the development of a biologic therapy comprising a cell-penetrating antibody conjugated to a peptide that specifically binds and inhibits the SET domain of WHSC1. The antibody...
2. The hypothesis proposes a nanobody-based biologic therapy targeting WHSC1 in cancer, utilizing nuclear localization signals and cell-penetrating peptides for specificity and efficient delivery. (Elo rating: 1244.4)
Content snippet: This improved hypothesis proposes the development of a novel biologic therapy using engineered nanobodies fused to a nuclear localization signal (NLS), designed to selectively bind and inhibit the SET...
3. Develop a fusion protein therapy using DARPins and CPP to target WHSC1's SET domain with high specificity and efficient cellular/nuclear delivery. (Elo rating: 1231.3)
Content snippet: This hypothesis proposes the development of a biologic therapy using a bifunctional fusion protein. This fusion protein consists of a designed ankyrin repeat protein (DARPins) targeting the SET domain...

Research overview:

Title: Designing a Novel Biologic Therapy for Targeting WHSC1 in Cancer

Executive summary: This research overview outlines strategies for developing a biologic therapy targeting the SET domain of the WHSC1 protein, a histone methyltransferase implicated in various cancers. The focus is on designing a therapy that is specific, effective, and capable of nuclear penetration, leveraging structural biology techniques, cell-based assays, and in vivo models.

... providing novel (hopefully less hallucinated) insights across diverse data streams



Protein design needs “special” prompts to be effective but lacks clear, direct incorporation of design constraints

“I need a mostly hydrophobic protein with 7 transmembrane alpha-helices that can mediate cell signaling”



LLaMA

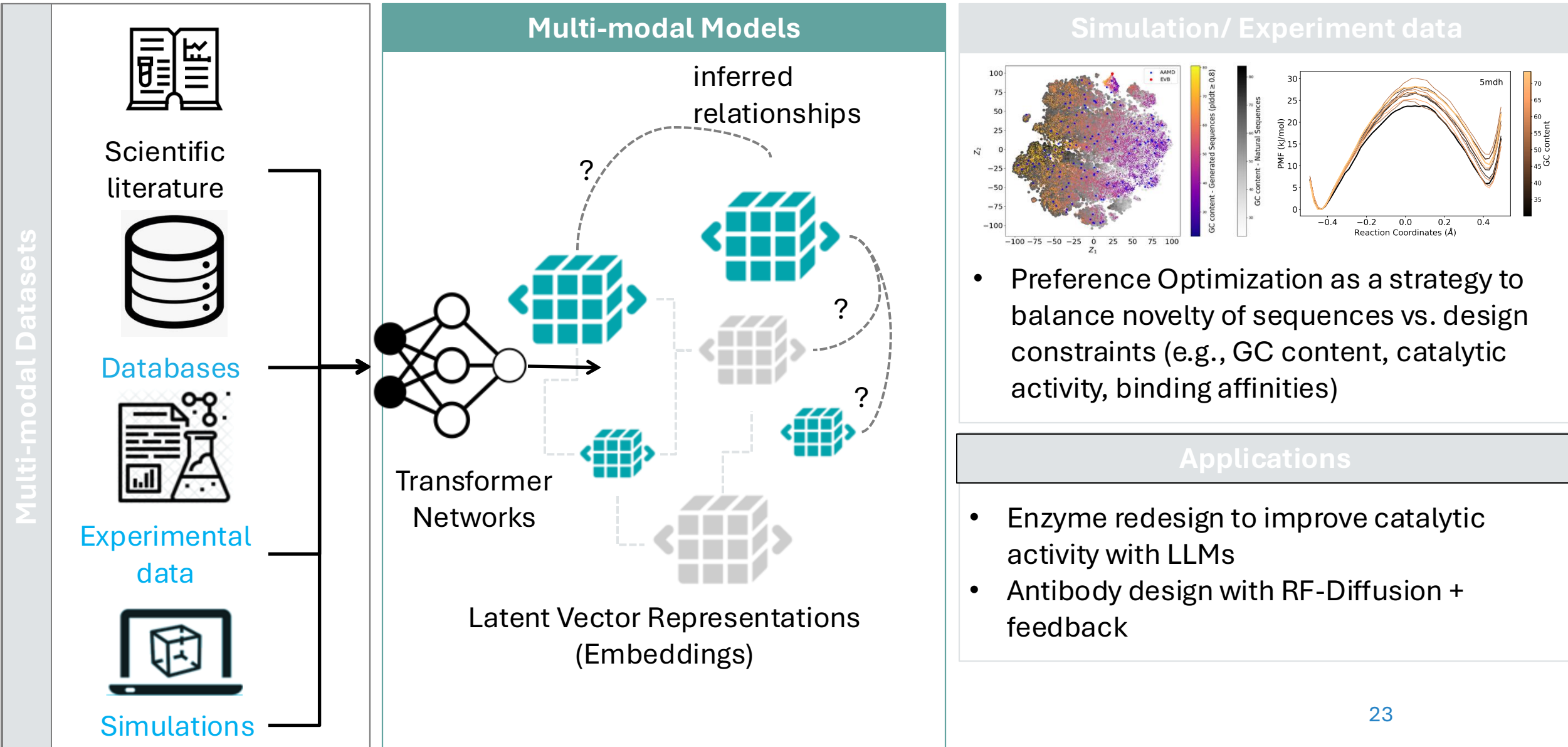
+ Proteins



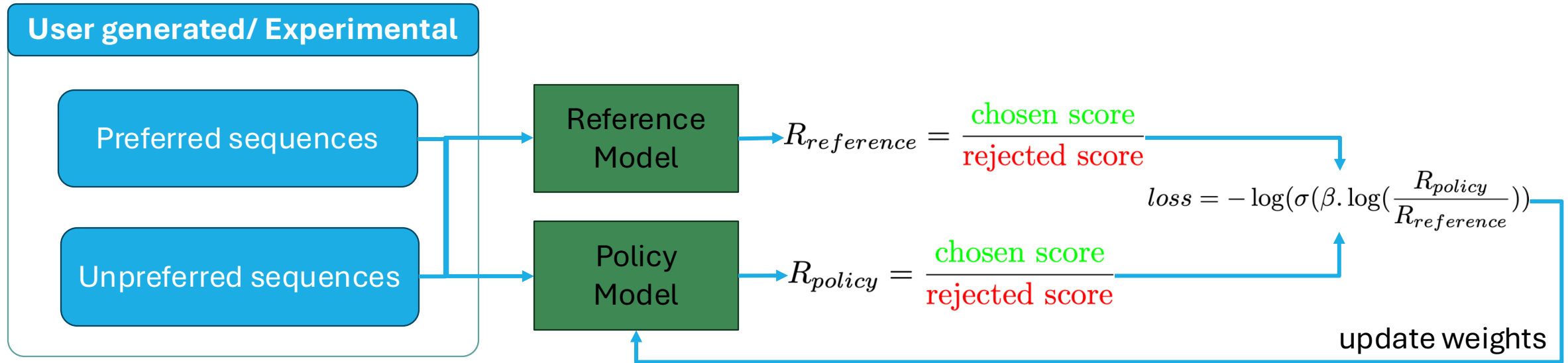
<https://310.ai/>

<https://www.evolutionaryscale.ai/blog/esm3-release>

Hypothesis: Multimodal language models can incorporate experimental observables to constrain the generation of protein sequences



Algorithmic innovation: Direct Preference Optimization strategy in two distinct modes aligns generative process w/ experimental data



- Incorporates feedback into the language model via preferences represented via user/ experimental data:
 - Mode 1: Encode user preferences via a classifier trained on experimental fitness datasets
 - Mode 2: Self-alignment where the language models “learns” the preferences via the generative process
- DPO loss function formulated to preferentially weigh ‘preferred’ samples over ‘unpreferred’ samples to update model weights
- Scaling of DPO implemented for the reference and policy models using Megatron-DeepSpeed framework:
 - fused kernels from NVIDIA’s Megatron-LM with the ZeRO optimization and Pipeline parallelism of DeepSpeed
 - FlashAttention-2 to improve throughput of training
 - sequence lengths of 512 and 1024 as the target protein families

Where is Attention going?

Head Pooling

Avg



Layer Pooling

None

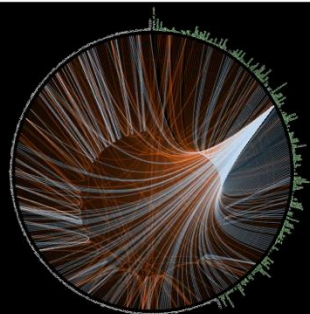


Select Layer

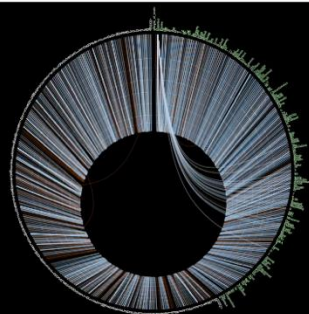
All



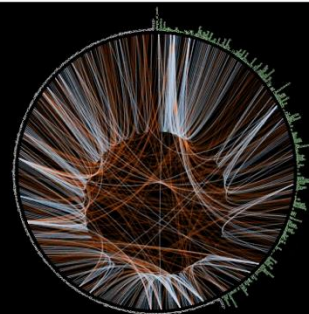
Submit



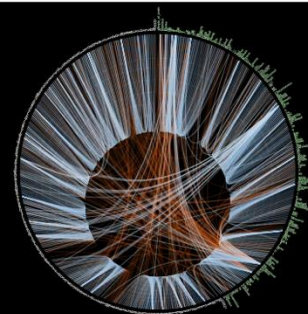
Layer 1



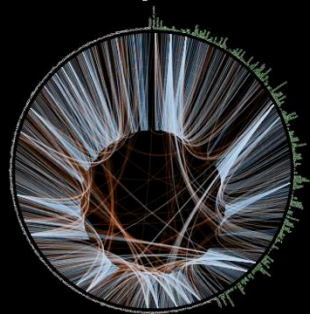
Layer 2



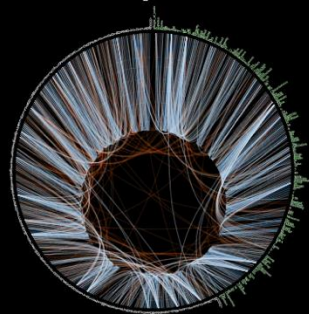
Layer 3



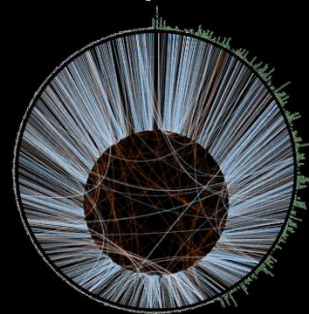
Layer 4



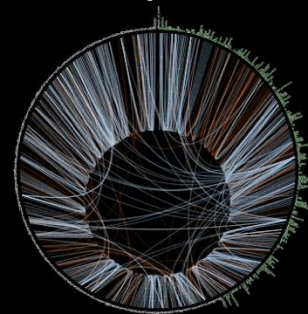
Layer 5



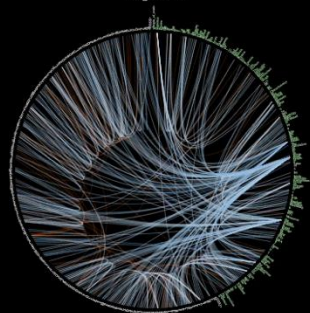
Layer 6



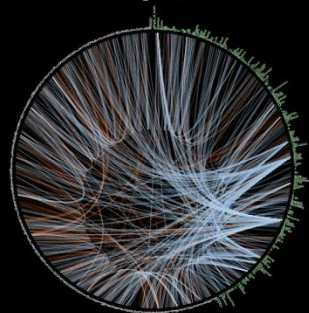
Layer 7



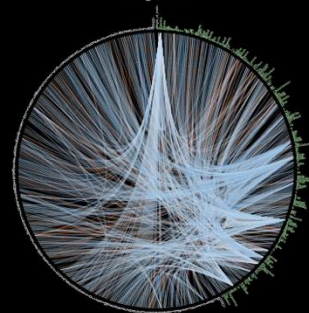
Layer 8



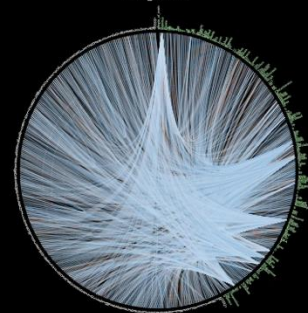
Layer 9



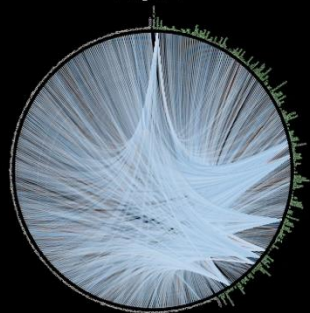
Layer 10



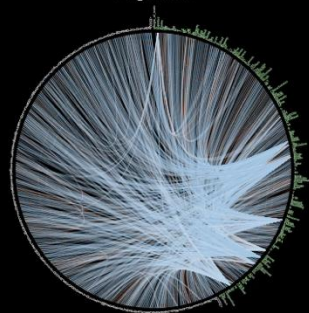
Layer 11



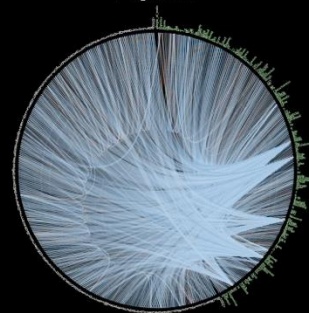
Layer 12



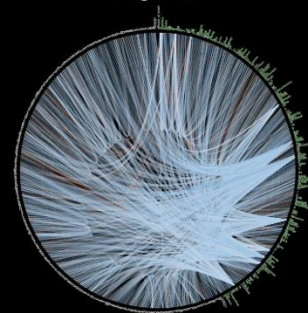
Layer 13



Layer 14

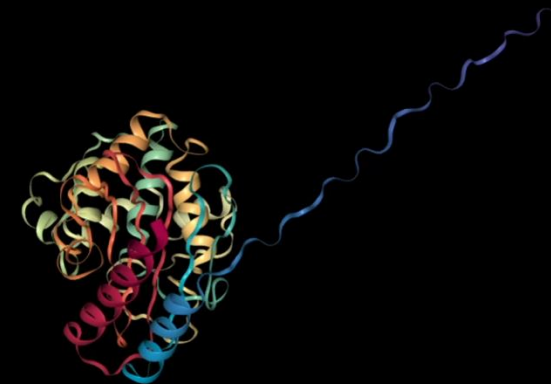


Layer 15

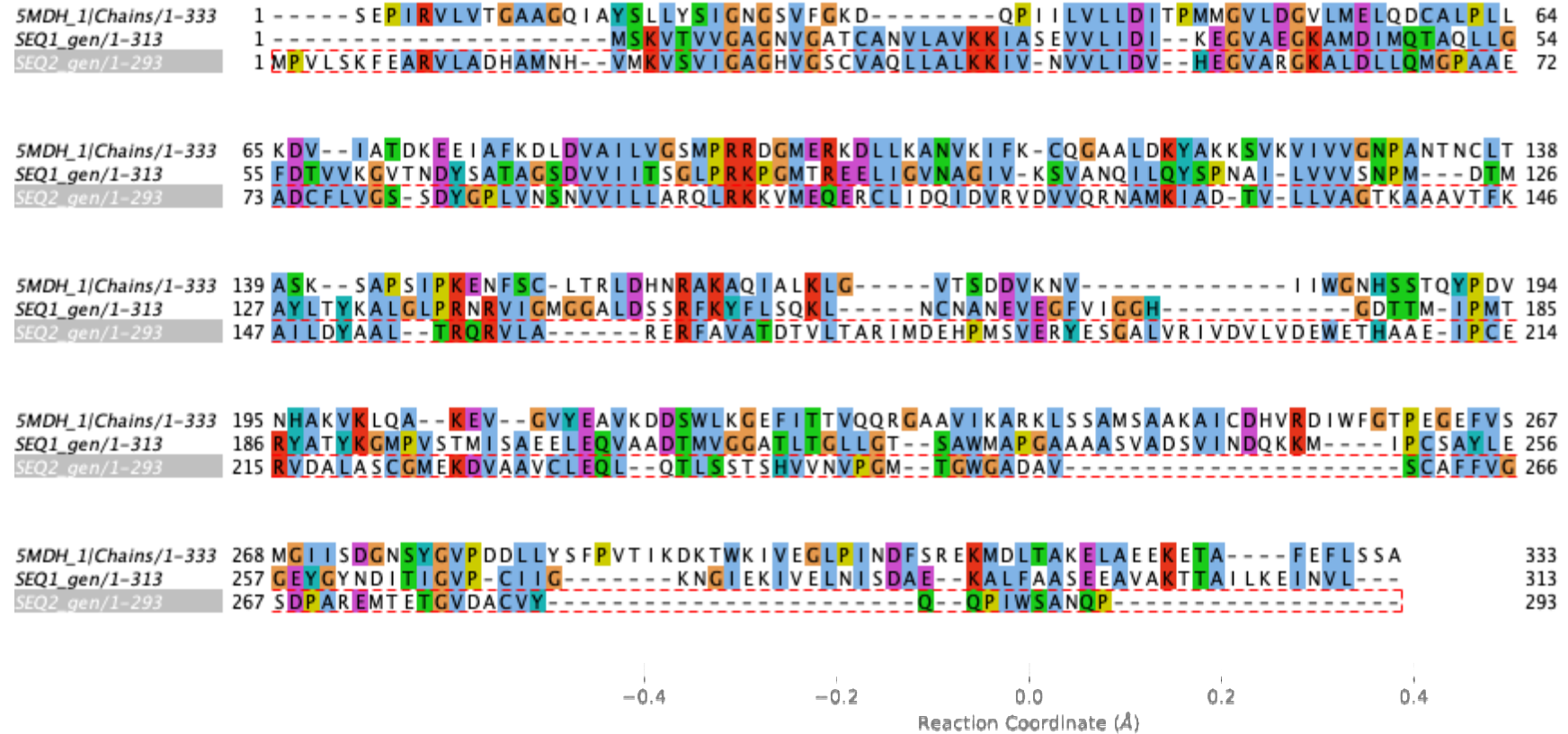
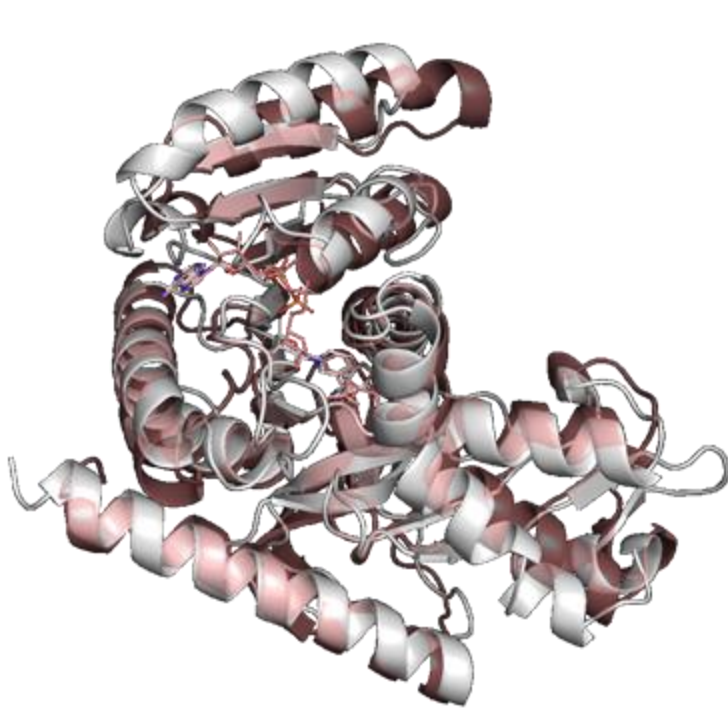


Layer 16

<|begin_of_text|> A sequence is known by the ID A 0 A 140 D 2 T 1 _Z IK V _M 207 . Its Property Name =< Deep Mut ational Sc anning (D MS) score > is valued at Property Val =< 1 . 22 > and shows this protein is Fitness =< fit > . It consists of 286 residues . The molecular structure has been determined to weigh 315 00 . 86 Da . Characteristics of this protein include an aromatic ity value of 6 . 29 and an instability index calculated at 39 . 8 , hint ing at its stable nature . The calculation of its is oe lectric point results in a value of 5 . 84 , while flexibility assessments reveal an average score of 1 . 0 with a standard deviation measuring 0 . 000 692 . Moreover , the protein exhibits a grand average of hy drop ath icity (GRA V Y) score of - 0 . 1 . Among its amino acid makeup , the predominant residues are Le uc ine , Alan ine , and Th reon ine , which constitute 36 . 36 % of its composition . The sequence is <SSEQ>MSIQHFRVALIPFFA AFCLPVFAHPETLVKVKDAETQLGARVGYIELDLNSGKILESFRPEERFPMMSTFKVLLCGAVLSRVDAGQEQLGRRIHYSQNDLVEYSPVTEKHLTDGMTVRELCSAAITMSDNTAANLLLTIGGPKELTAFLHNMGDHVTRLDRWPELNEAIPNDERDTTMPAAMATTLRKLLTGELLTLASRQLIDWMEADKVAGPLLRSA LPAGWFIADKSGAGERGSRGIIAALGPDGKPSRIVVIYTTGSQATMDERNRQIAEIGASLIKHW <ESEQ>

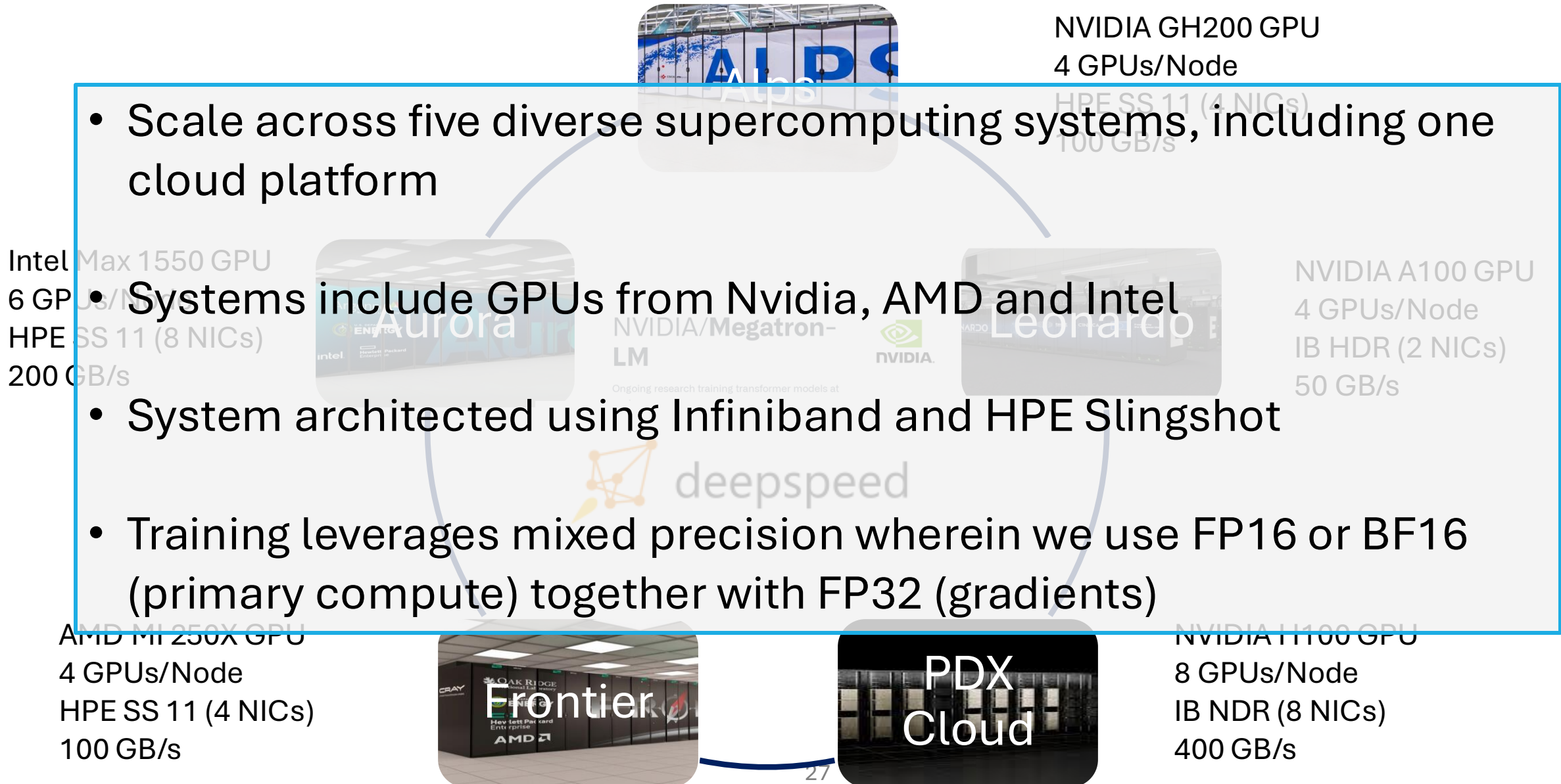


MProt-DPO for Malate Dehydrogenase (MDH) improves predicted activity ~3.0 fold increase with preliminary experimental confirmation

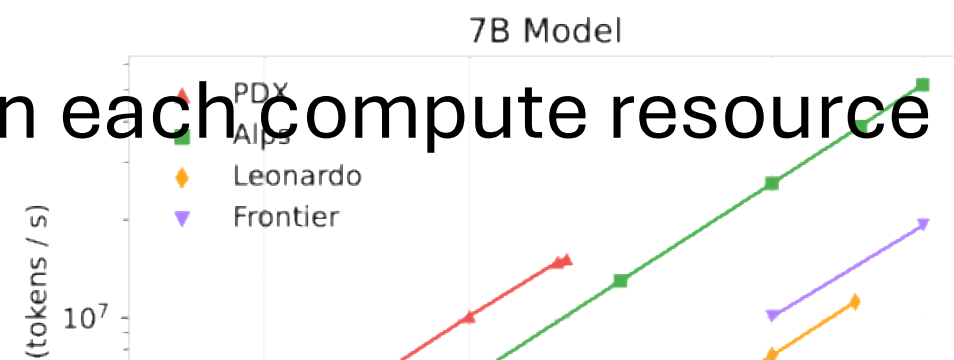


- Despite lower sequence similarity, key motifs in the structures are conserved; average sequence lengths 300, median ~25 residues
- 48 variants validated in the laboratory of which 13 sequences exhibit enhanced MDH activity compared to wildtype designs (round 1)

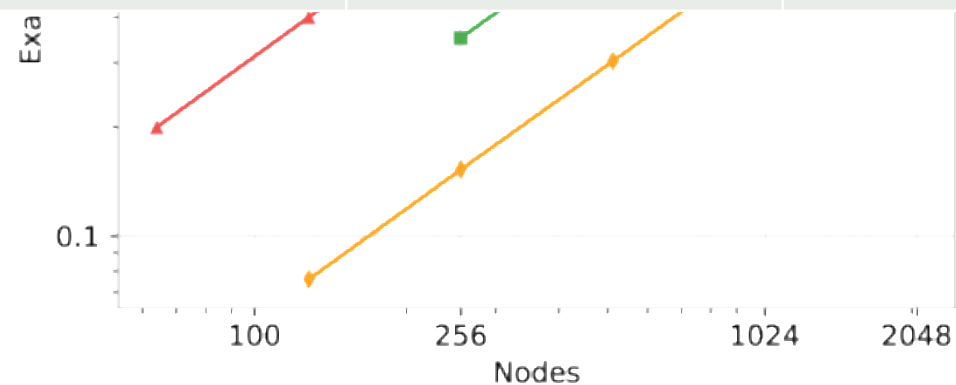
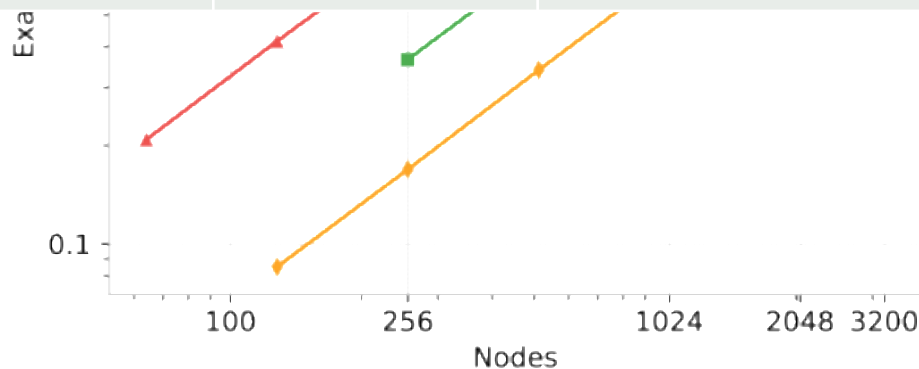
Scaling Mprot-DPO on supercomputers



... Achieves >1 Exaflops (MP) on each compute resource



System	Nodes	# of GPUs	Sustained EFLOPS (MP)	% Model Flop Utilization (MFU)	Peak EFLOPS (MP)
Aurora	3200	19200	4.11	44.5	5.57
Alps	2060	8240	2.92	41.7	3.16
Frontier	2048	8192	1.06	33.8	1.18
PDX	400	3200	1.29	48.4	1.39



Summary

- AI techniques in the loop can help accelerate therapeutic discovery processes
- We can design novel types of therapeutics with generative models:
 - peptides, antibodies,
 - proteins with non-standard amino-acids,
 - PNA (peptide nucleic acid)
- Self-driving labs is at the intersection of AI + high performance computing + real-life interactions
 - Transforming an entire generation of workforce
- HPC environments are critical resources for generative design:
 - a promptable engine for biologics design
 - inverse design is also a plausible strategy for targeting IDPs



Acknowledgements

Funding

- DOE BER AI Pilot projects
- CEPI
- *DOE ASCR MEDAL Project*
- *ARPA-H*

Computing Time

- Argonne Leadership Computing (Theta/Theta-GPU/ AI-testbed)
- INCITE
- Cerebras/Nvidia

Data/ Code/ Models

- <https://github.com/ramanathanlab/genSLM>
- Access to model weights will also be available via API

Colleagues

- Richard Scheuermann
- James Olds
- Wesley Scott
- Anda Trifan
- Ashka Shah
- Ozan Gokdemir
- Mike Tynes

Questions/Comments

ramanathana@anl.gov