Artificial Intelligence-Enabled Multi-Scale Simulations for COVID-19 Drug Discovery

Arvind Ramanathan

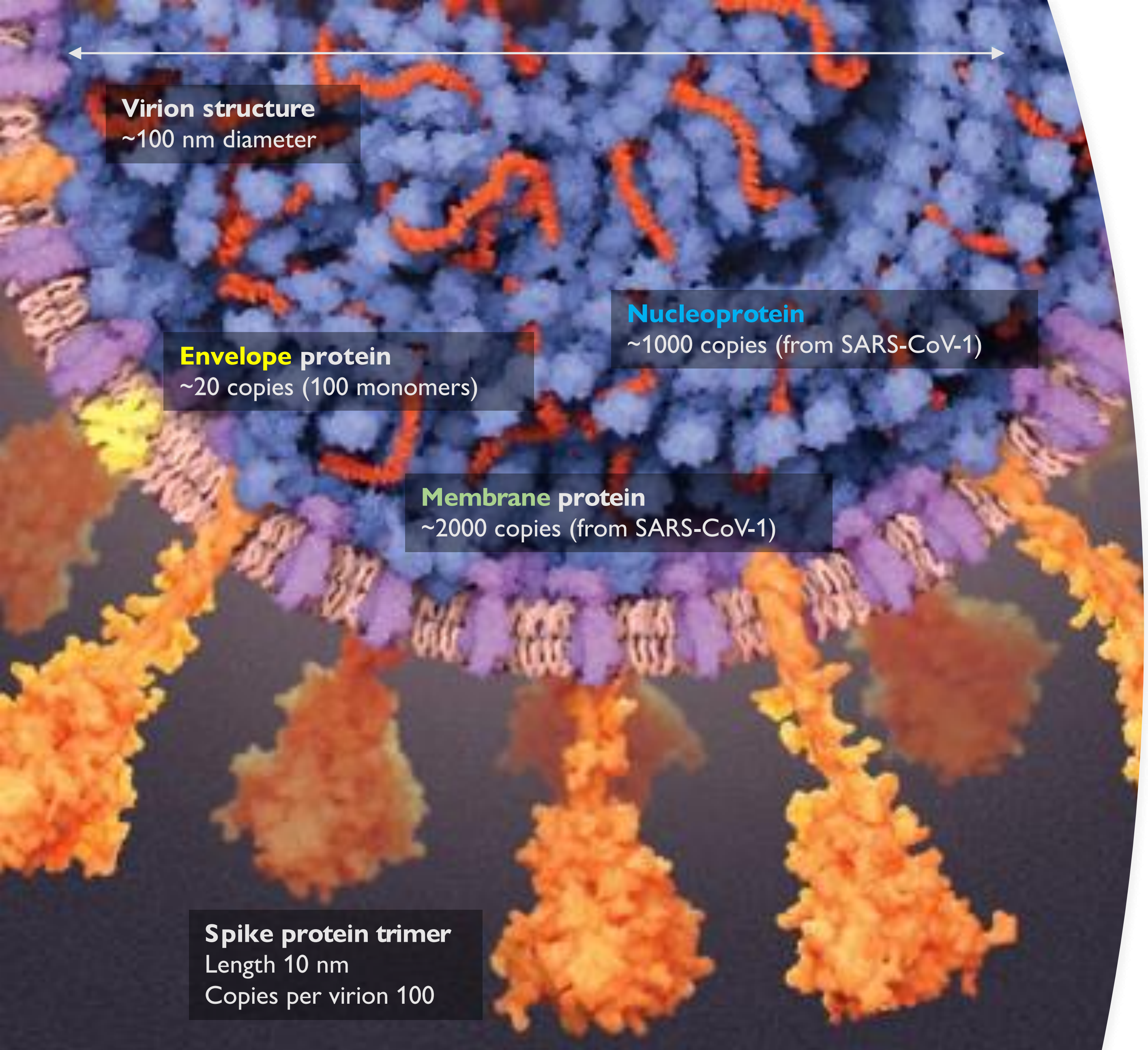
[1]Data Science and Learning Division, Argonne National Laboratory

[4]University of Chicago

ramanathana@anl.gov | https://ramanathanlab.org

Some take home messages…

- AI/ML approaches can interface with rigorous physics-based methods to address drug discovery challenges
- Emerging AI/ML approaches impose interesting “co-design” requirements for HPC
    - on existing supercomputing platforms
    - on emerging heterogeneous platforms
- Discovery of novel biological aspects related to SARS-CoV-2
    - small molecules that can bind to and inhibit SARS-CoV-2
    - insights into how SARS-CoV-2 binds to the ACE2 receptor



Veronica Falconieri Hays; Source: Lorenzo Casalino, Zied Gaieb and Rommie Amaro, U.C. San Diego (*spike model with glycosylations*)

https://www.scientificamerican.com/article/a-visual-guide-to-the-sars-cov-2-coronavirus/

Introduction to Covid-19 and SARS-COV-2

- Observed first in Wuhan (Dec 2019)
  - Quickly spread to the province of Hubei and then onto the world
- Spreads via close contact or through respiratory particles
- Virus is larger and far more stable than its counterparts (SARS and MERS)
  - can live on surfaces for a while
- Need a comprehensive strategy to identify small molecules (or other therapeutic strategies) to treat infection

published in eLife, March 31st, 2020
https://elifesciences.org/articles/57309

🟩 Structures solved at APS

🟦 Plausible Drug Targets

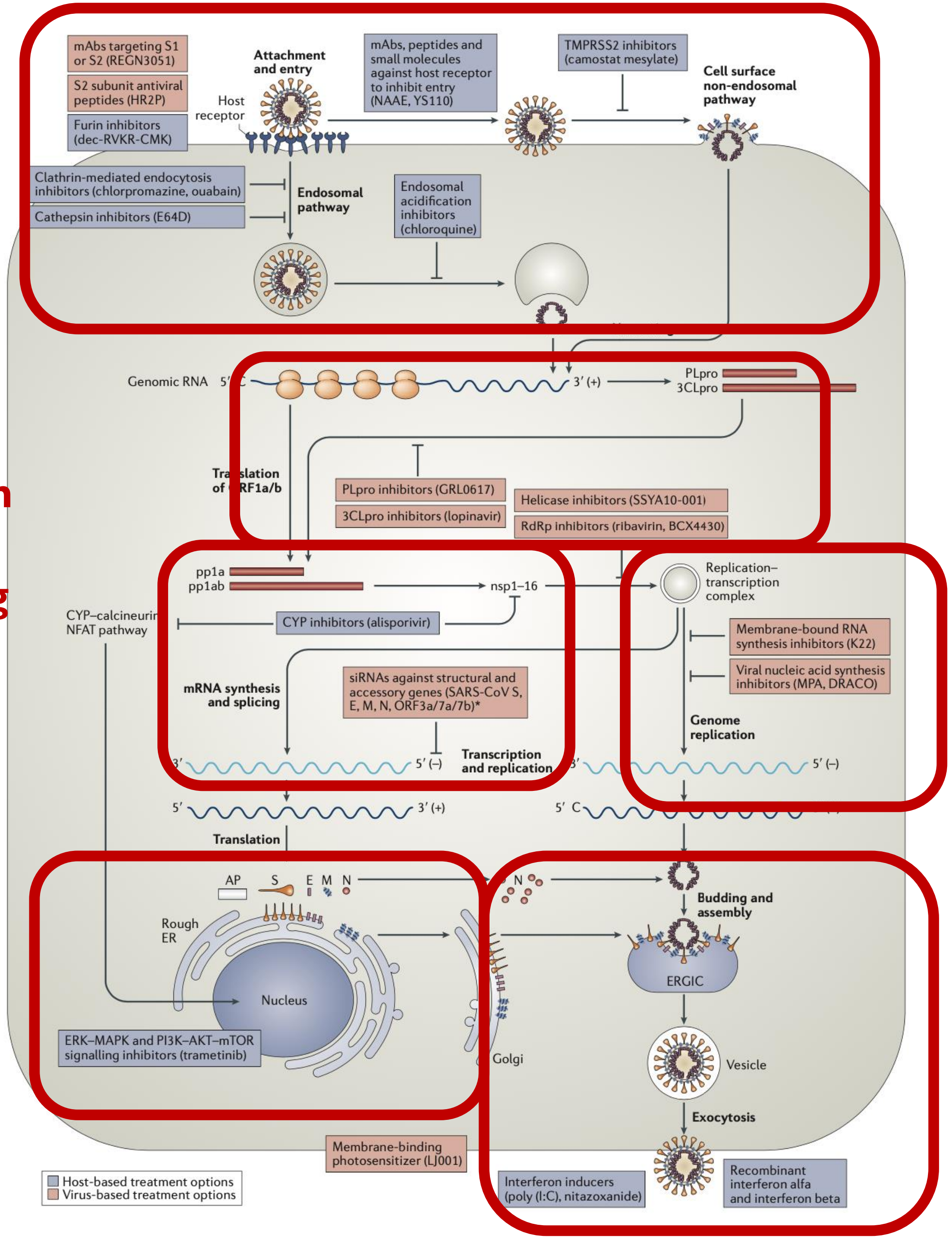
🟥 Priority Drug Targets

	Protein	Mol. weight (kDa)	Seq. similarity with SARS-CoV	Description	
	Nsp1	19.8	91.1%	Suppresses host antiviral response	🟦
	Nsp2	70.5	82.9%		
🟩	Nsp3	217.3	86.5%	Nsp3-Nsp4-Nsp6 complex involved in viral replication	🟥
	Nsp4	56.2	90.8%	Nsp3-Nsp4-Nsp6 complex involved in viral replication	
🟩	Nsp5	33.8	98.7%	Main protease (3C-like)	🟥
	Nsp6	33.0	94.8%	Nsp3-Nsp4-Nsp6 complex involved in viral replication	
🟩	Nsp7	9.2	100.0%	Nsp7-Nsp8 complex is part of RNA polymerase	🟥
🟩	Nsp8	21.9	99.0%	Nsp7-Nsp8 complex is part of RNA polymerase	🟦
	Nsp9	12.4	98.2%	ssRNA binding	
🟩	Nsp10	14.8	99.3%	Essential for Nsp16 methyltransferase activity	🟥
	Nsp11	1.3	92.3%	Short peptide	
	Nsp12	106.7	98.3%	RNA polymerase	🟦
	Nsp13	66.9	100.0%	Helicase/triphosphatase	
	Nsp14	59.8	98.7%	3′–5′ exonuclease	🟦
🟩	Nsp15	38.8	95.7%	Uridine-specific endoribonuclease	🟥
	Nsp16	33.3	98.0%	RNA-cap methyltransferase	🟥
	S	141.2	87.0%	Spike protein, mediates binding to ACE2	
	Orf3a	31.1	85.1%	Activates the NLRP3 inflammasome	🟦
	Orf3b	6.5	9.5%		
	E	8.4	96.1%	Envelope protein, involved in virus morphogenesis and assembly	🟦
	M	25.1	96.4%	Membrane glycoprotein, predominant component of the envelope	🟦
	Orf6	7.3	85.7%	Type I IFN antagonist	
🟩	Orf7a	13.7	90.2%	Virus-induced apoptosis	🟥
	Orf7b	5.2	84.1%		
	Orf8	13.8	45.3%		
🟩	N	45.6	94.3%	Nucleocapsid phosphoprotein, binds to RNA genome	🟦
	Orf9b	10.8	84.7%	Type I IFN antagonist	
	Orf9c	8.0	78.1%		
	Orf10	4.4	-		

Viral Entry

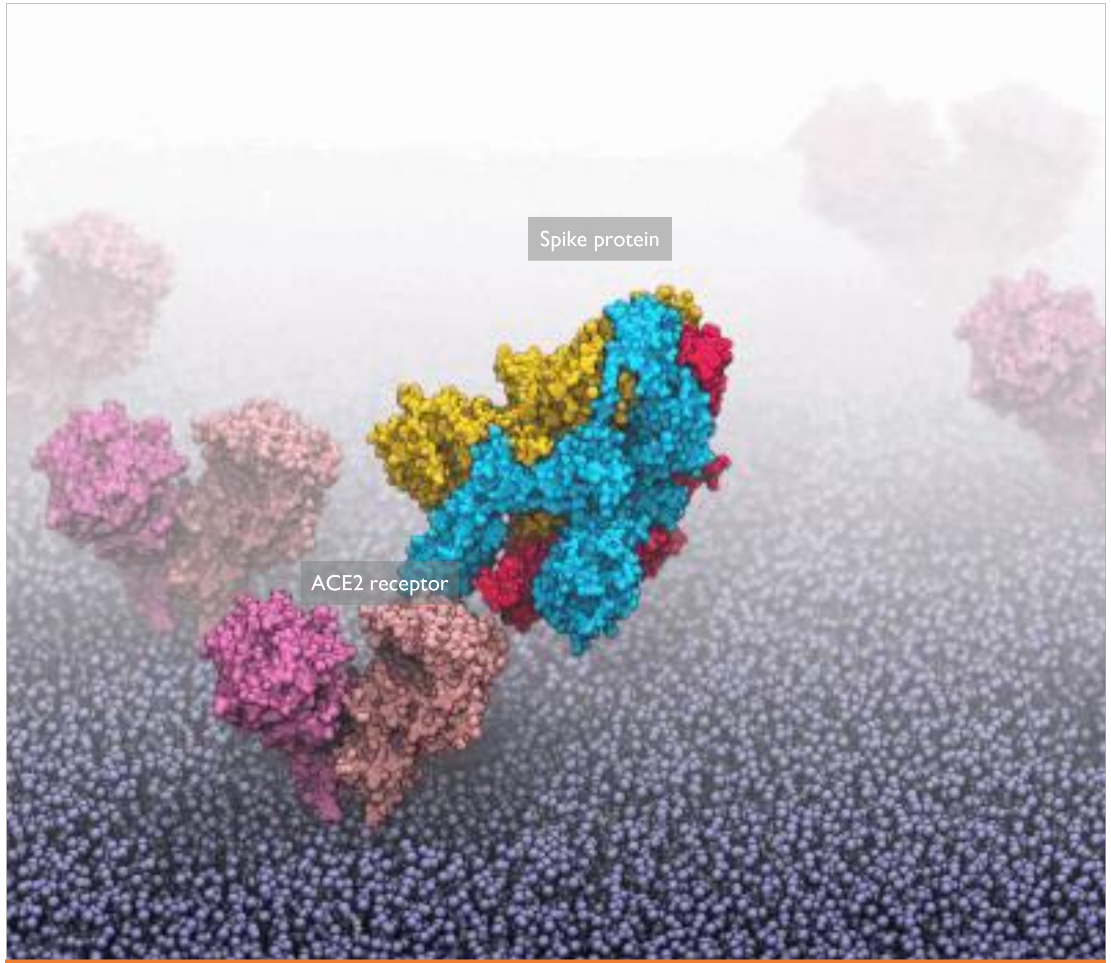
Viral Replication and RNA Processing

Host Signaling Processes and Viral Exit



mAbs targeting S1 or S2 (REGN3051)
S2 subunit antiviral peptides (HR2P)
Furin inhibitors (dec-RVKR-CMK)
Attachment and entry
Host receptor
mAbs, peptides and small molecules against host receptor to inhibit entry (NAAE, YS110)
TMPRSS2 inhibitors (camostat mesylate)
Cell surface non-endosomal pathway
Clathrin-mediated endocytosis inhibitors (chlorpromazine, ouabain)
Cathepsin inhibitors (E64D)
Endosomal pathway
Endosomal acidification inhibitors (chloroquine)
Genomic RNA
5′
3′ (+)
PLpro
3CLpro
Translation of ORF1a/b
PLpro inhibitors (GRL0617)
3CLpro inhibitors (lopinavir)
Helicase inhibitors (SSYA10-001)
RdRp inhibitors (ribavirin, BCX4430)
pp1a
pp1ab
nsp1–16
Replication–transcription complex
CYP–calcineurin NFAT pathway
CYP inhibitors (alisporivir)
Membrane-bound RNA synthesis inhibitors (K22)
Viral nucleic acid synthesis inhibitors (MPA, DRACO)
mRNA synthesis and splicing
siRNAs against structural and accessory genes (SARS-CoV S, E, M, N, ORF3a/7a/7b)*
Genome replication
3′
5′ (–)
Transcription and replication
5′
3′ (+)
Translation
AP
S
E
M
N
Rough ER
Nucleus
Golgi
Budding and assembly
ERGIC
Vesicle
Exocytosis
ERK–MAPK and PI3K–AKT–mTOR signalling inhibitors (trametinib)
Membrane-binding photosensitizer (LJ001)
Interferon inducers (poly (I:C), nitazoxanide)
Recombinant interferon alfa and interferon beta
Host-based treatment options
Virus-based treatment options

Outline (1)



Simulations driven by AI depict how the CoV-2 spike protein attaches to the human ACE2 receptor protein (Carlos Simmerling, Stony Brook)

How do we accelerate simulations of complex biological phenomena?

Collaboration between 10 institutions, 30 scientists across the globe!

Rommie Amaro

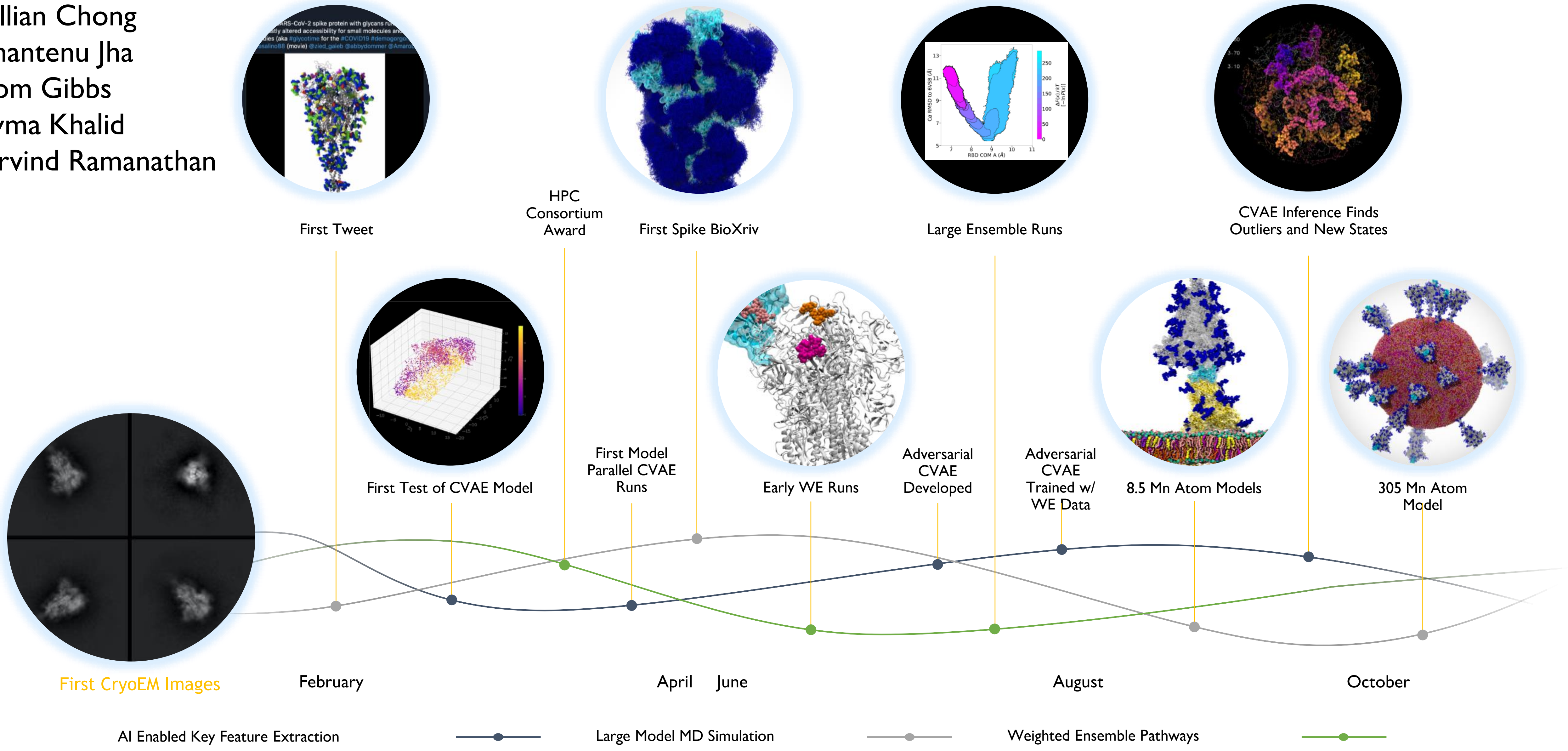
Lillian Chong

Shantenu Jha

Tom Gibbs

Syma Khalid

Arvind Ramanathan



ARS-CoV-2 spike protein with glycans ru
stly altered accessibility for small molecules and
dies (aka #glycotime for the #COVID19 #demogorgo
salino88 (movie) @zied_gaieb @abbydommer @Amaro
Cα RMSD to 6VSB (Å)
RBD COM A (Å)
ΔF(x) / kT [−ln P(x)]

First CryoEM Images

February

First Tweet

First Test of CVAE Model

HPC Consortium Award

First Model Parallel CVAE Runs

April

First Spike BioXriv

June

Early WE Runs

Adversarial CVAE Developed

Large Ensemble Runs

Adversarial CVAE Trained w/ WE Data

August

8.5 Mn Atom Models

CVAE Inference Finds Outliers and New States

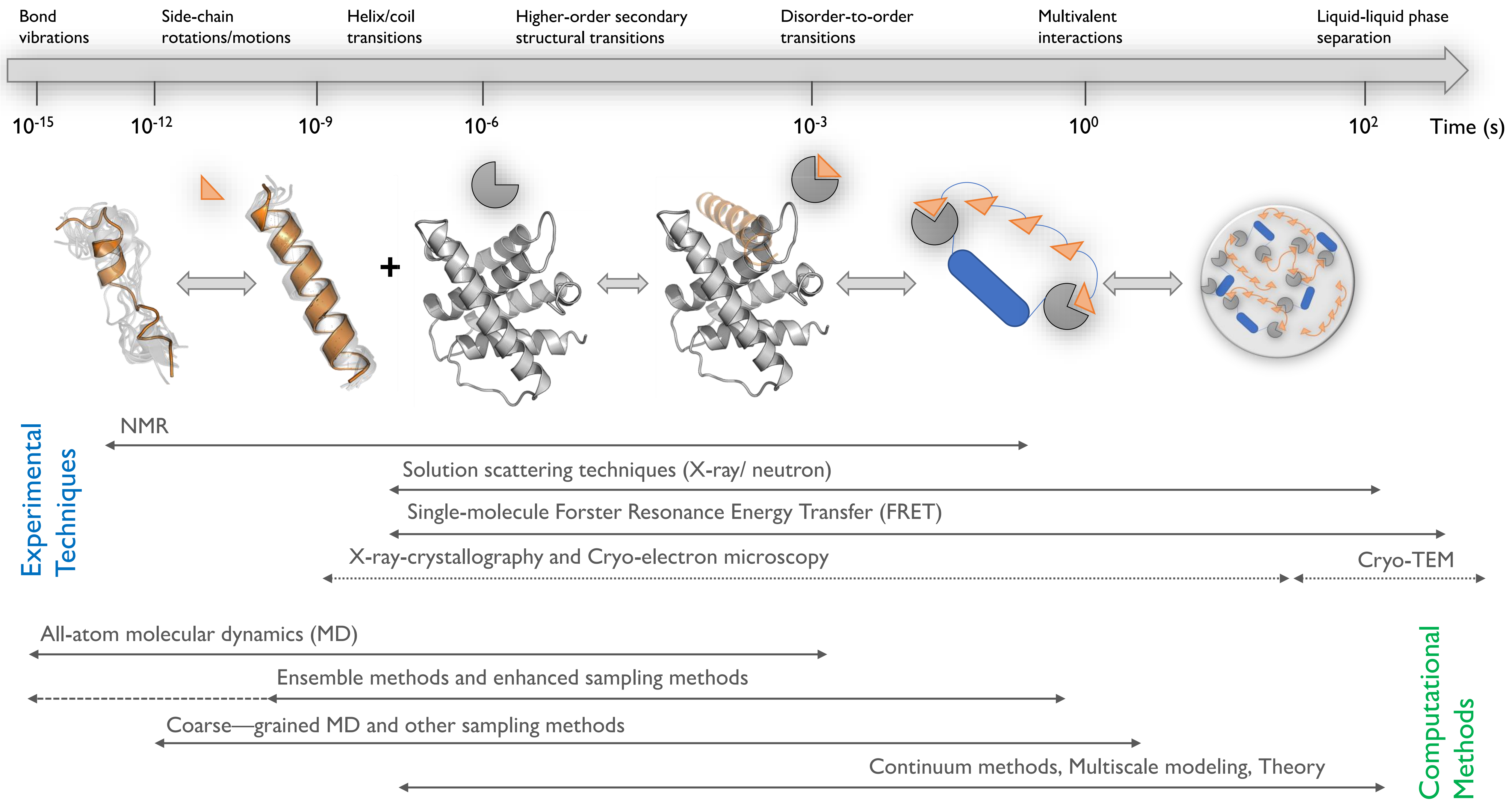
October

305 Mn Atom Model

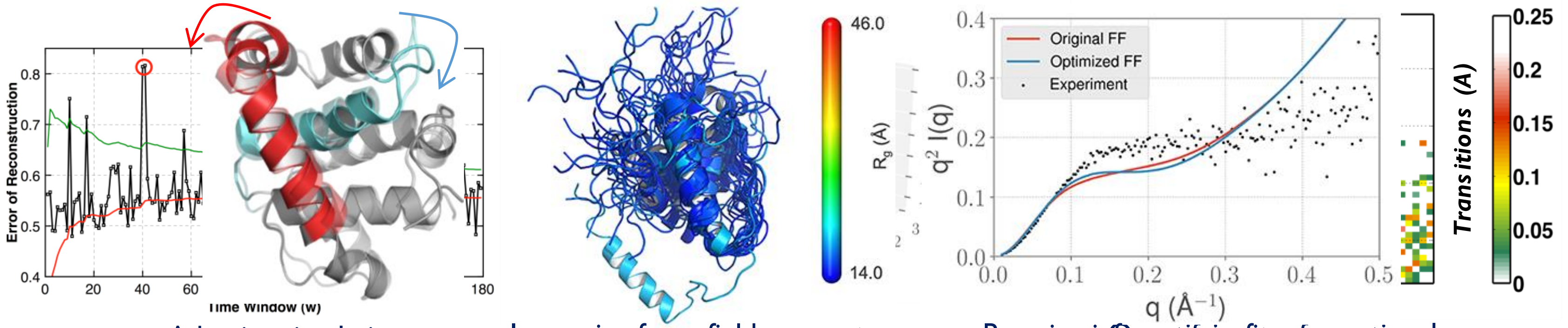
AI Enabled Key Feature Extraction

Large Model MD Simulation

Weighted Ensemble Pathways


Bond vibrations
Side-chain rotations/motions
Helix/coil transitions
Higher-order secondary structural transitions
Disorder-to-order transitions
Multivalent interactions
Liquid-liquid phase separation
10-15
10-12
10-9
10-6
10-3
100
102
Time (s)
+
Experimental Techniques
NMR
Solution scattering techniques (X-ray/ neutron)
Single-molecule Forster Resonance Energy Transfer (FRET)
X-ray-crystallography and Cryo-electron microscopy
Cryo-TEM
All-atom molecular dynamics (MD)
Ensemble methods and enhanced sampling methods
Coarse—grained MD and other sampling methods
Continuum methods, Multiscale modeling, Theory
Computational Methods

Statistical Inference: glue information across scales



Adaptive simulations + multi-scaling

Event detection

Improving force field parameters for simulations

Using manifold learning and clustering

Bayesian inference to fit sparse experimental observables

Quantifying conformational transitions

AI, probabilistic models, Bayesian inference

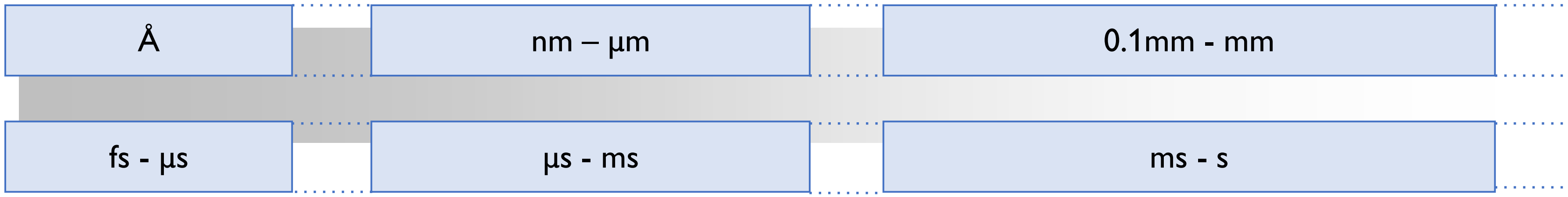
ML and deep learning approaches

Molecular & Macromolecular

Subcellular

Cellular

Spatial and Temporal Scales



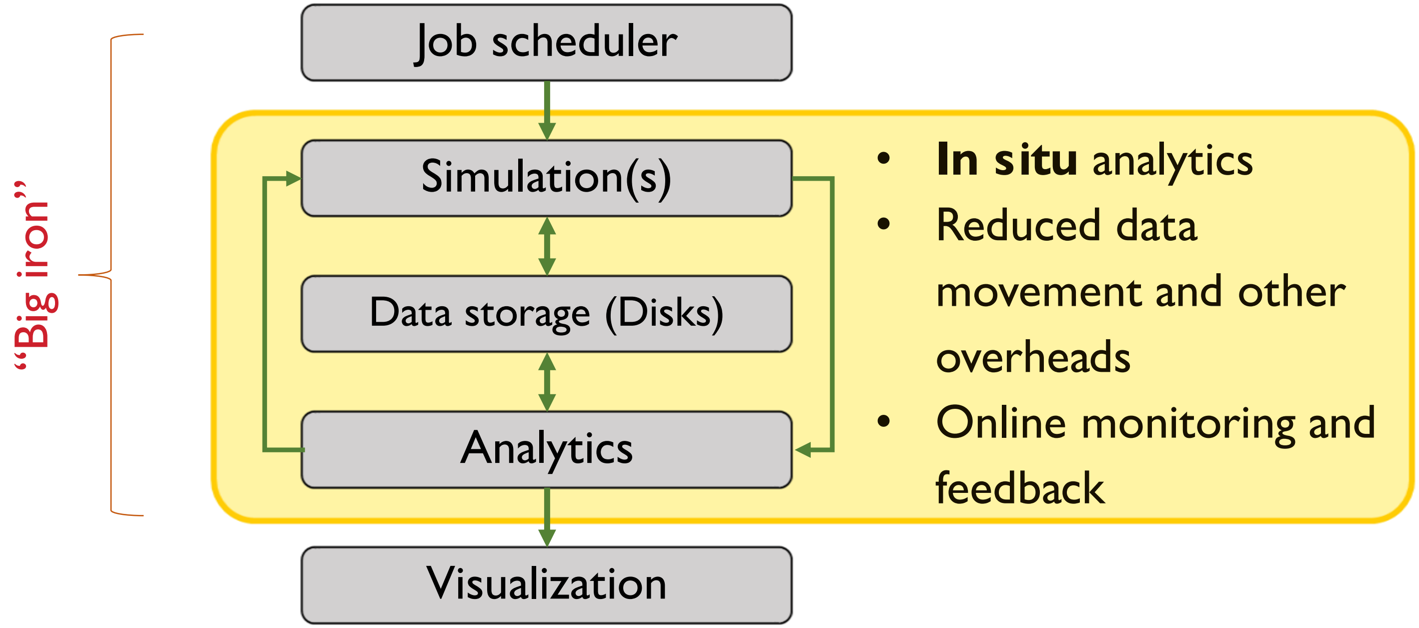
Standard simulations

Enhanced sampling workflows

AI/ML-driven workflows?

Standard simulation approaches face significant data movement and parallel analytics challenges

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

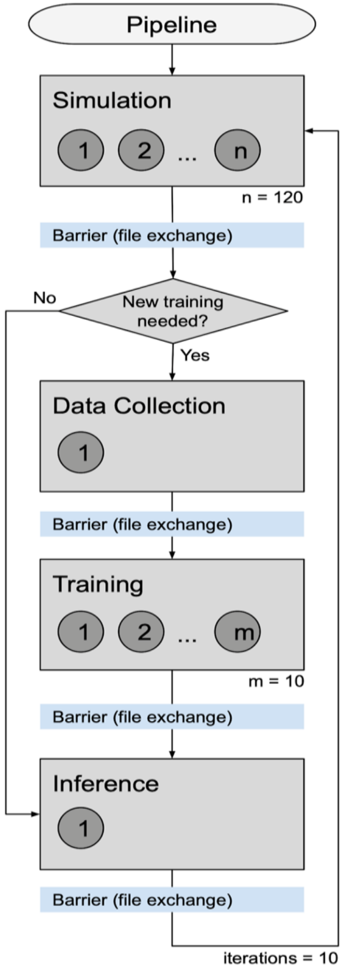


- **In situ** analytics
- Reduced data movement and other overheads
- Online monitoring and feedback

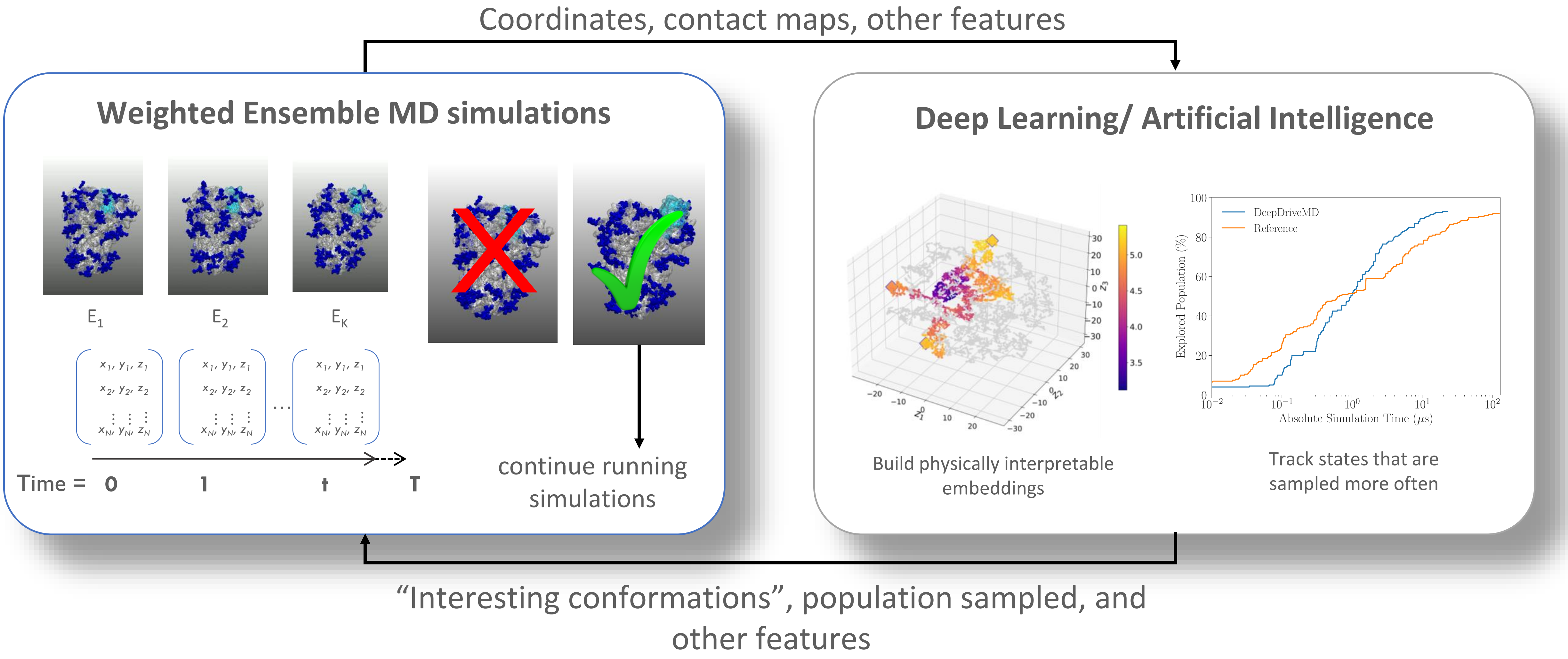
- Large simulations generate **> O(100 TB)** of data
- Humanly impossible to peek into "biologically" interesting events!
- http://deepdrivemd.github.io

- Ma, Lee, et al. PARCO (2019)
- Lee, Ma, et al. Workshop on Deep Learning on Supercomputers, Supercomputing (2019)

Ensemble Toolkit Workflow



Combining AI with HPC: AI-driven MD simulations -- DeepDriveMD



Coordinates, contact maps, other features

Weighted Ensemble MD simulations

continue running simulations

Deep Learning/ Artificial Intelligence

Build physically interpretable embeddings

Track states that are sampled more often

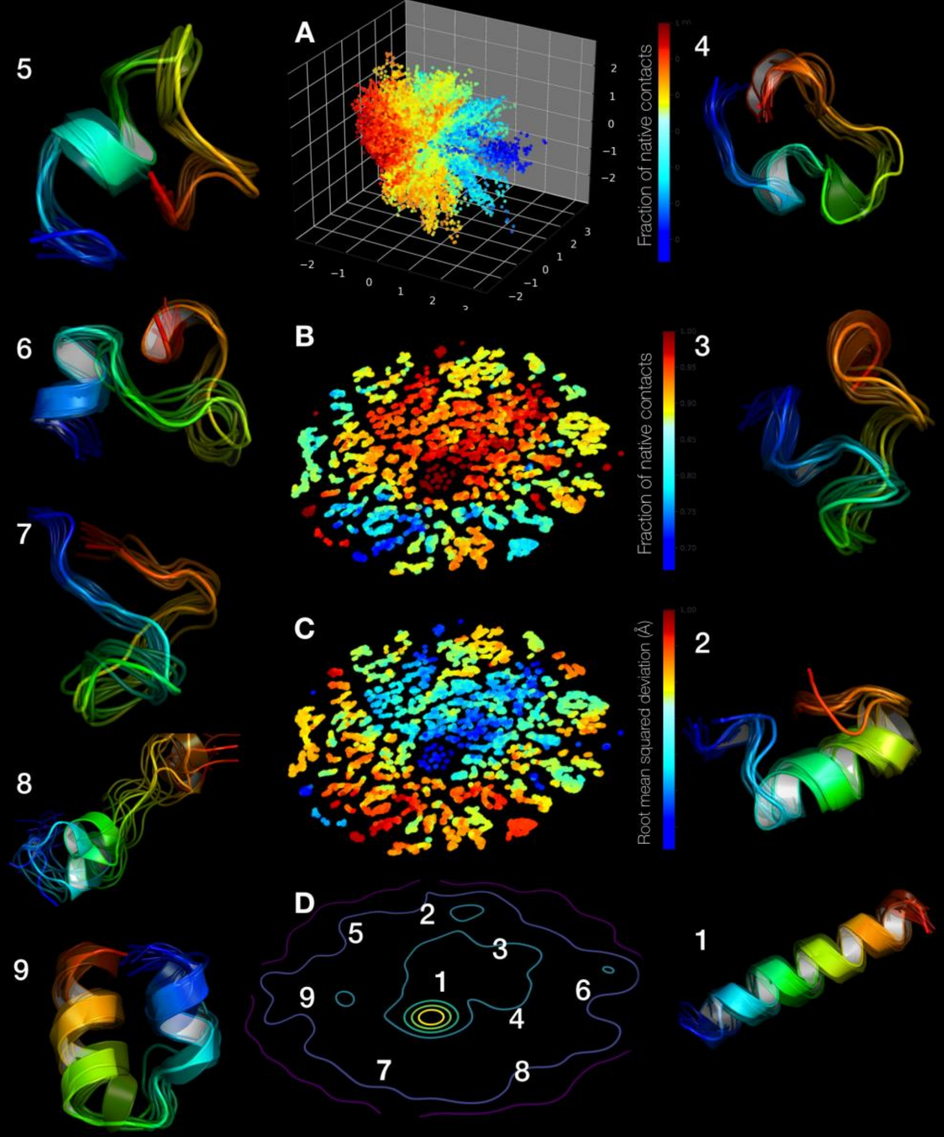
“Interesting conformations”, population sampled, and other features

Learning Everywhere

- Jha & Fox. In Visionary Track”, 15th International Conference eScience (2019), San Diego, California
- Jha & Fox. 15th International Conference eScience (2019), San Diego, California

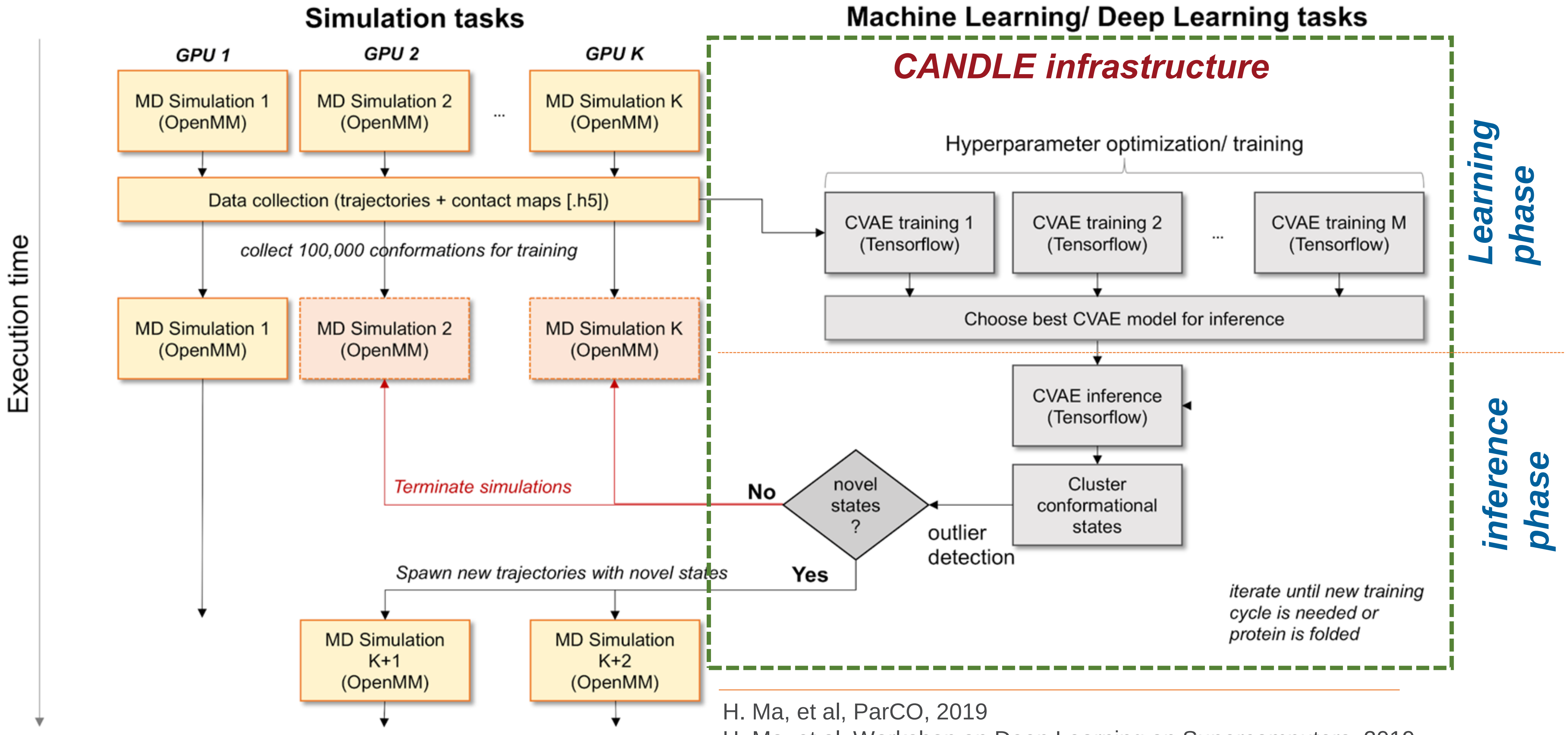
Deep clustering of protein folding simulations

- Convolutional Variational Auto Encoders (CVAE)
  - Low dimensional representations of states from simulation trajectories.
  - CVAE can transfer learned features to reveal novel states across simulations
- On folding trajectories:
  - identify intermediate states in an unsupervised manner
- Applied across multiple protein systems can provide a general way to extract ”reaction coordinates”



Bhowmik, D., et al, BMC Bioinformatics (2018).

DeepDriveMD: DL driven Adaptive Ensembles MD



Simulation tasks
Machine Learning/ Deep Learning tasks
CANDLE infrastructure
Execution time
GPU 1
GPU 2
GPU K
MD Simulation 1 (OpenMM)
MD Simulation 2 (OpenMM)
...
MD Simulation K (OpenMM)
Data collection (trajectories + contact maps [.h5])
collect 100,000 conformations for training
MD Simulation 1 (OpenMM)
MD Simulation 2 (OpenMM)
MD Simulation K (OpenMM)
Terminate simulations
No
Spawn new trajectories with novel states
Yes
MD Simulation K+1 (OpenMM)
MD Simulation K+2 (OpenMM)
Hyperparameter optimization/ training
CVAE training 1 (Tensorflow)
CVAE training 2 (Tensorflow)
...
CVAE training M (Tensorflow)
Choose best CVAE model for inference
Learning phase
CVAE inference (Tensorflow)
Cluster conformational states
novel states ?
outlier detection
inference phase
iterate until new training cycle is needed or protein is folded

H. Ma, et al, ParCO, 2019

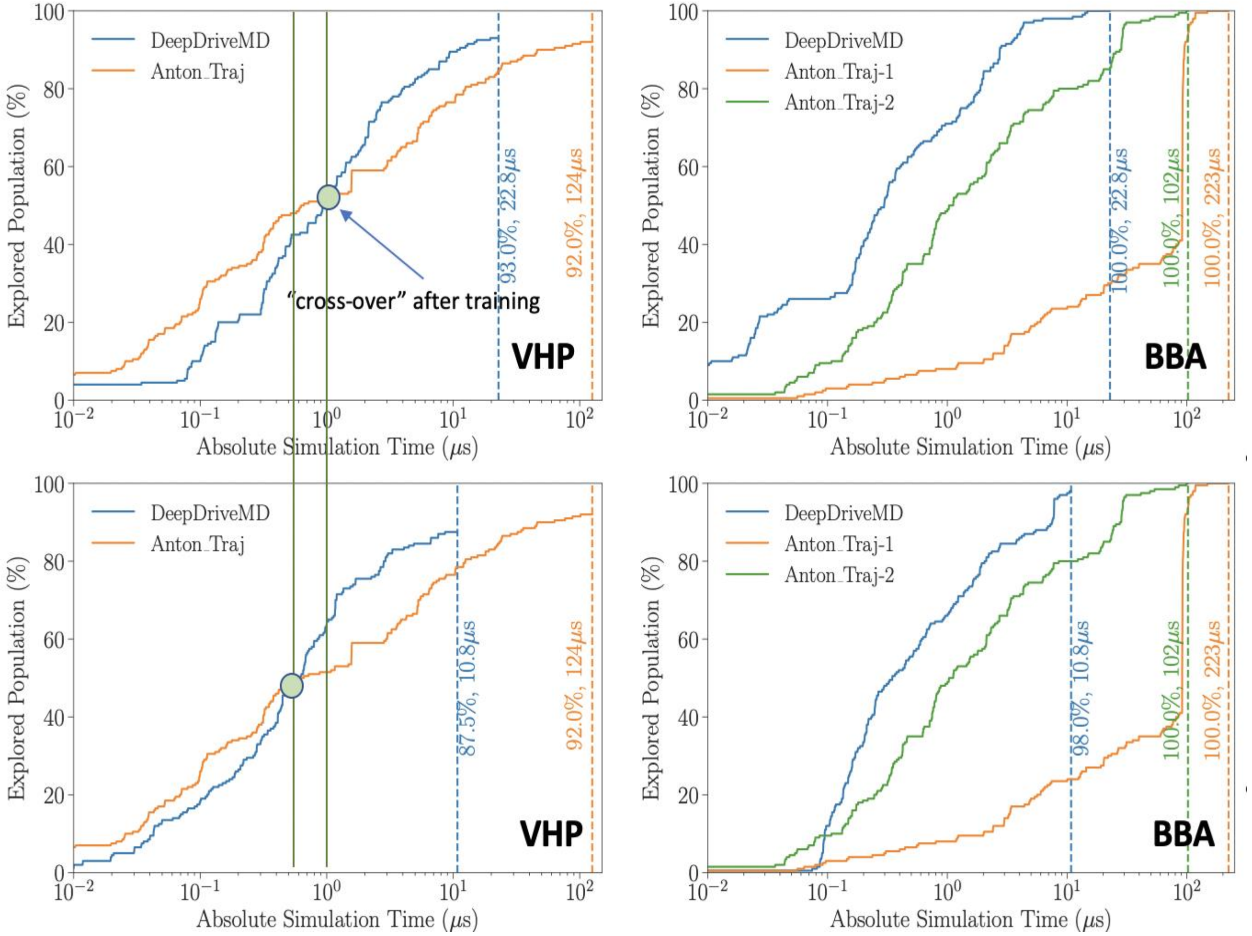
H. Ma, et al, Workshop on Deep Learning on Supercomputers, 2019

Collaboration with Shantenu Jha (Rutgers/ Brookhaven) and RADICAL team

DeepDriveMD is at least an order of magnitude better than traditional sampling

- Crossover point where DeepDriveMD based sampling is: (i) accelerated (ii) improves over "classical" methods
- O(100) greater sampling efficiency without considering time to train (for BBA protein)
    - If reference trajectories take O(microsecond) to sample a particular state, DeepDriveMD samples in O(100 ns)
    - For BBA, 98% sampled states are observed within 10 microseconds!
- Greater efficiency gains with larger proteins and complex dynamics
- Requires multiple and distinct levels of parallelism for "balanced" performance

DeepDriveMD: Deep-Learning Driven Adaptive Molecular Simulations for Protein Folding, **Workshop on Deep Learning on Supercomputers, SC'19** https://arxiv.org/abs/1909.07817



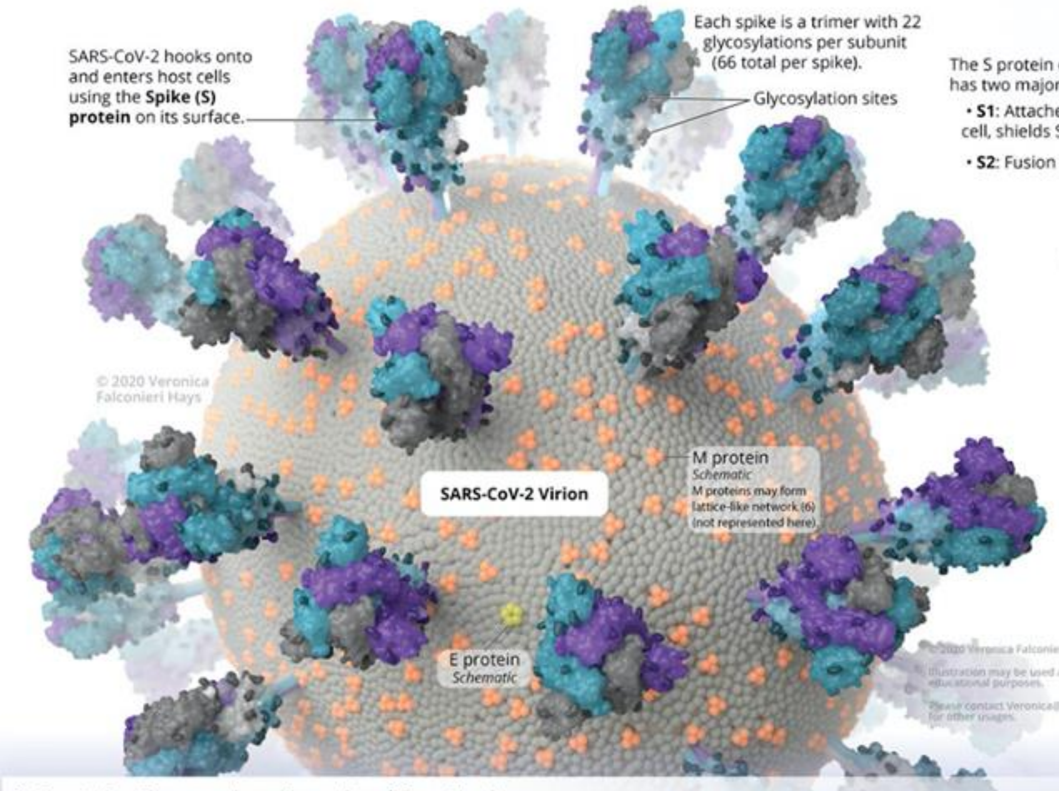
SARS-CoV-2 Spike Protein

Structure and pre-fusion processing

Because of its location and function, the spike is the target of neutralizing antibodies, and the focus of vaccine design.

See something inaccurate? Please contact Veronica@FalconieriVisuals.com to let us know. Thank you!

Last updated 4/2/2020



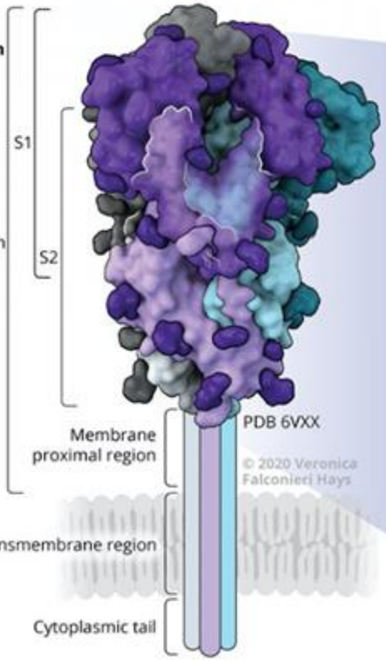
SARS-CoV-2 hooks onto and enters host cells using the **Spike (S) protein** on its surface.

Each spike is a trimer with 22 glycosylations per subunit (66 total per spike).

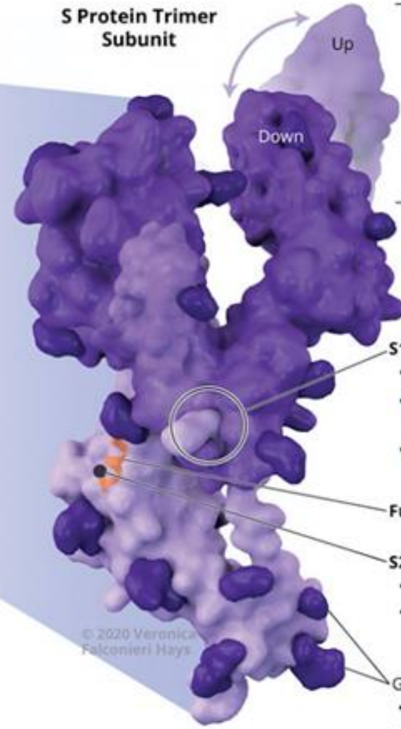
S Protein Trimer

The S protein **ectodomain** has two major regions:

- **S1**: Attaches to host cell, shields S2
- **S2**: Fusion machinery



S Protein Trimer Subunit



RBD

- Receptor binding domain. Part of S1.
- SARS-CoV-2's RBD binds host receptor (ACE2) with 10-20x the affinity of SARS-CoV's RBD (1)
- Flexible. Some spikes have one subunit with RBD "up" (PDB 6VSB), even without ACE2 present (1, 2)

S1/S2 "Priming" cleavage site

- Between the S1 and S2 domains
- SARS-CoV-2: Furin recognition site (Polybasic: RRAR)
- SARS-CoV: No furin recognition site (Monobasic: R)

Fusion peptide

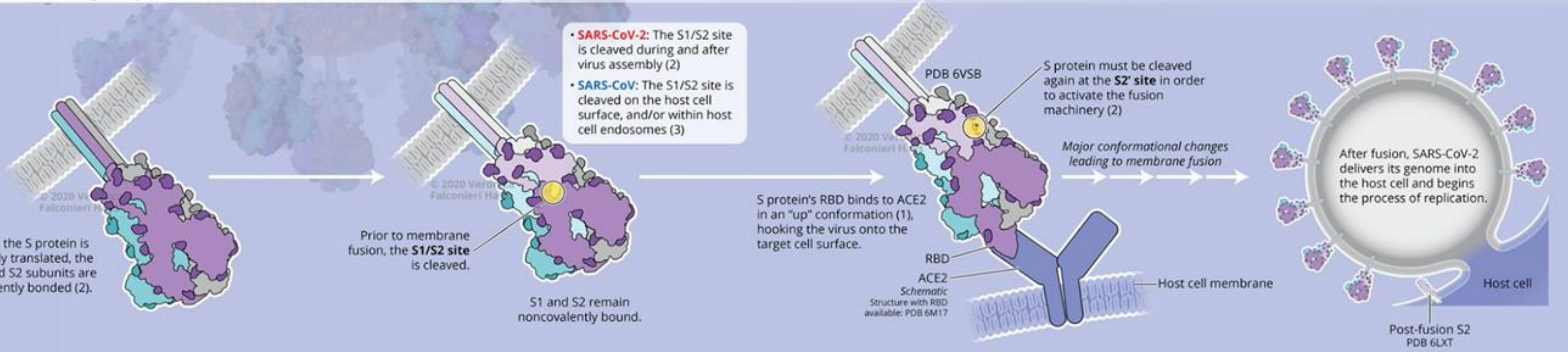
S2' "Activating" cleavage site

- Immediately next to the fusion peptide.
- Transmembrane protease (TMPRSS) and/or Cathepsin L recognition site

Glycosylation sites

- 16 / 22 resolved in PDB 6VXX
- 15 / 22 in PDB 6VSB

S Protein Processing Required for Fusion



When the S protein is initially translated, the S1 and S2 subunits are covalently bonded (2).

Prior to membrane fusion, the **S1/S2 site** is cleaved.

S1 and S2 remain noncovalently bound.

- SARS-CoV-2: The S1/S2 site is cleaved during and after virus assembly (2)
- SARS-CoV: The S1/S2 site is cleaved on the host cell surface, and/or within host cell endosomes (3)

S protein's RBD binds to ACE2 in an "up" conformation (1), hooking the virus onto the target cell surface.

S protein must be cleaved again at the **S2' site** in order to activate the fusion machinery (2)

Major conformational changes leading to membrane fusion

After fusion, SARS-CoV-2 delivers its genome into the host cell and begins the process of replication.

Abbreviations

SARS-CoV-2: Severe acute respiratory syndrome coronavirus-2. The virus that causes the disease COVID-19. Previously known as 2019-nCoV. Often referred to as "the coronavirus", "novel coronavirus".

SARS-CoV: The virus that caused the outbreak of SARS between 2002 - 2003. Closely related to SARS-CoV-2.

RBD: Receptor binding domain. Also known as Domain B (S^B).

ACE2: Angiotensin-converting enzyme 2 (ACE2). Target receptor on host cells required for entry (2).

References

Spike Structure and Function

1. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science (80-). 2020 Mar 13;367(6483):1260–3.

2. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell [Internet]. 2020 Mar 6 [cited 2020 Mar 28]; Available from: http://www.ncbi.nlm.nih.gov/pubmed/32155444

3. Glowacka I, Bertram S, Muller MA, Allen P, Soilleux E, Pfefferle S, et al. Evidence that TMPRSS2 Activates the Severe Acute Respiratory Syndrome Coronavirus Spike Protein for Membrane Fusion and Reduces Viral Control by the Humoral Immune Response. J Virol. 2011 May 1;85(9):4122–34. (SARS-CoV S1/S2 processing location)

SARS-CoV-2 Ultrastructure

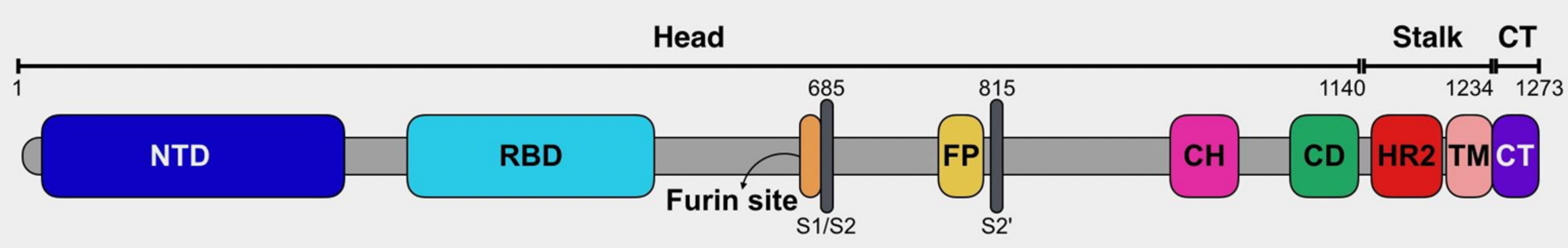
4. Novel Coronavirus SARS-CoV-2: Transmission electron micrograph of SARS-CoV-2 virus particles, isolated from a patient. NIAID. Available from: https://www.flickr.com/photos/niaid/49645120251/in/album-72157712914621487/

5. Coronavirus Illustration by David S. Goodsell, RCSB Protein Data Bank; doi: 10.2210/rcsb_pdb/goodsell-gallery-019. Available from: http://pdb101.rcsb.org/sci-art/goodsell-gallery/coronavirus

6. Neuman BW, Adair BD, Yoshioka C, Quispe JD, Orca G, Kuhn P, et al. Supramolecular Architecture of Severe Acute Respiratory Syndrome Coronavirus Revealed by Electron Cryomicroscopy. J Virol. 2006 Aug 15;80(16):7918–28. (SARS-CoV data: Size, S, M, and E protein stoichiometry)

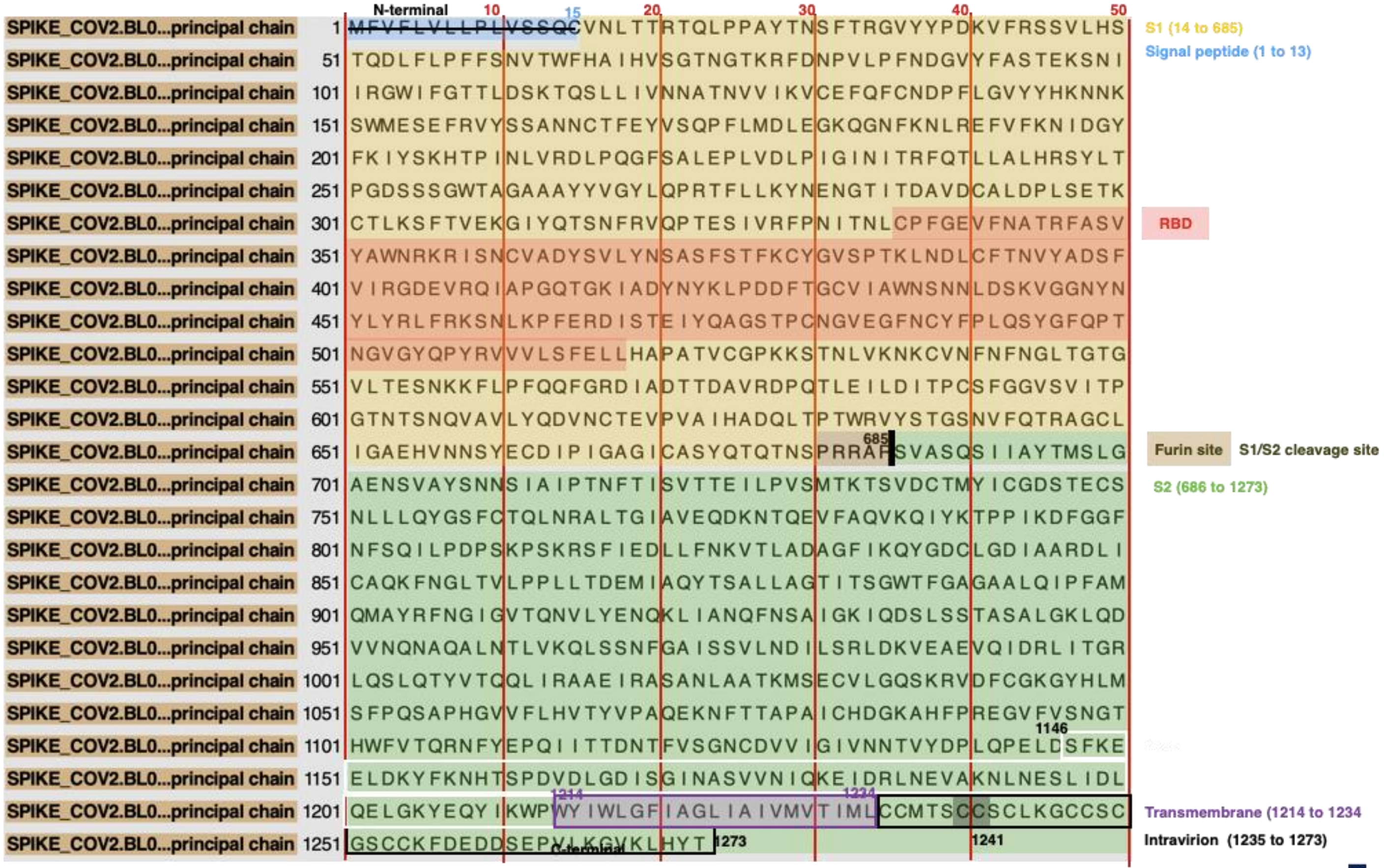
7. Siu YL, Teoh KT, Lo J, Chan CM, Kien F, Escriou N, et al. The M, E, and N Structural Proteins of the Severe Acute Respiratory Syndrome Coronavirus Are Required for Efficient Assembly, Trafficking, and Release of Virus-Like Particles. J Virol. 2008 Nov 15;82(22):11318–30. (M protein SARS CoV)

8. Torres J, Parthasarathy K, Lin X, Saravanan R, Kukol A, Ding XL. Model of a putative pore: The pentameric α-helical bundle of SARS coronavirus E protein in lipid bilayers. Biophys J. 2006;91(3):938–47. (E protein SARS CoV)

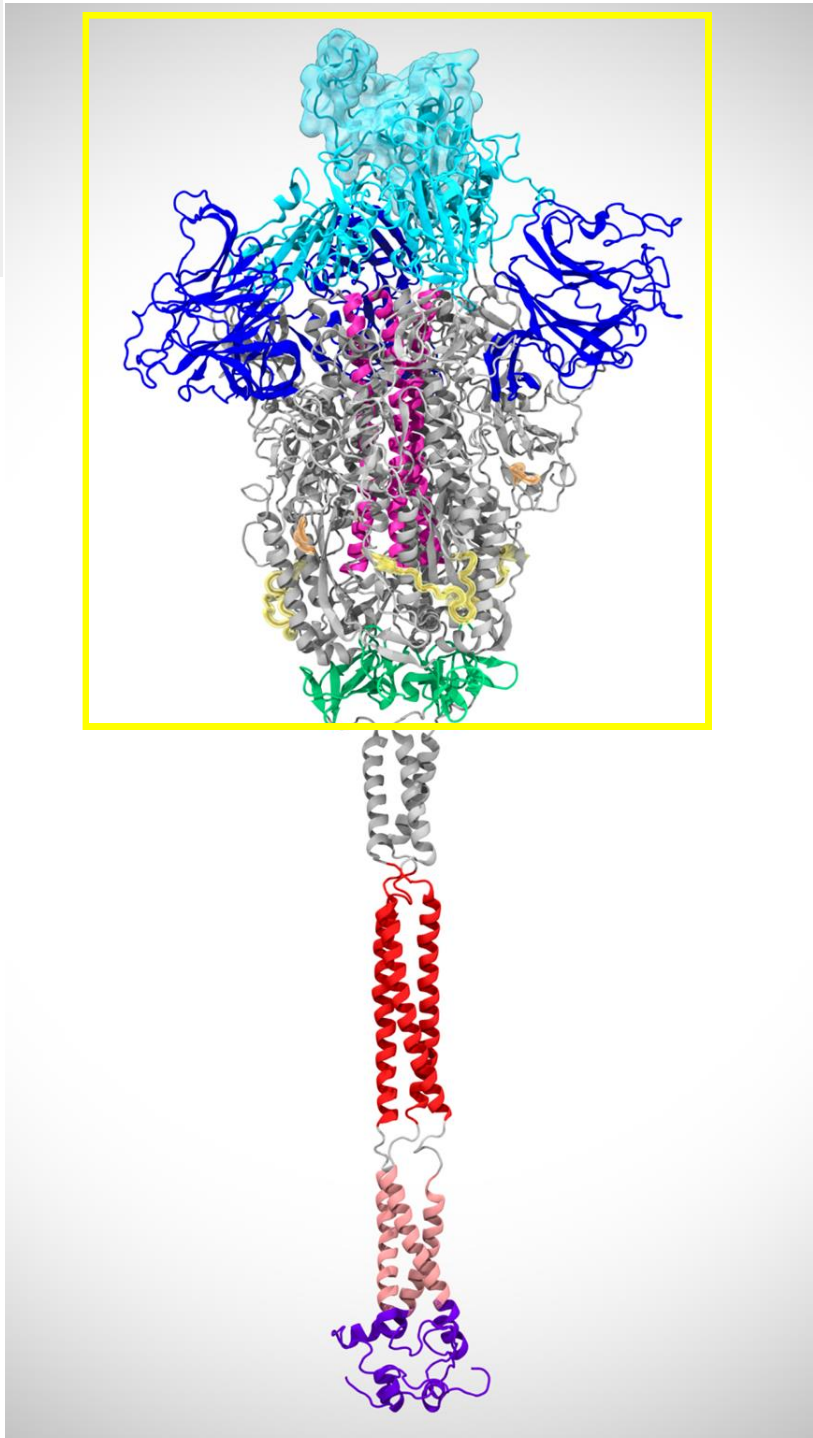


SARS-CoV-2 Spike sequence

- **Sequence of 6vsb (with gaps):**



SPIKE_COV2.BL0...principal chain	1	~~MFVFLVLLPLVSSQ~~CVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHS	S1 (14 to 685) Signal peptide (1 to 13)
SPIKE_COV2.BL0...principal chain	51	TQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNI	
SPIKE_COV2.BL0...principal chain	101	IRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNK	
SPIKE_COV2.BL0...principal chain	151	SWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGY	
SPIKE_COV2.BL0...principal chain	201	FKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLT	
SPIKE_COV2.BL0...principal chain	251	PGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETK	
SPIKE_COV2.BL0...principal chain	301	CTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASV	RBD
SPIKE_COV2.BL0...principal chain	351	YAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF	
SPIKE_COV2.BL0...principal chain	401	VIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN	
SPIKE_COV2.BL0...principal chain	451	YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPT	
SPIKE_COV2.BL0...principal chain	501	NGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTG	
SPIKE_COV2.BL0...principal chain	551	VLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP	
SPIKE_COV2.BL0...principal chain	601	GTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCL	
SPIKE_COV2.BL0...principal chain	651	IGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLG	Furin site S1/S2 cleavage site
SPIKE_COV2.BL0...principal chain	701	AENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECS	S2 (686 to 1273)
SPIKE_COV2.BL0...principal chain	751	NLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGF	
SPIKE_COV2.BL0...principal chain	801	NFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLI	
SPIKE_COV2.BL0...principal chain	851	CAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAM	
SPIKE_COV2.BL0...principal chain	901	QMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQD	
SPIKE_COV2.BL0...principal chain	951	VVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGR	
SPIKE_COV2.BL0...principal chain	1001	LQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLM	
SPIKE_COV2.BL0...principal chain	1051	SFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGT	
SPIKE_COV2.BL0...principal chain	1101	HWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKE	
SPIKE_COV2.BL0...principal chain	1151	ELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDL	
SPIKE_COV2.BL0...principal chain	1201	QELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSC	Transmembrane (1214 to 1234
SPIKE_COV2.BL0...principal chain	1251	GSCCKFDEDDSEPVLKGVKLHYT	Intravirion (1235 to 1273)



Weighted Ensemble (WE) method

- From Huber and Kim *Biophys. Journal* (1996)
- Instead of running one long simulation, runs many short simulations (“walkers”) with probabilities
- Samples the free energy landscape defined by chosen progress coordinates □ landscape is divided into “bins” and user chooses which trajectories to continue based on how they are progressing
- Trick is that you miss out on the ‘waiting times’ or the dwell times in energy wells for molecular events

- Why use the WE method?
  1. No statistical bias is added to the system
  2. Can sample **both** thermodynamic and kinetic properties
  3. Continuous, unbiased pathways can be obtained
  4. Monitoring evolution and convergence of properties is possible
  5. Adjusting bins and other parameters “on-the-fly” is possible

Collaboration with Terra Stzain, Shirley Ahn, Antony Bogetti, Lillian Chong

Weighted Ensemble Simulations

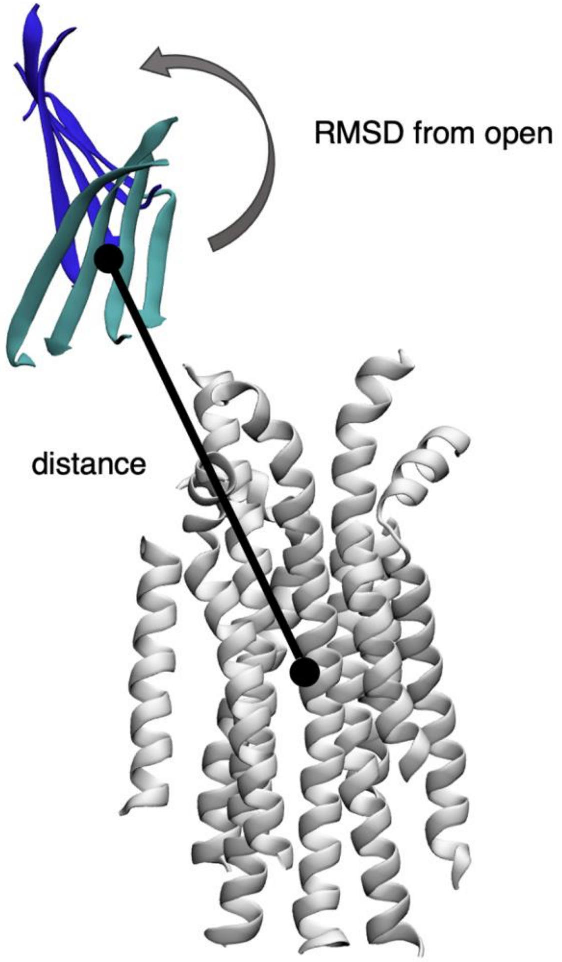
- ~ 600,000 atoms, the largest system **by an order of magnitude** that has been simulated using the WE method
- Initial state: 6VXX (closed)
- Weighted Ensemble Simulation Toolkit with Parallelization and Analysis (WESTPA)
- Initial runs on SDSC Comet, NVIDIA P100 GPUs
- Longhorn system at the Texas Advanced Computing Center (TACC)
- AMBER 18 MD engine, GPU optimized pmemd.cuda on 100 NVIDIA V100 GPUs

Aggregate sampling: ~200 microseconds actual simulation time

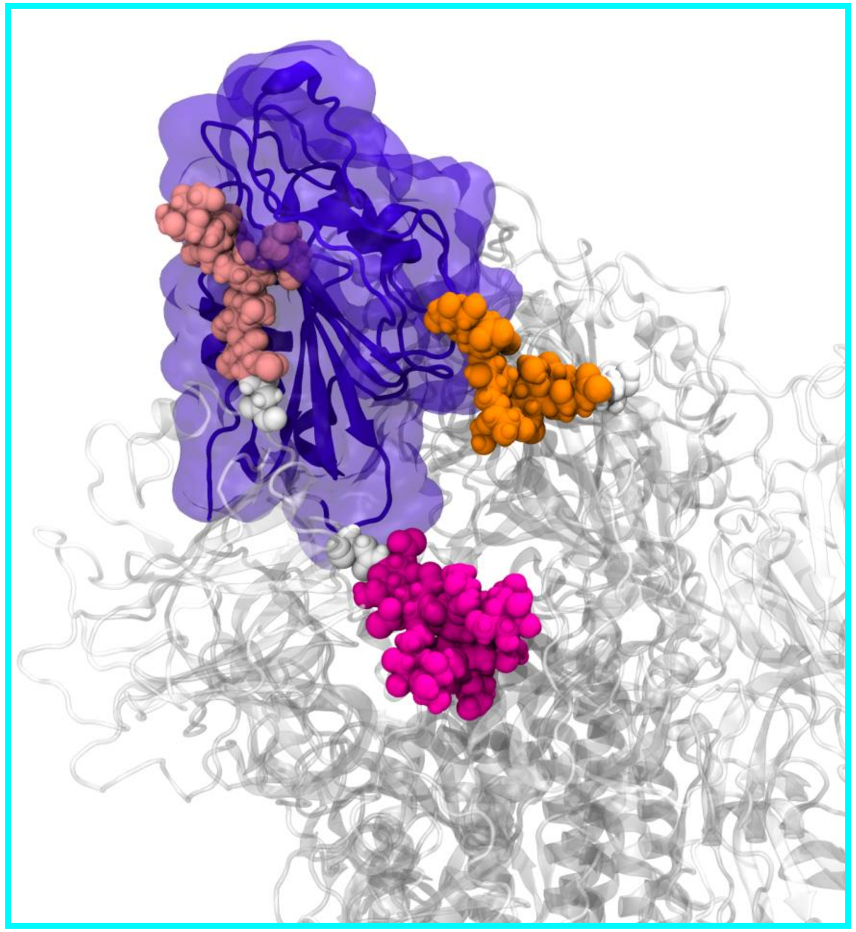
Equivalent of ~ 100s of milliseconds of timescale sampling with WE

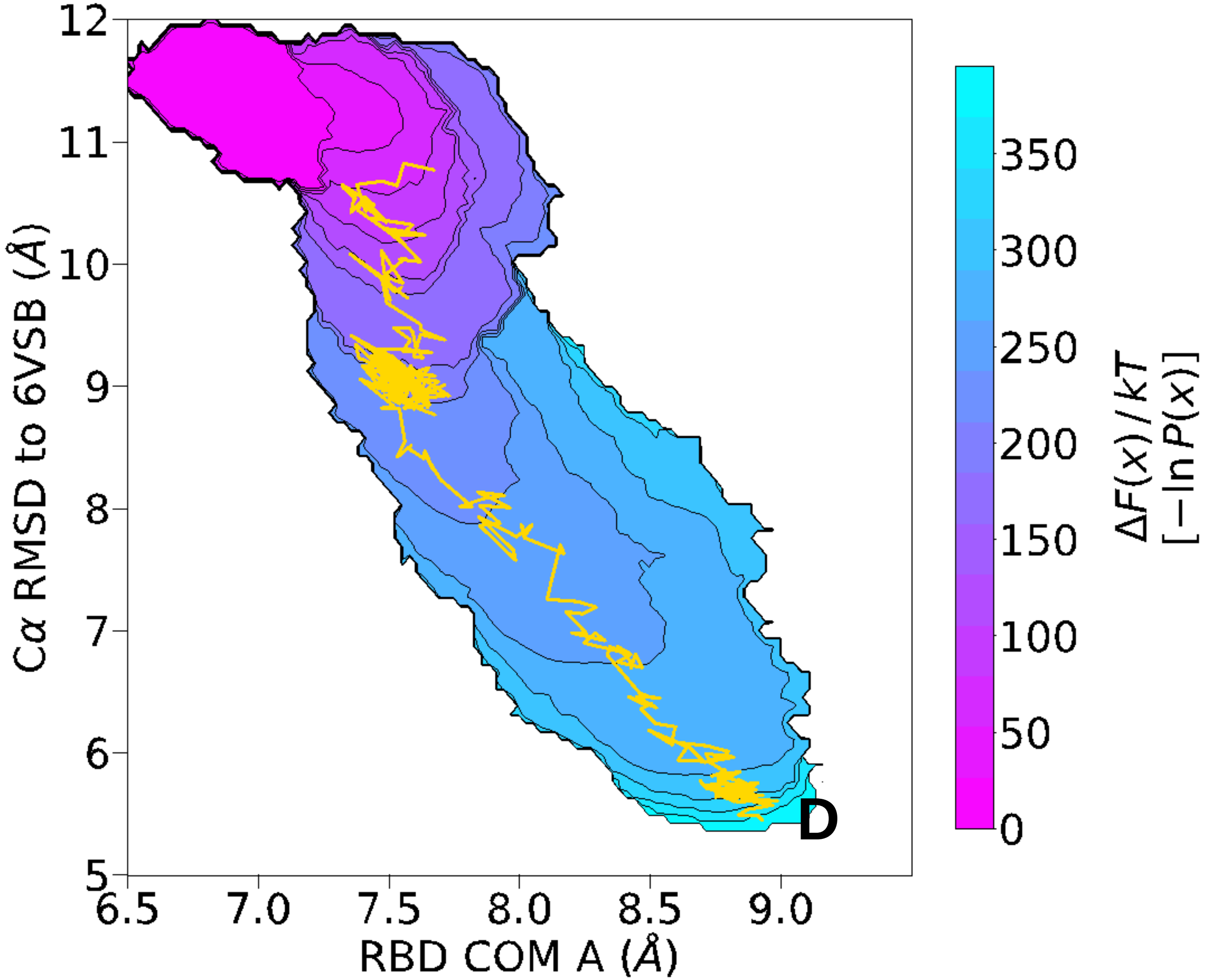
~ 100TB of data with compression, w/o solvent (protein only)

Collaboration with Terra Stzain, Shirley Ahn, Antony Bogetti, Lillian Chong

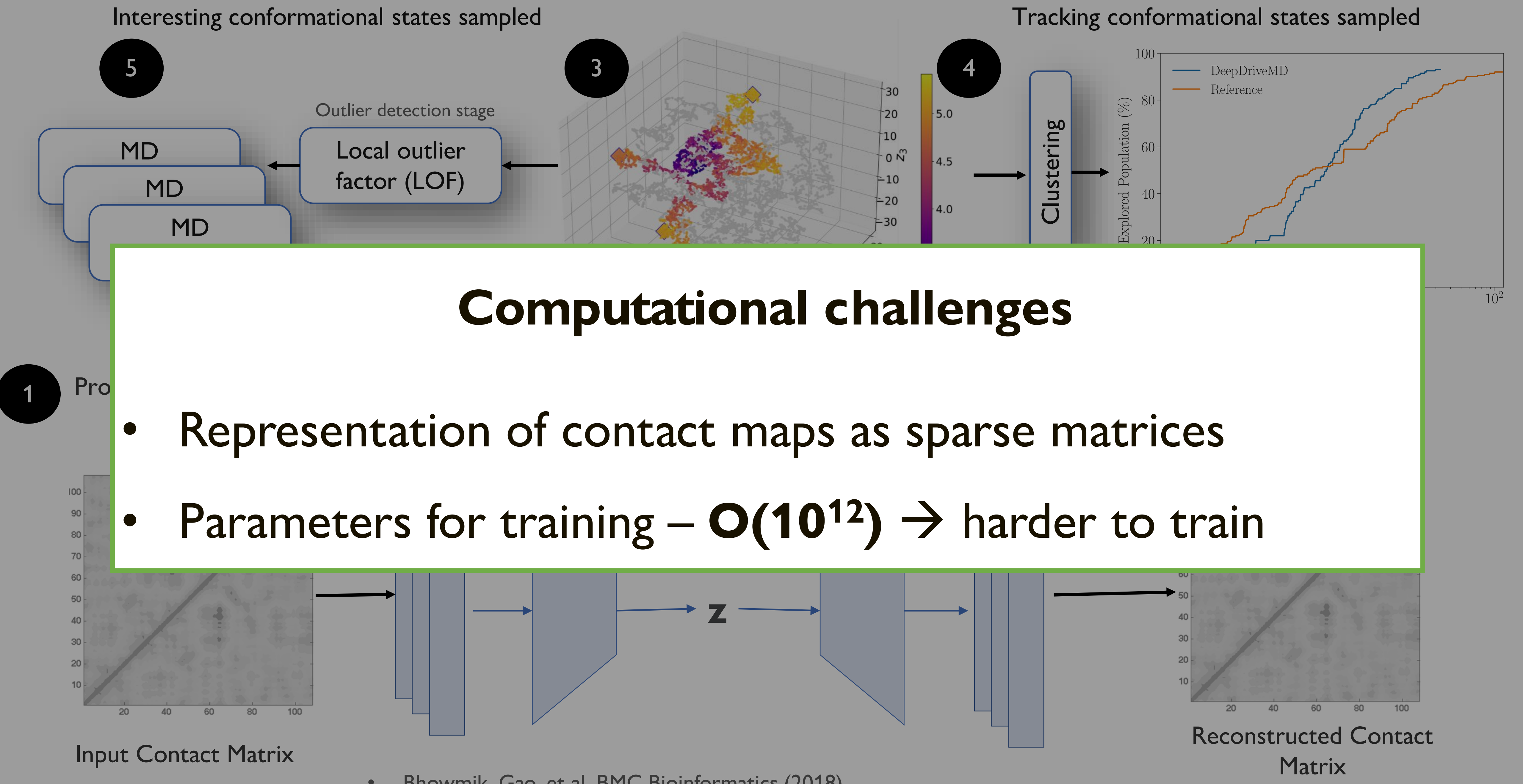


Continuous unbiased spike opening





Collaboration with Terra Stzain, Shirley Ahn, Antony Bogetti, Lillian Chong

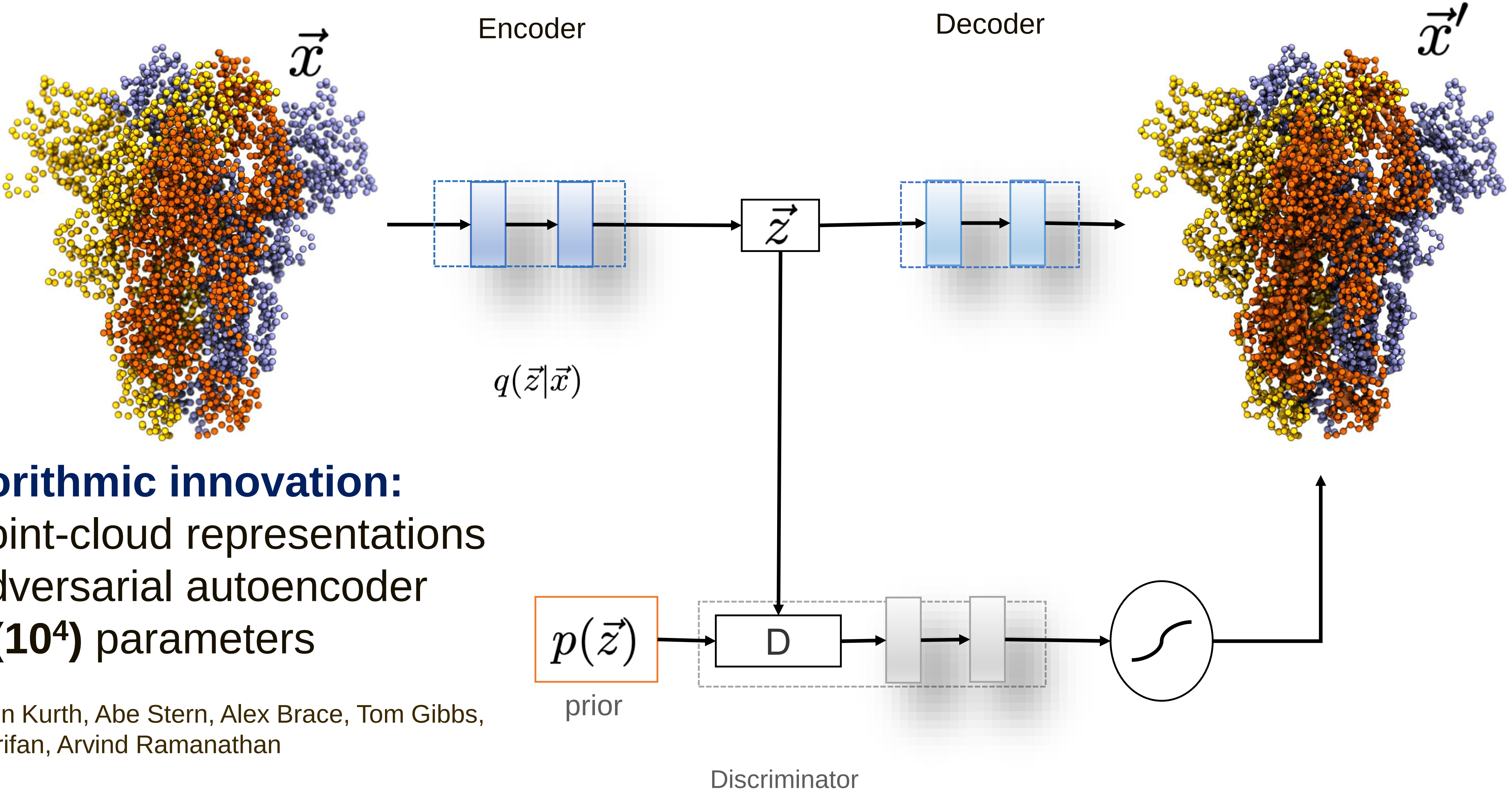


Computational challenges

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

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

Adversarial autoencoders for efficient analysis

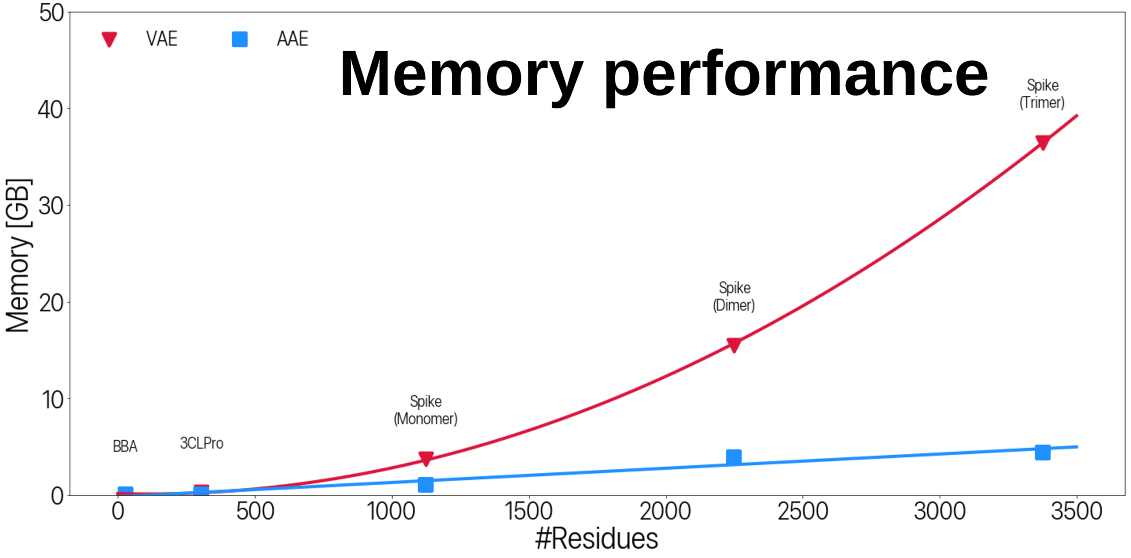


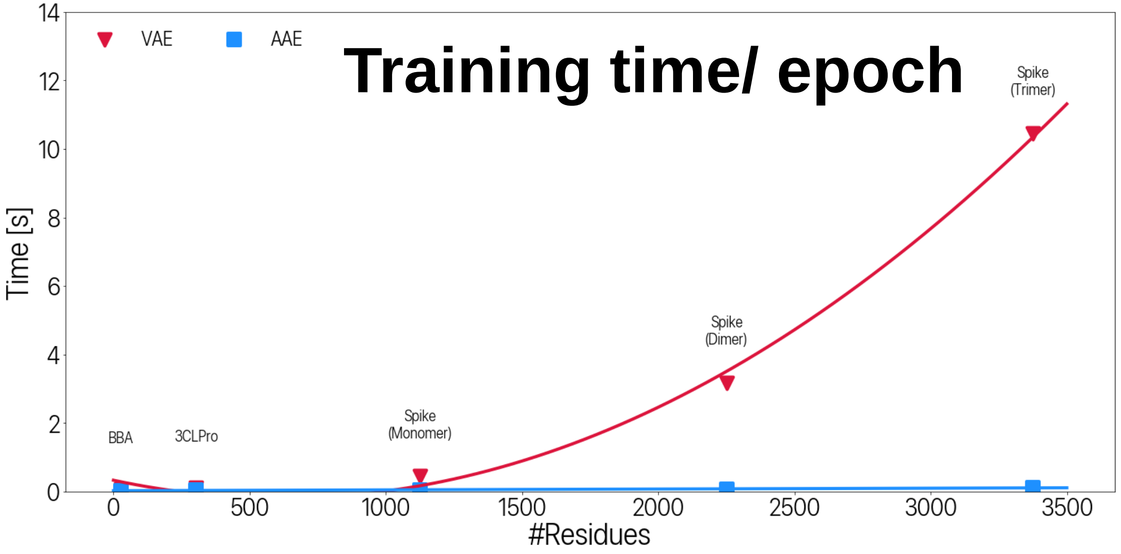
Algorithmic innovation:

- Point-cloud representations
- Adversarial autoencoder
- **O(10^4)** parameters

Thorsten Kurth, Abe Stern, Alex Brace, Tom Gibbs, Anda Trifan, Arvind Ramanathan

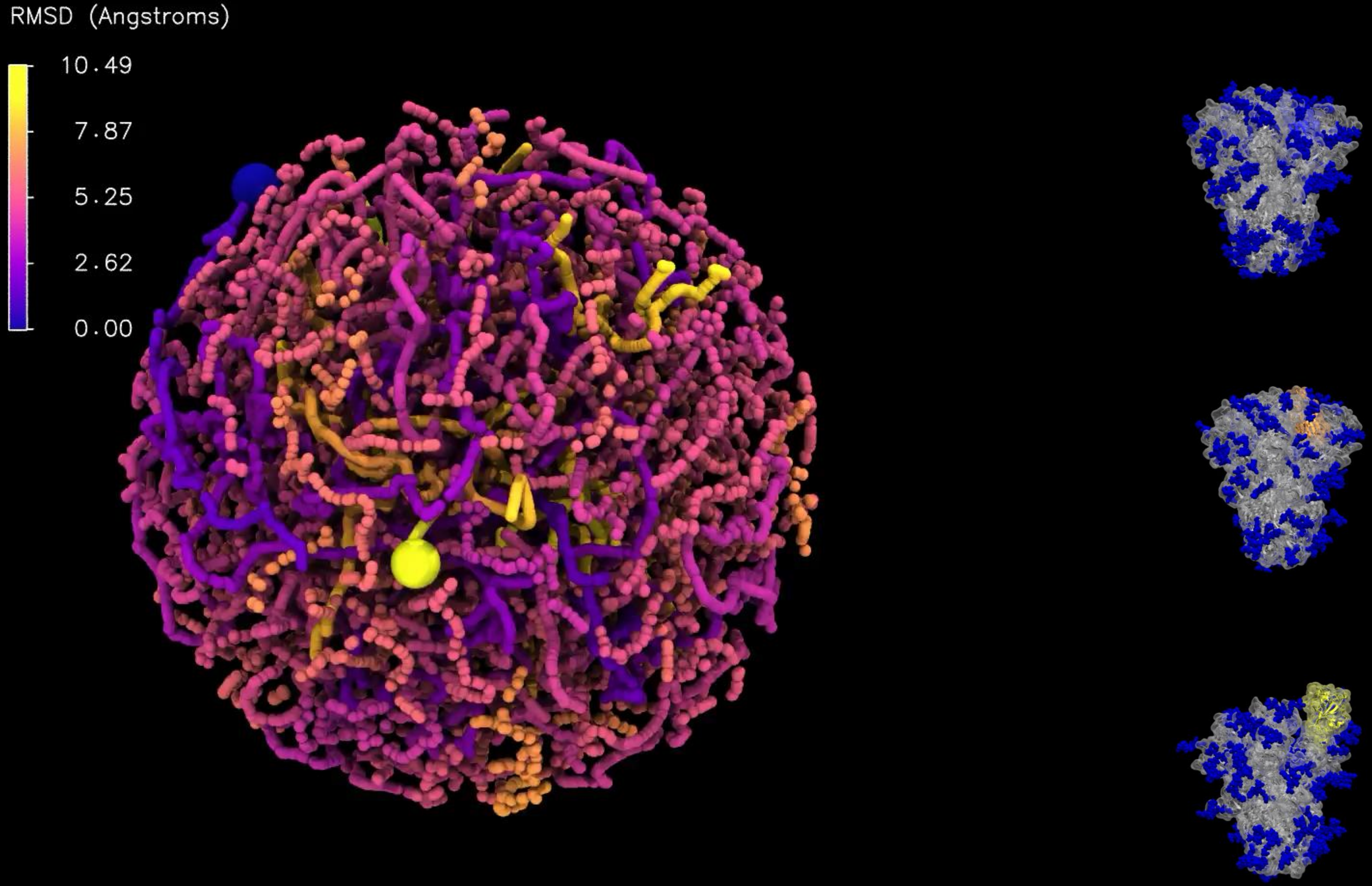
DeepDriveMD: Computational Performance

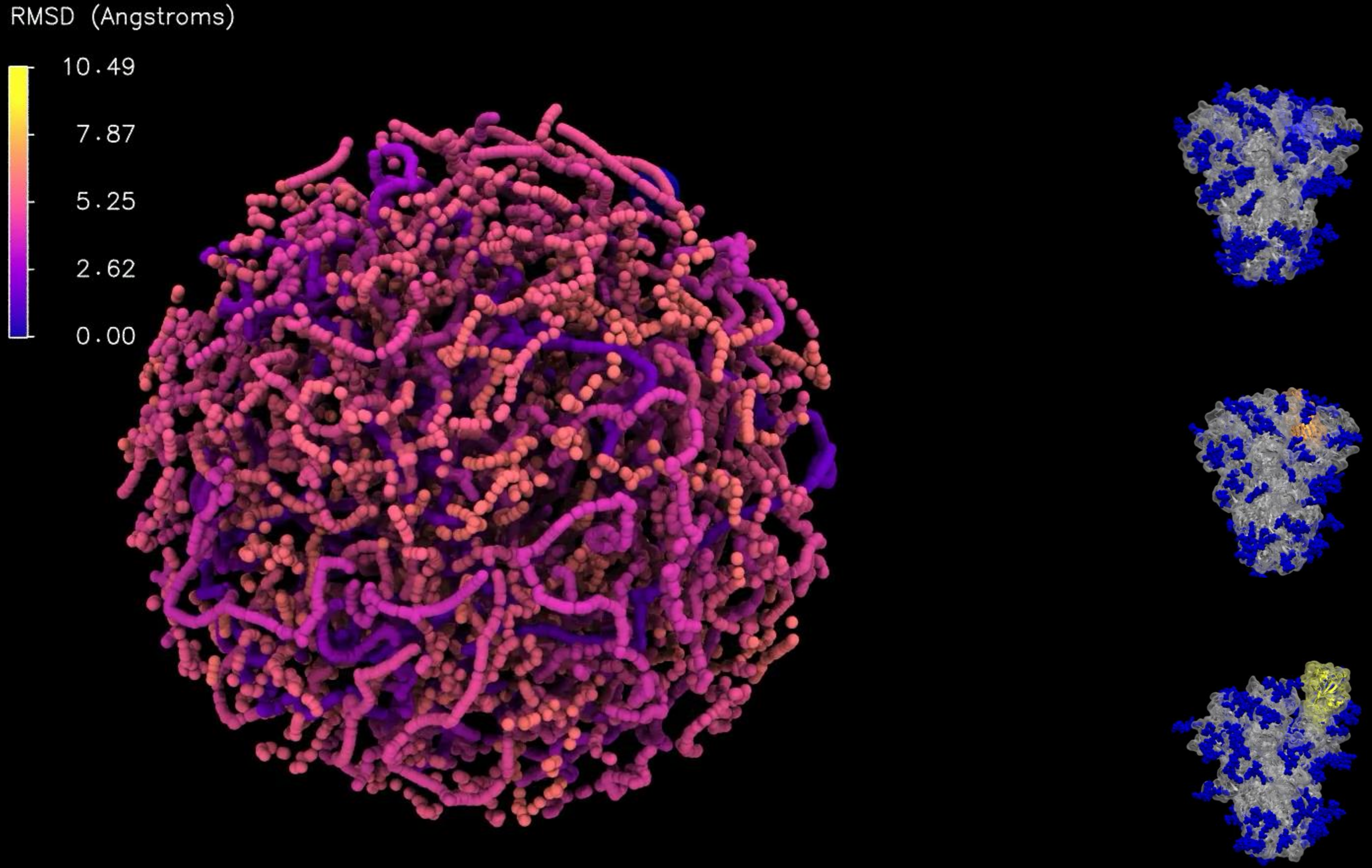


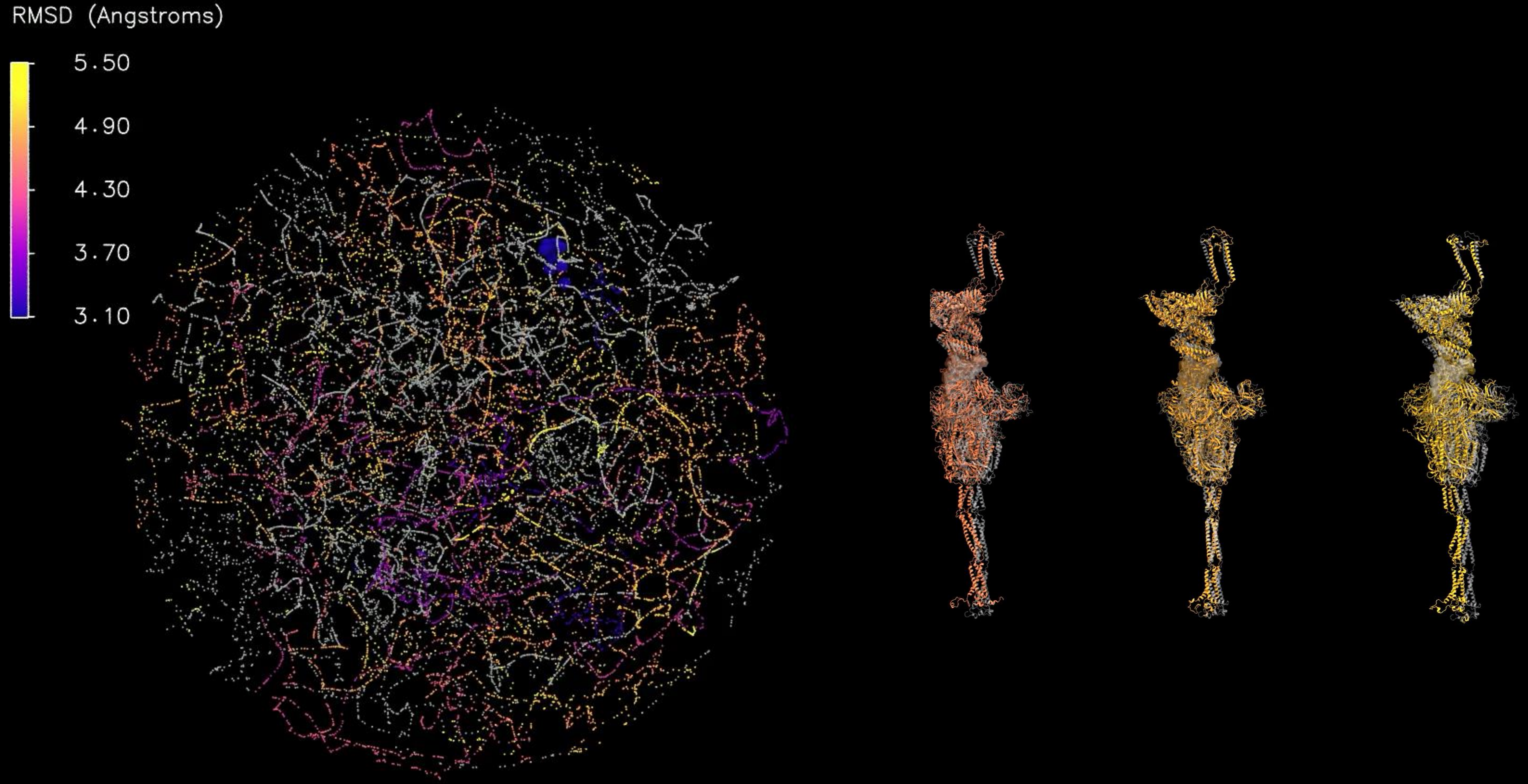


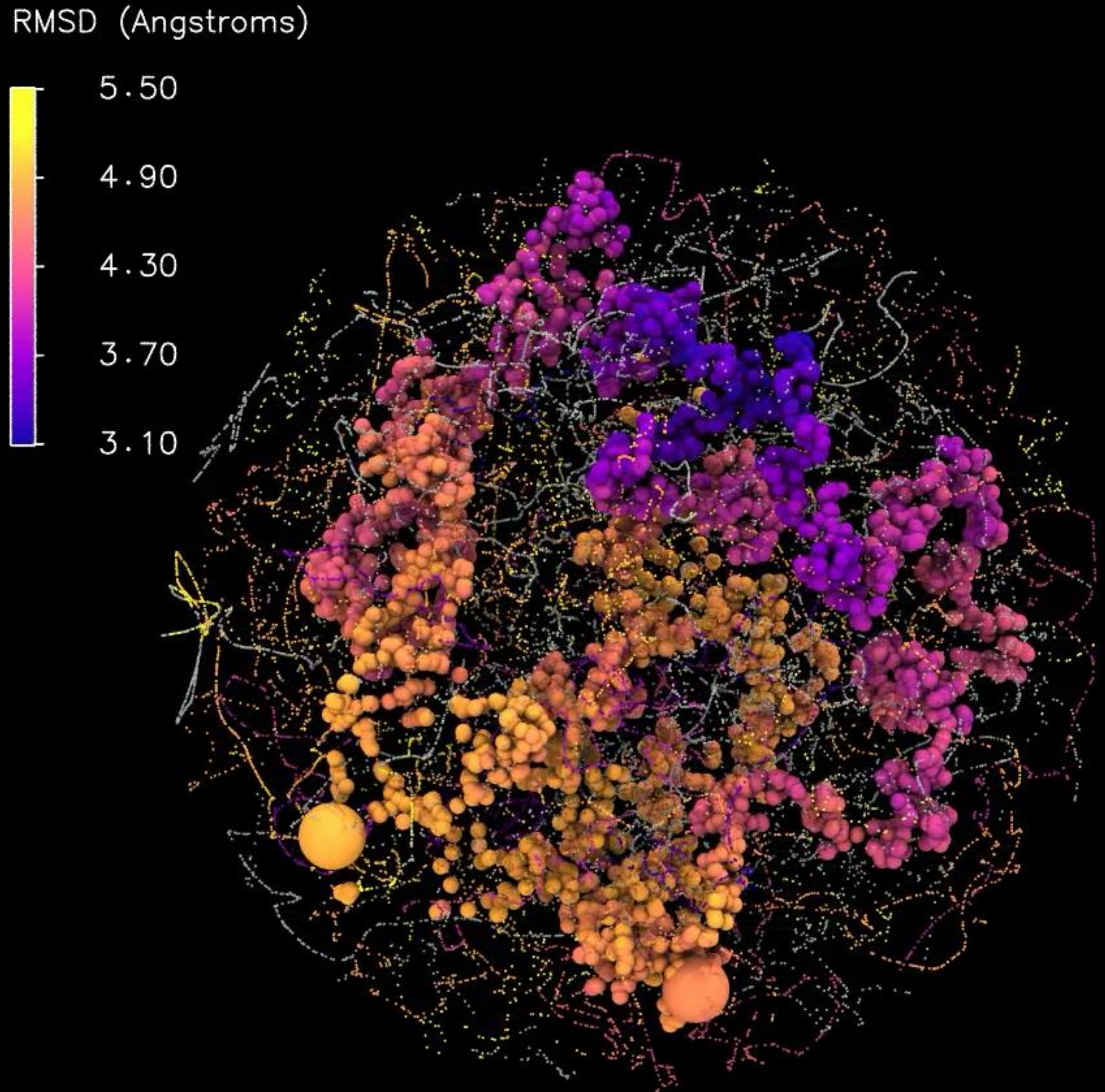
- **VAE**: larger memory footprint and longer training times
- **AAE**: can scale to much larger protein sizes and far more efficient in training time
  - linear increase in memory utilization
  - almost constant cost in training time (better scaling)

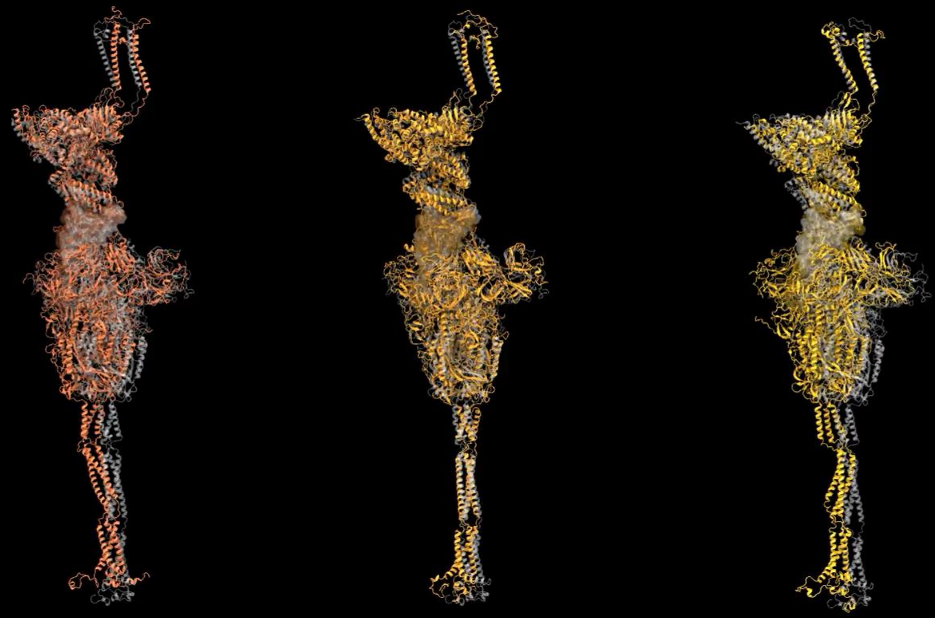
- Machine learning for protein folding and dynamics. Current Opinion in Structural Biology, (2020).
- Discovering protein conformational flexibility through artificial intelligence-aided molecular dynamics. Journal of Physical Chemistry (2019).
- Reinforcement learning based adaptive sampling: Reaping rewards by exploring protein conformational landscapes. The Journal of Physical Chemistry B (2018).


RMSD (Angstroms)
10.49
7.87
5.25
2.62
0.00

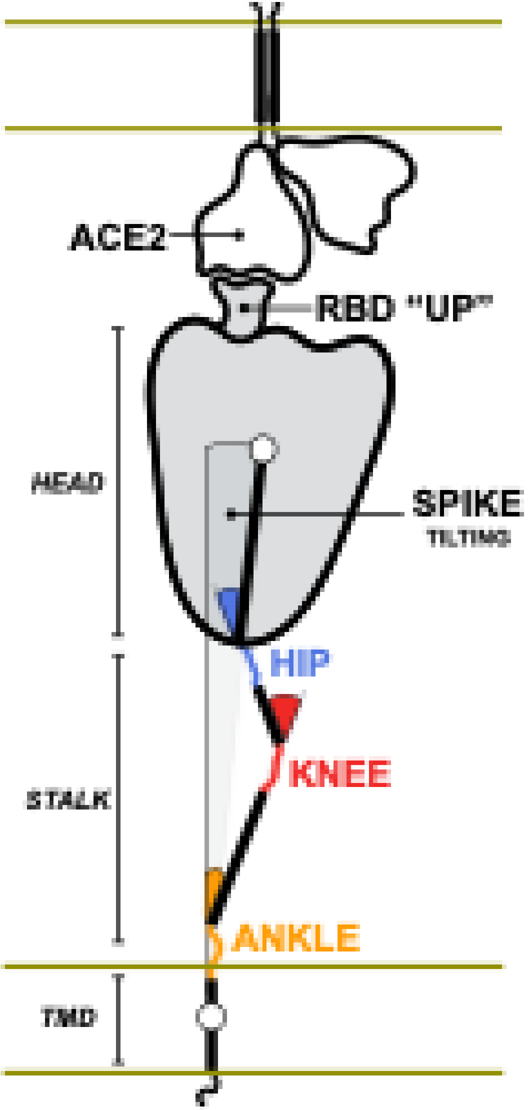

RMSD (Angstroms)
10.49
7.87
5.25
2.62
0.00

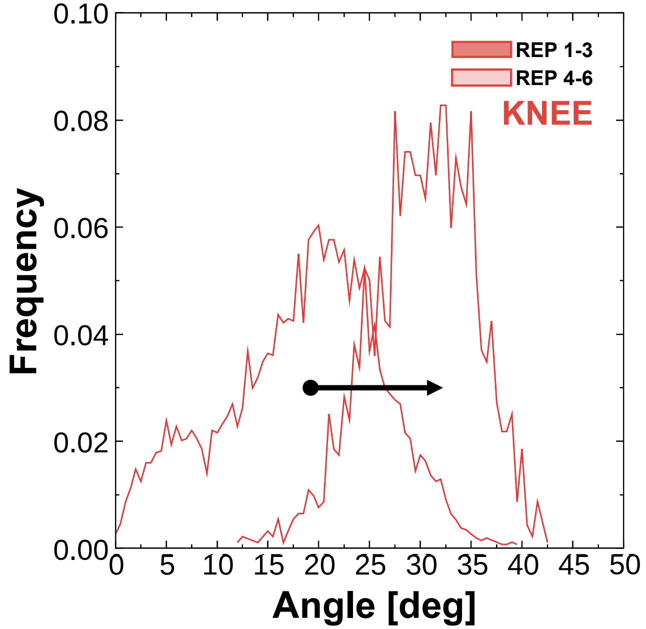

RMSD (Angstroms)
5.50
4.90
4.30
3.70
3.10

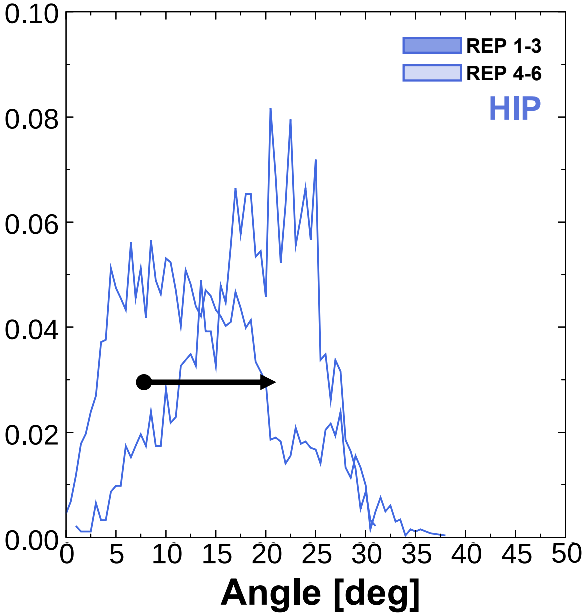

RMSD (Angstroms)
5.50
4.90
4.30
3.70
3.10



DeepDriveMD: Effective ”Scientific” Performance

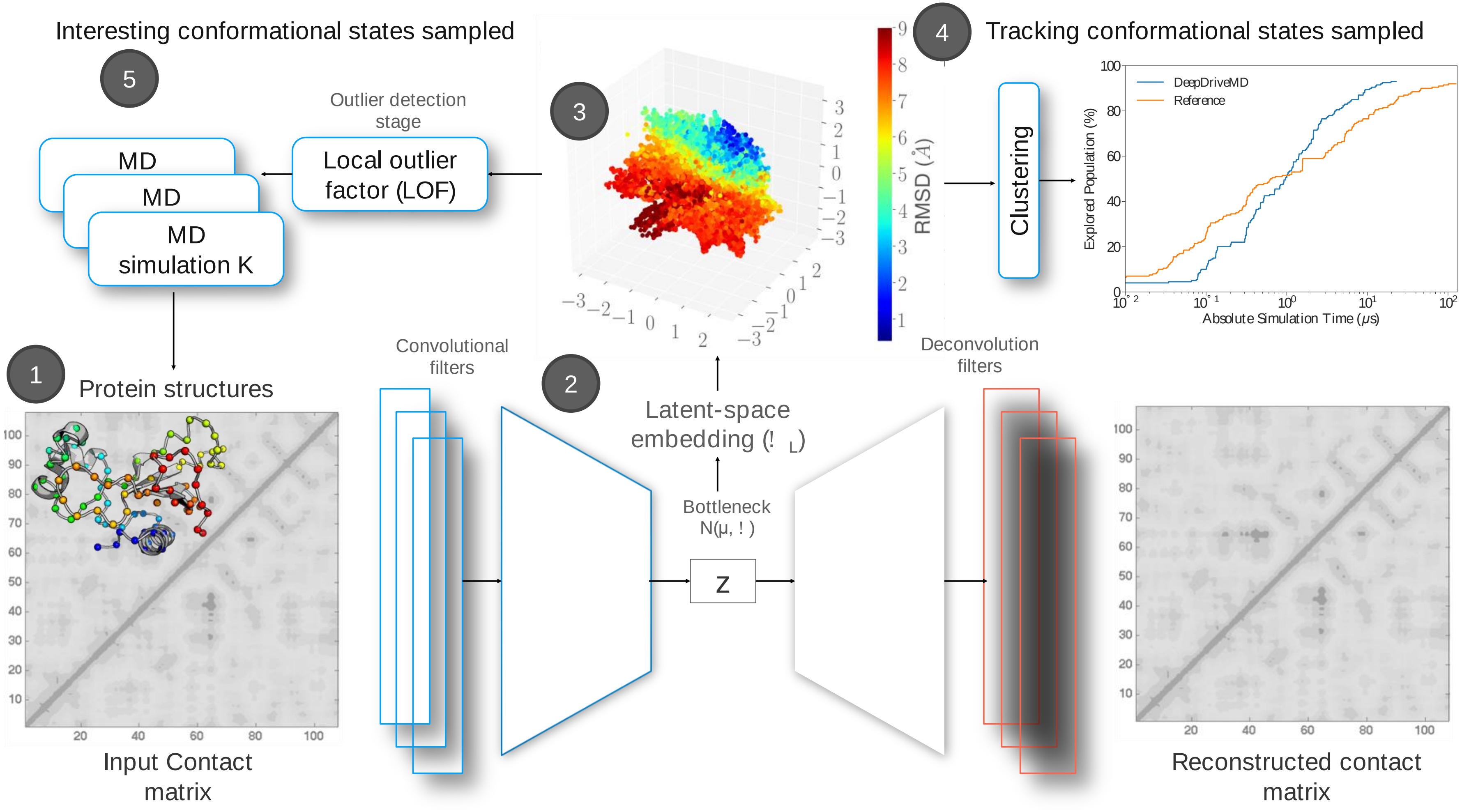






- Effective speedup of **o(8.3X)** sampling efficiency
  - without DeepDriveMD: 0.5 μs
  - with DeepDriveMD: 0.06 μs
- **Observed 25% more conformations of the knee bending in only 12% of the time!**
- Has been scaled to 1024 nodes of Summit for large ensembles

Enabling streaming AI/ML with multiscale simulations



Interesting conformational states sampled

5

Outlier detection stage

MD

MD

MD simulation K

Local outlier factor (LOF)

3

4

Tracking conformational states sampled

Clustering

1

Protein structures

Convolutional filters

2

Latent-space embedding (! L)

Bottleneck N(μ, !)

z

Deconvolution filters

Input Contact matrix

Reconstructed contact matrix

Cerebras CS-1: A 15 RU System for Training & Inference in the Data Center

Powered by the Cerebras Wafer Scale Engine (WSE):

- **400,000** AI optimized cores
- **18 GB** on chip memory—all 1 clock cycle from the cores,
  - 4 Billion parameters for training (FP 16); 16B inference (8int)
- **9 PByte/s** memory bandwidth
- **100 Pbit/s** fabric bandwidth

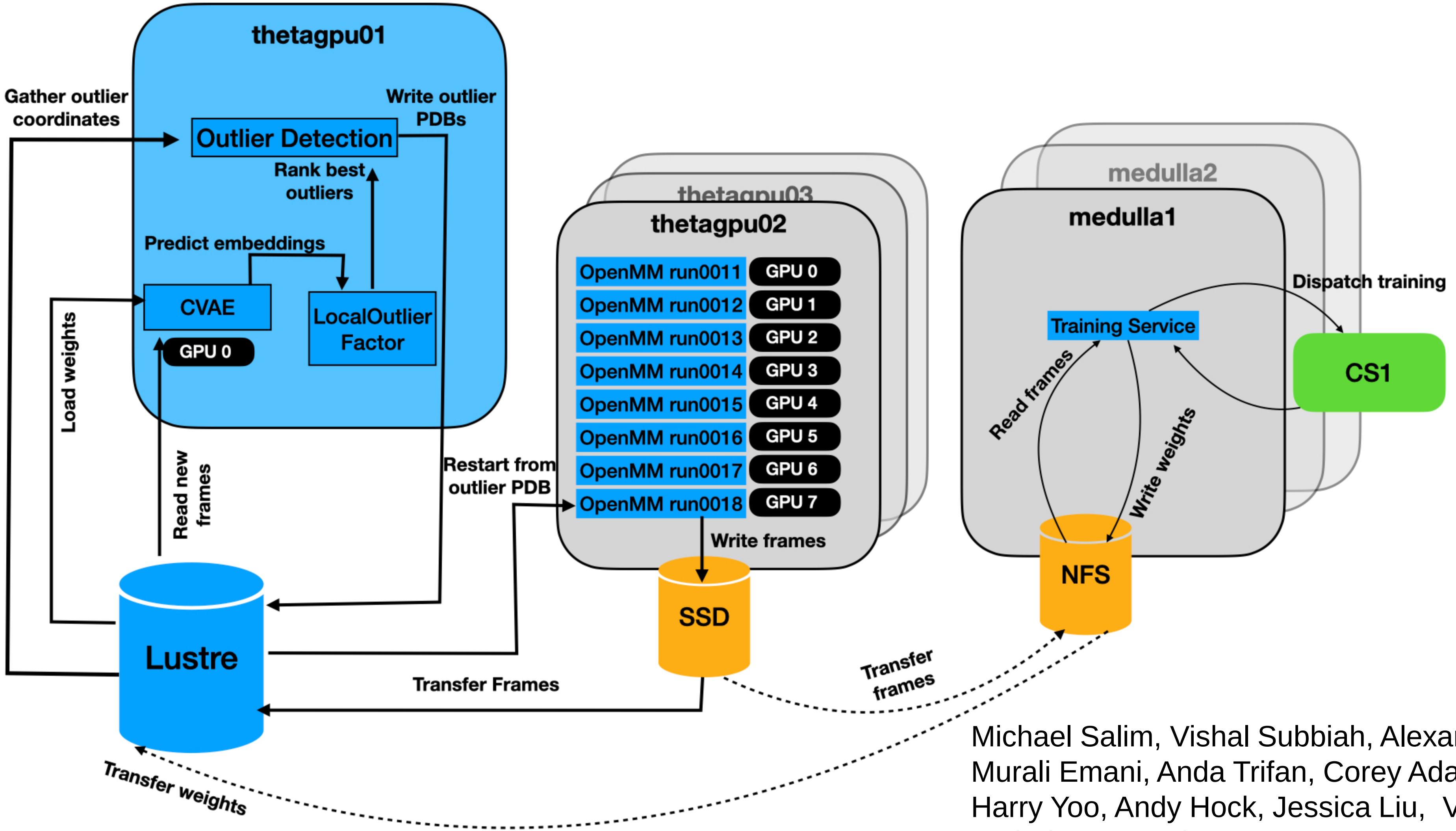
System IO: **12 x 100 GbE**

System power: **20 kW**

Ingests TensorFlow, PyTorch, etc.

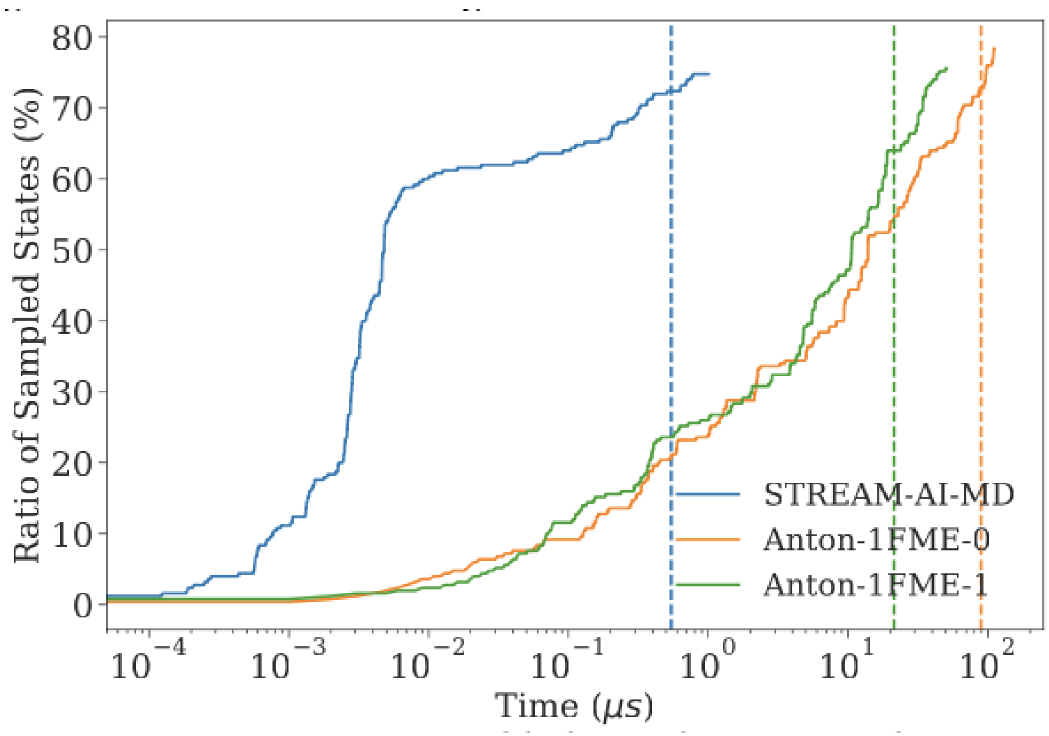
Courtesy: Cerebras Systems Inc.

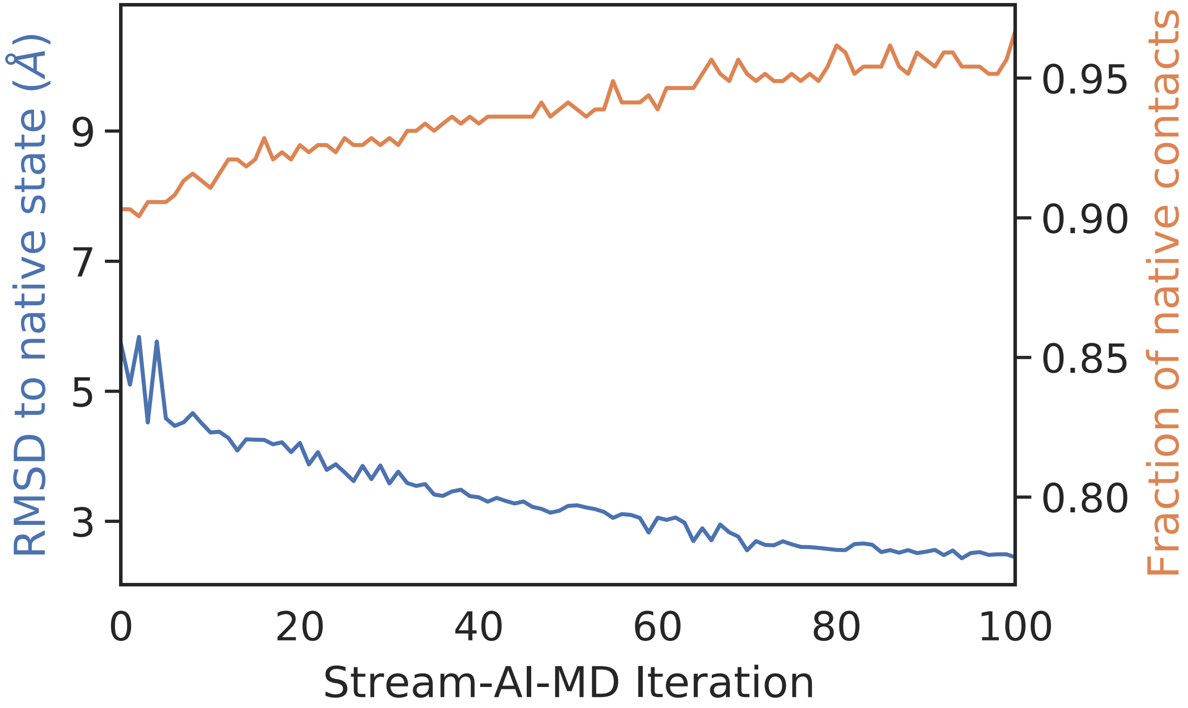
Bringing together heterogenous hardware to enable streaming analysis



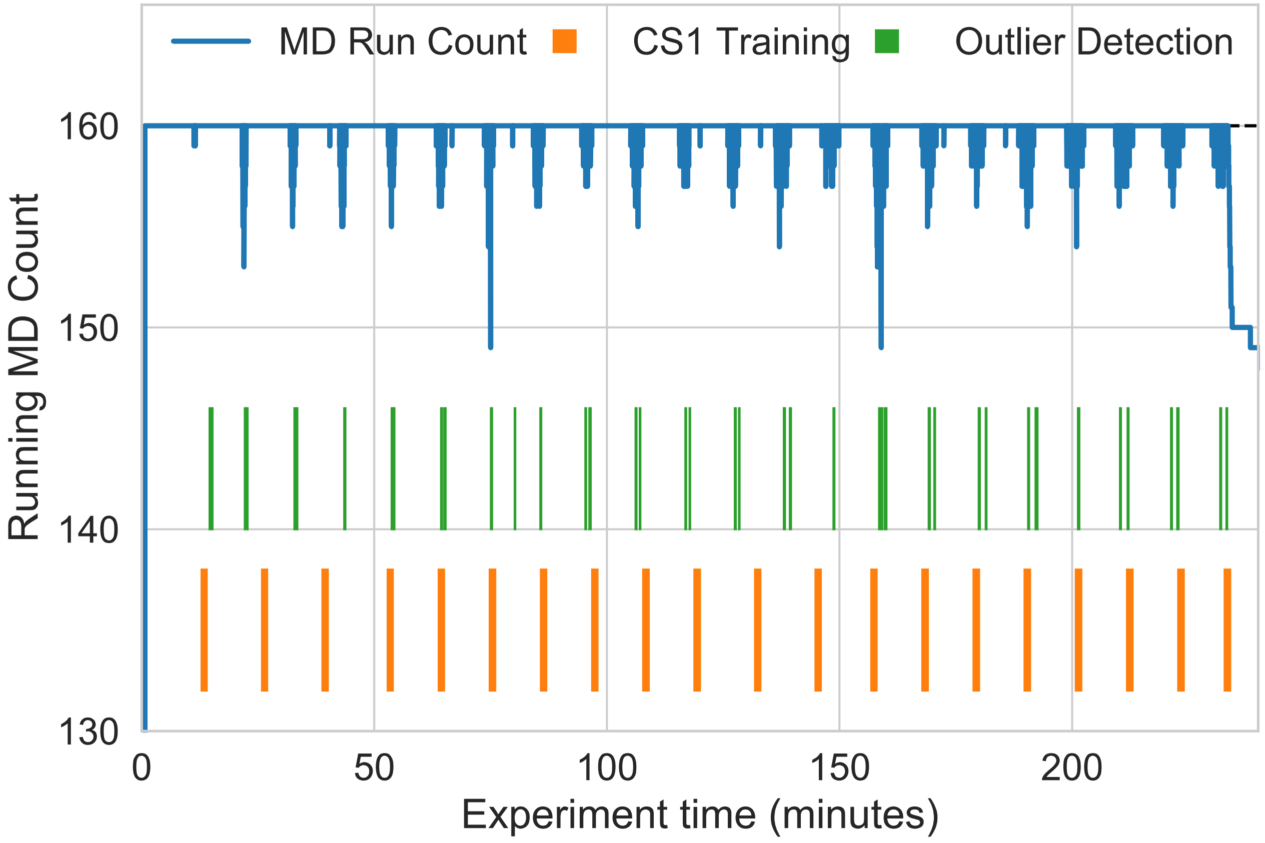
Michael Salim, Vishal Subbiah, Alexander Brace, Heng Ma, Murali Emani, Anda Trifan, Corey Adams, Thomas Uram, Harry Yoo, Andy Hock, Jessica Liu, Vernkat Vishwanath, Arvind Ramanathan

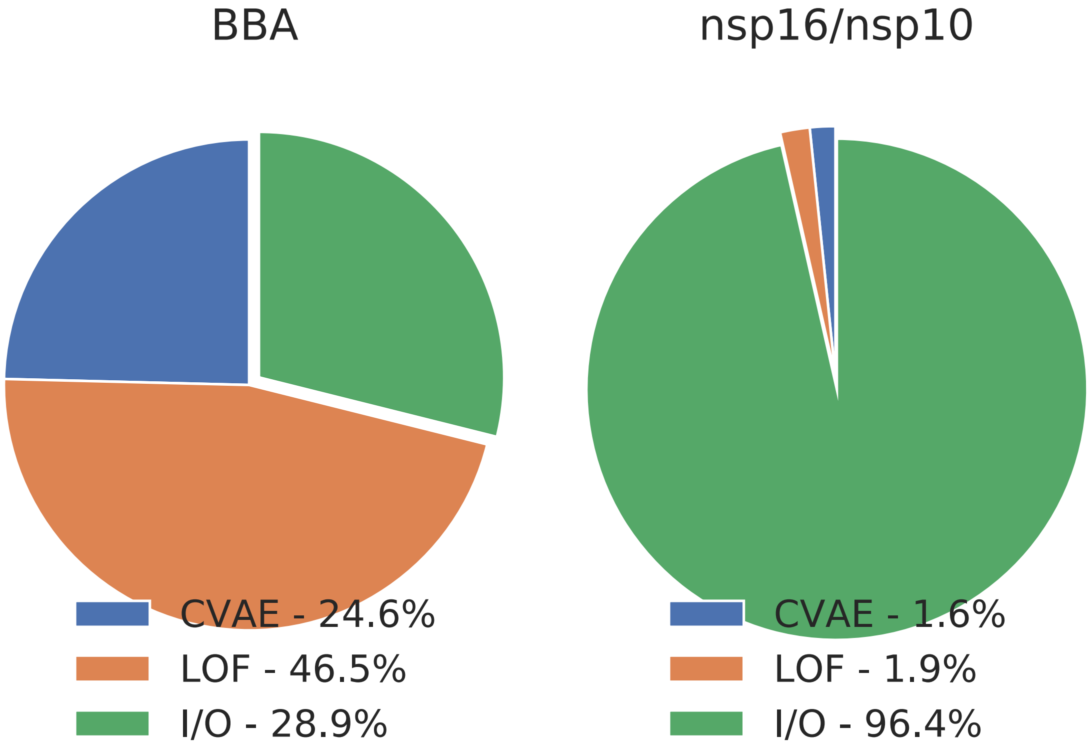
Stream-AI-MD enables at least 2 orders of magnitude faster sampling of folded states



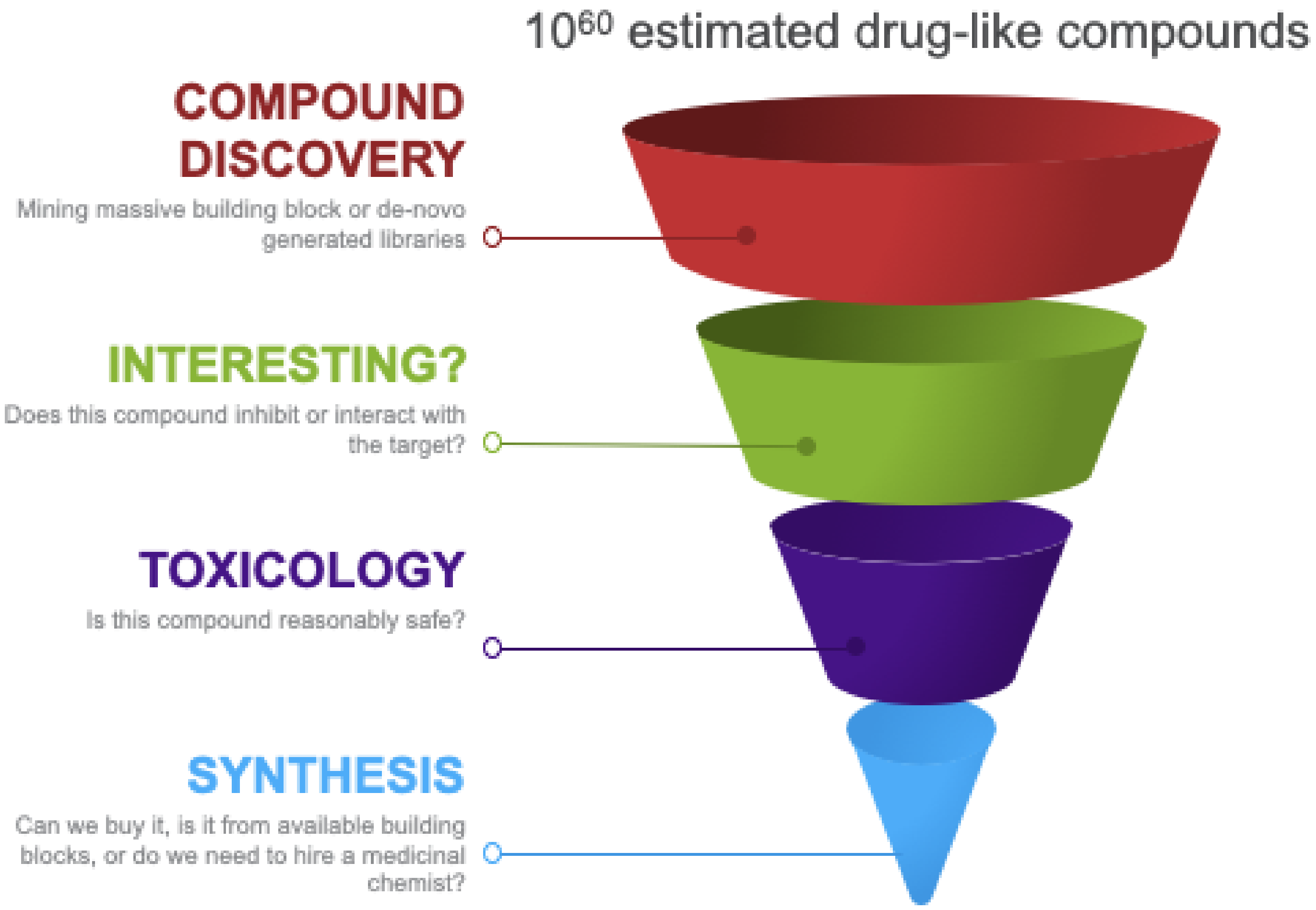


Utilization of hardware resources can vary depending on how tasks are scheduled





Outline (2)



10^{60} estimated drug-like compounds

COMPOUND DISCOVERY

Mining massive building block or de-novo generated libraries

INTERESTING?

Does this compound inhibit or interact with the target?

TOXICOLOGY

Is this compound reasonably safe?

SYNTHESIS

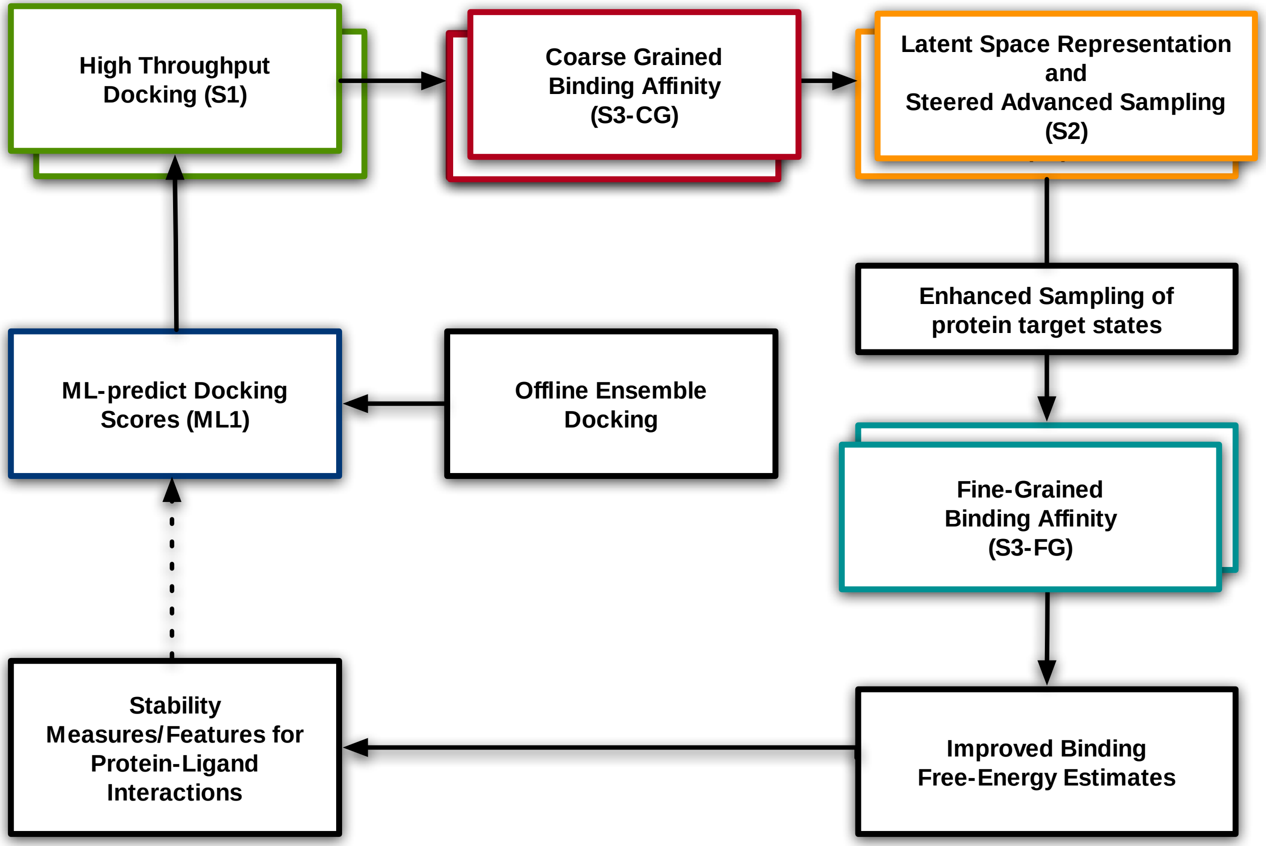
Can we buy it, is it from available building blocks, or do we need to hire a medicinal chemist?

How to search billions of molecules to find drug candidates?

Improving docking and finding better ligands that bind to SARS-COV-2 proteome

Multi-stage campaign employed to select promising drug candidates:

- **Stage-1**: High-throughput ensemble docking to identify small molecules ("hits")
- **Stage-2**: AI-driven Molecular Dynamics for modeling specific binding regions and understanding mechanistic changes involving drugs
- **Stage-3**: Binding Free Energy calculations of promising leads and (expensive) lead optimization



High Throughput Docking (S1)
Coarse Grained Binding Affinity (S3-CG)
Latent Space Representation and Steered Advanced Sampling (S2)
Enhanced Sampling of protein target states
ML-predict Docking Scores (ML1)
Offline Ensemble Docking
Fine-Grained Binding Affinity (S3-FG)
Stability Measures/Features for Protein-Ligand Interactions
Improved Binding Free-Energy Estimates

Aymen Al Saadi, Dario Alfe, Yadu Babuji, Agastya Bhati, Ben Blaiszik, Thomas Brettin, Ryan Chard, **Anda Trifan**, Alex Brace, Austin Clyde, Ian Foster, Tom Gibbs, Kristopher Keipert, Thorsten Kurth, Dieter Kranzlmüller, Hyungro Lee, Heng Ma, Andre Merzky, Gerald Matthias, Alexander Partin, Junqi Qiu, Ashka Shah, Abraham Stern, Li Tan, Mikhail Titov, Aristedis Tsaris, Matteo Turilli1, Huub Van Dam, Shunzhou Wan, David Wifling, Shantenu Jha*, Peter Coveney∗, Rick Stevens*, Arvind Ramanathan∗

Why not dock every available compound?

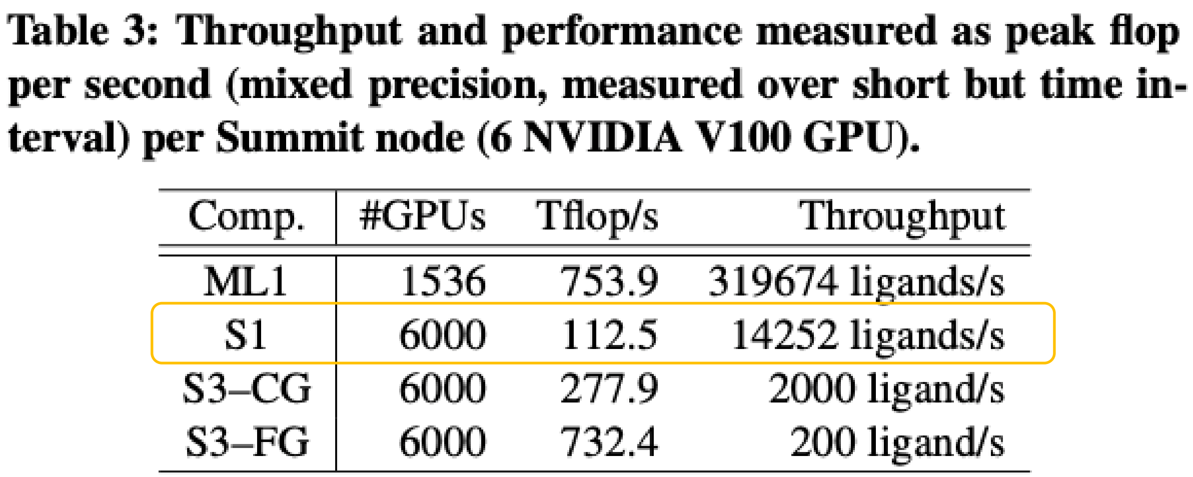


Table 3: Throughput and performance measured as peak flop per second (mixed precision, measured over short but time interval) per Summit node (6 NVIDIA V100 GPU).

Comp.	#GPUs	Tflop/s	Throughput
ML1	1536	753.9	319674 ligands/s
S1	6000	112.5	14252 ligands/s
S3–CG	6000	277.9	2000 ligand/s
S3–FG	6000	732.4	200 ligand/s

- S1 → O(15,000) ligands/sec on 6 GPUs ➔ all of Summit will still take ~6.8 – 8 hours to compute!!
- This is on one receptor → 100 receptors is not feasible

The COVID'19 data pipeline:

Developing machine readable datasets for small molecule libraries

CHEMICAL LIBRARY DATABASE

4B known molecules

Enamine

DRUGBANK

BindingDB GDB

eMolecules

cureFFI MOSES

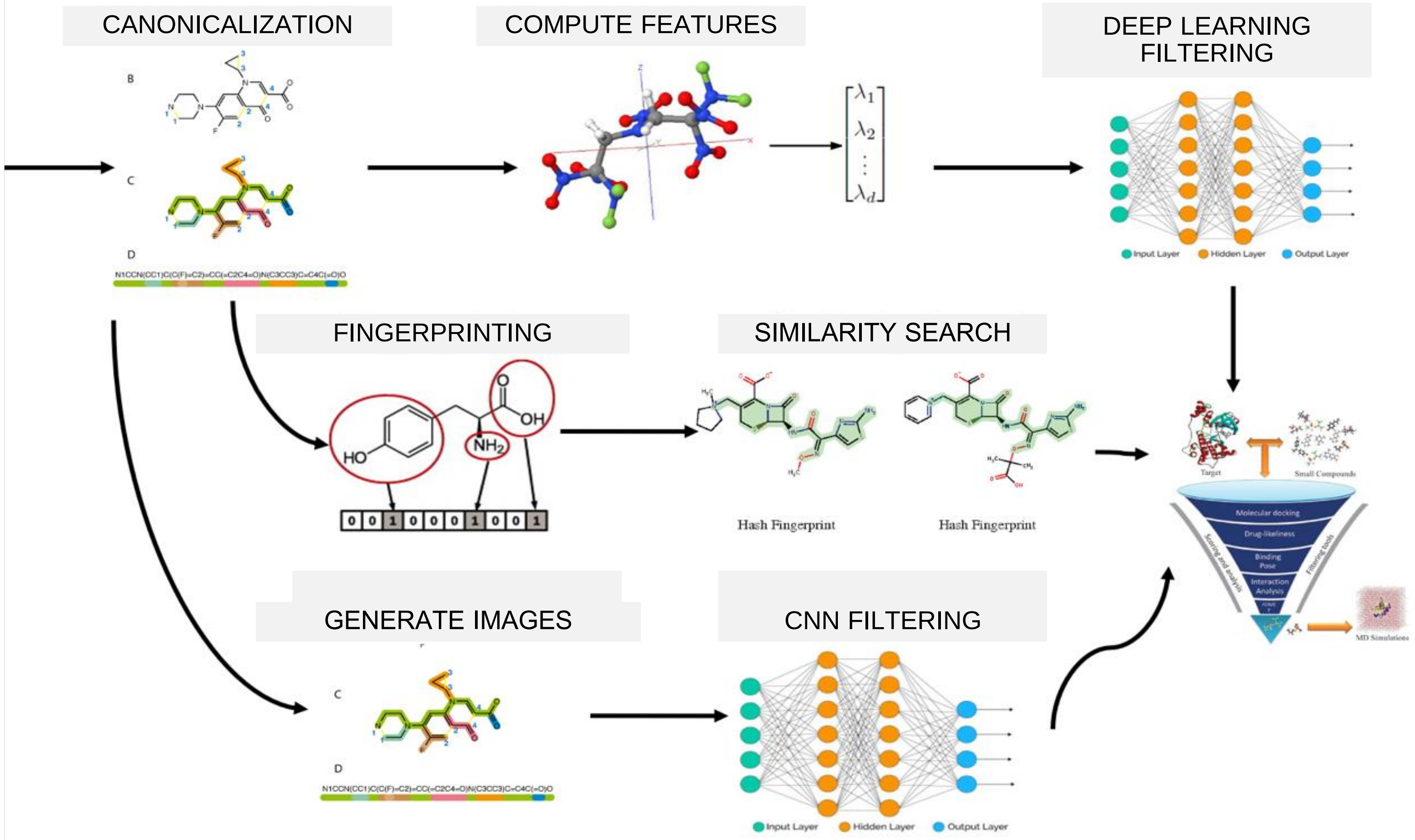
ZINC15

LINCS

SureChEMBL

PubChem

AND MORE


CANONICALIZATION

COMPUTE FEATURES

DEEP LEARNING FILTERING

FINGERPRINTING

SIMILARITY SEARCH

GENERATE IMAGES

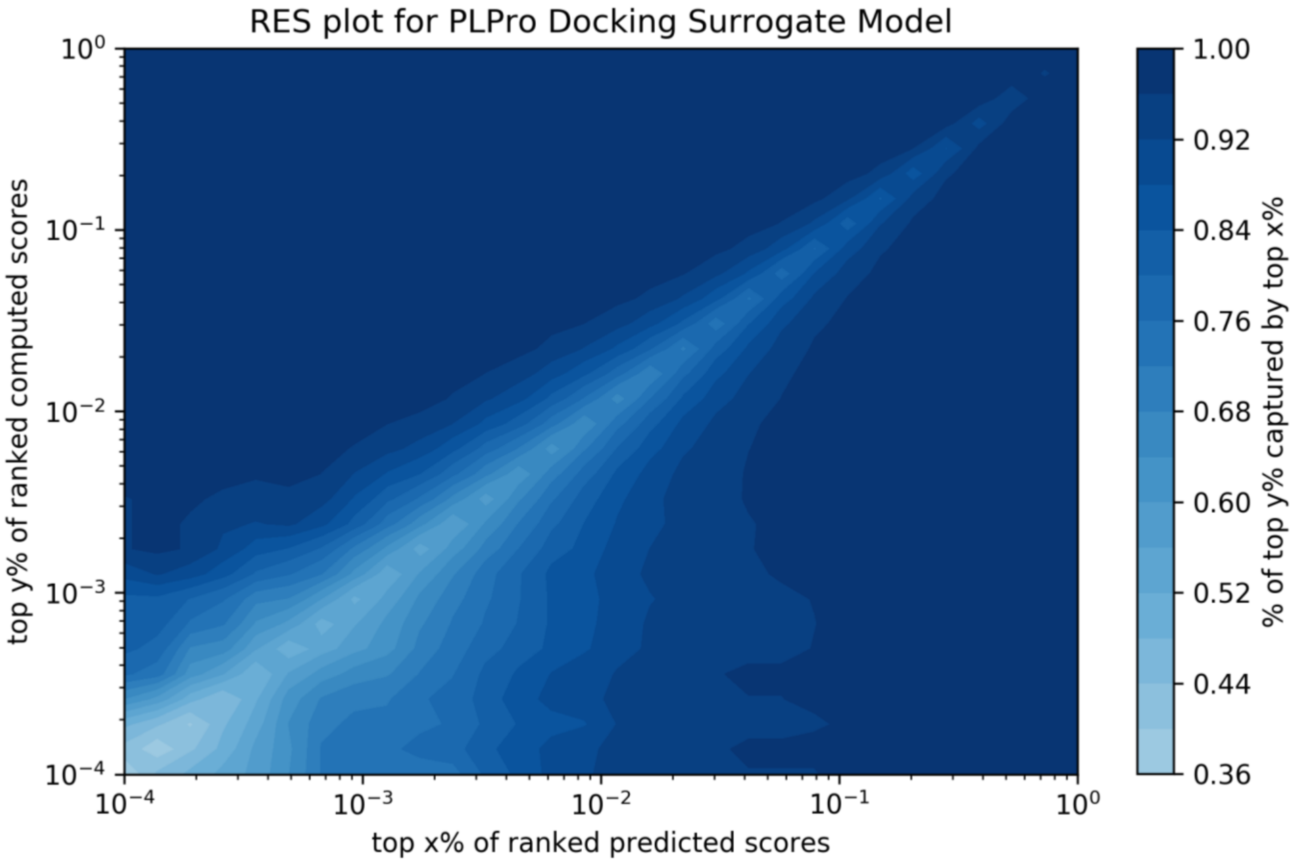
CNN FILTERING

COMPUTING RESOURCES

Yadu Babuji, Ben Blaiszik, Kyle Chard, Ryan Chard, Ian Foster, Logan Ward, Tom Brettin et al

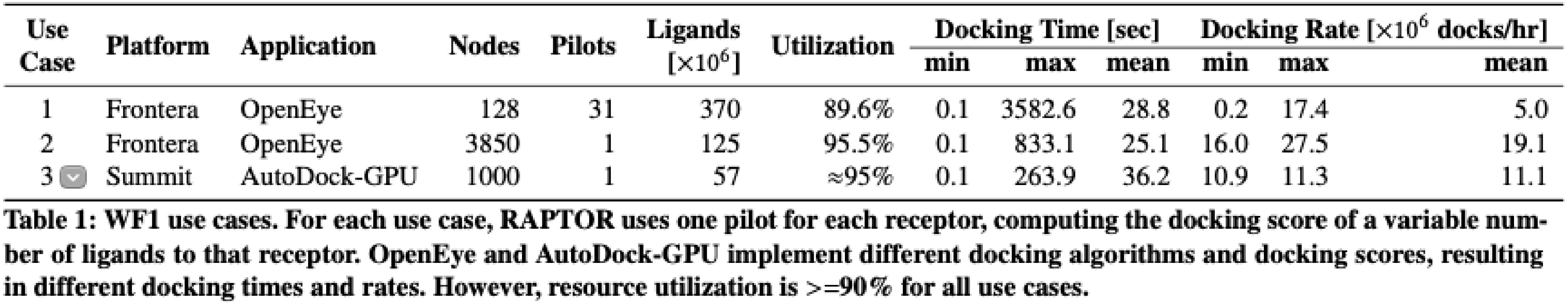
ML to the rescue! Increased scientific throughput for virtual screening

- Instead of docking or predicting the docking pose, predict:
  - the docking score: a regression problem
  - whether a molecule will bind to a given protein target
- ML problem formulation: how many compounds can we find at the top-ranking list given some training data?
  - still uses the regression problem
  - instead of ranking we provide a bound for saying how many compounds we need to dock before we get ”true hits”
- Leverage image-based models (CNNs on image with rotation invariant formulations) that are well optimized



A. Clyde, R. Stevens, Regression Enrichment Surfaces, https://arxiv.org/abs/2006.01171
https://github.com/aclyde11/regression_enrichment_surface

Computational performance



Use Case	Platform	Application	Nodes	Pilots	Ligands [$\times 10^6$]	Utilization	Docking Time [sec]			Docking Rate [$\times 10^6$ docks/hr]		
							min	max	mean	min	max	mean
1	Frontera	OpenEye	128	31	370	89.6%	0.1	3582.6	28.8	0.2	17.4	5.0
2	Frontera	OpenEye	3850	1	125	95.5%	0.1	833.1	25.1	16.0	27.5	19.1
3	Summit	AutoDock-GPU	1000	1	57	≈95%	0.1	263.9	36.2	10.9	11.3	11.1

Table 1: WF1 use cases. For each use case, RAPTOR uses one pilot for each receptor, computing the docking score of a variable number of ligands to that receptor. OpenEye and AutoDock-GPU implement different docking algorithms and docking scores, resulting in different docking times and rates. However, resource utilization is >=90% for all use cases.

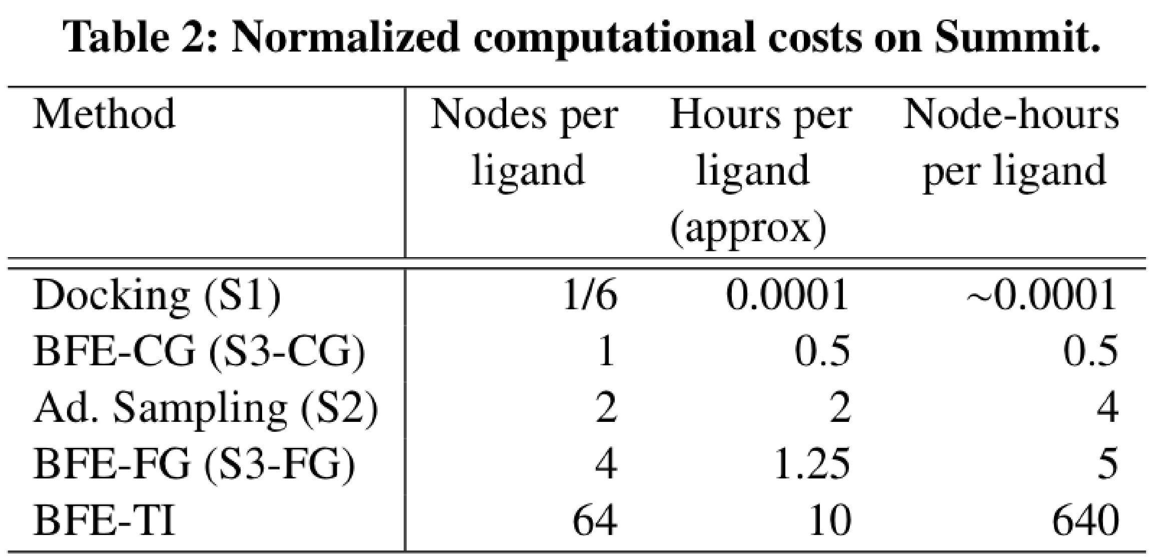


Table 2: Normalized computational costs on Summit.

Method	Nodes per ligand	Hours per ligand (approx)	Node-hours per ligand
Docking (S1)	1/6	0.0001	~0.0001
BFE-CG (S3-CG)	1	0.5	0.5
Ad. Sampling (S2)	2	2	4
BFE-FG (S3-FG)	4	1.25	5
BFE-TI	64	10	640

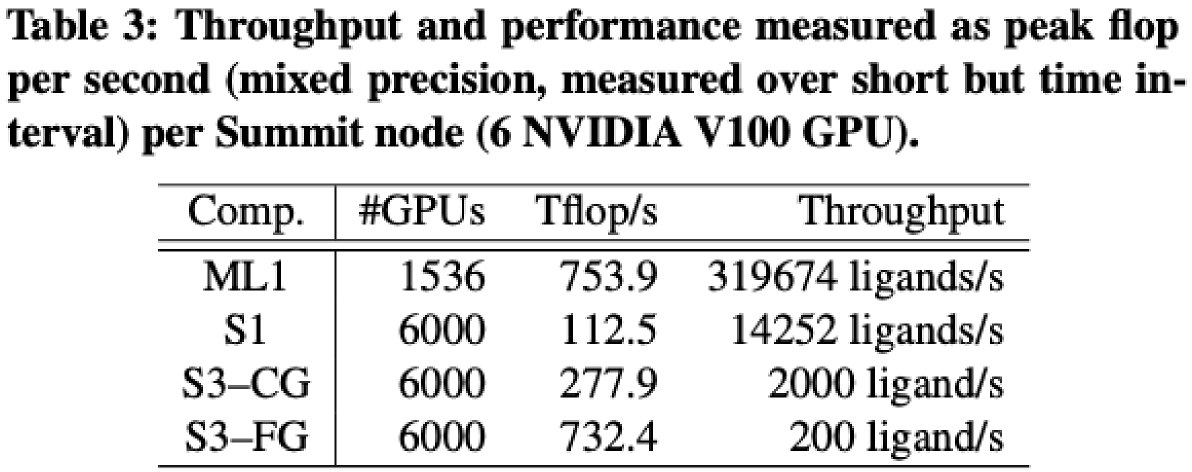
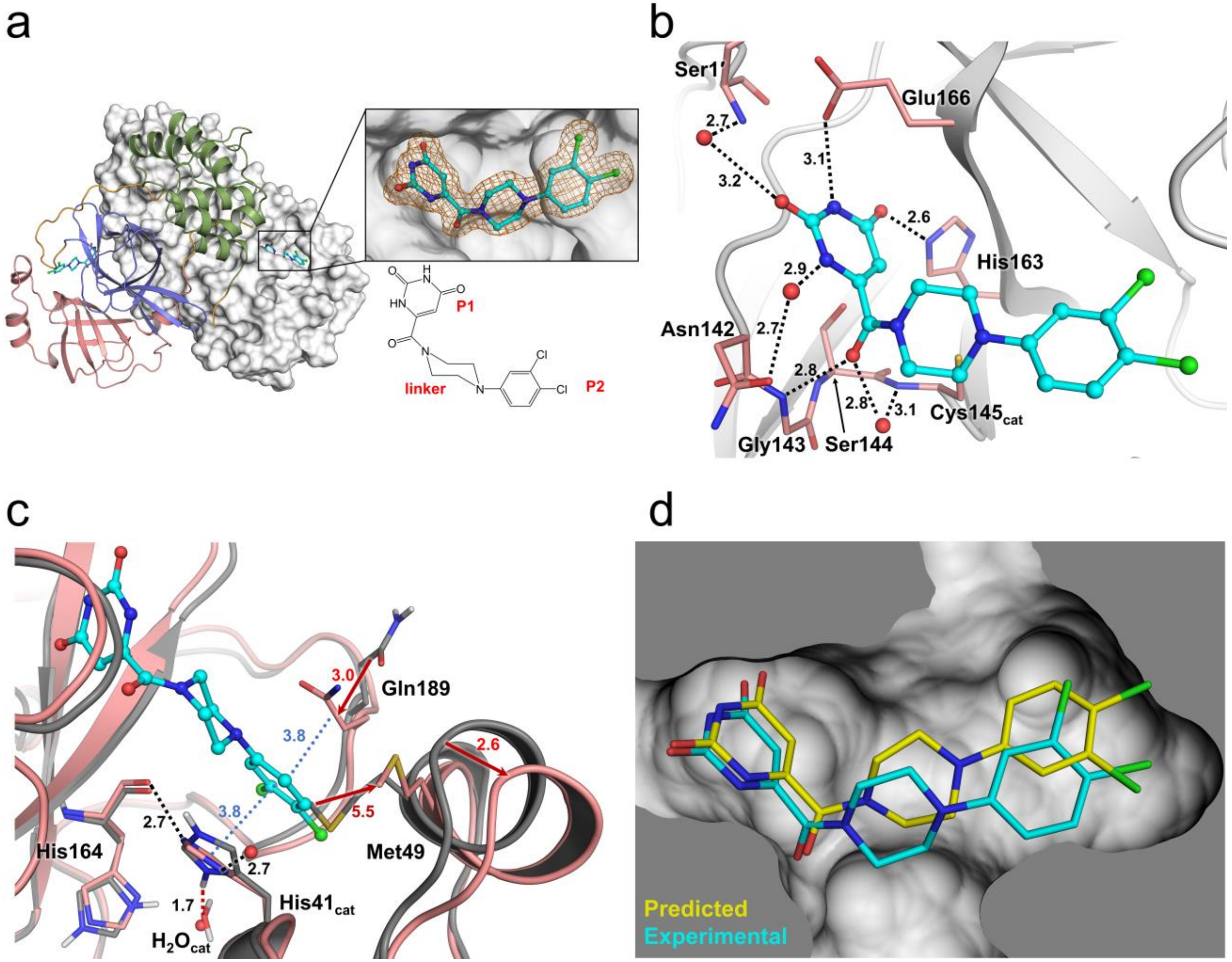


Table 3: Throughput and performance measured as peak flop per second (mixed precision, measured over short but time interval) per Summit node (6 NVIDIA V100 GPU).

Comp.	#GPUs	Tflop/s	Throughput
ML1	1536	753.9	319674 ligands/s
S1	6000	112.5	14252 ligands/s
S3–CG	6000	277.9	2000 ligand/s
S3–FG	6000	732.4	200 ligand/s

Shantenu Jha and team

Our workflow results in better binding compounds …



- From the 1000 compounds were ordered for whole-cell assays, ~50 of compounds show viral inhibition activity
- Several compounds have already been processed for X-ray crystallography efforts (at Argonne and NSLS-II)
- Synthetic chemistry efforts are being driven across labs to either optimize compounds

Impacting SARS-CoV-2 Medical Therapeutics

- Scale of operation:
  - ~10^{11} docking calculations using OpenEye and Autodock in ratio 10:1
  - Thousands of DeepDriveMD calculations over multiple platforms (Summit, Lassen, …)
  - 5 x 10^4 Binding Free Energy Calculations across machines
  - 2.5 x 10^6 node-hours (equal to ~25 days of 100% of Summit)
    - Assuming 5-year lifetime of Summit at $500M → $6M cost of computing!
- For S1, we estimate 1.25 x 10^6 node hours (lower bound)
  - Peak Performance: ~4000 nodes for docking studies on Frontera (06 Sep 2020),
- Robust and Extensible Computational Infrastructure and Capabilities
  - Campaign -- 24x7 operation over multiple heterogeneous resources
  - AI-methods & Software Systems can be extended to ATOM, and other drug discovery pipelines
  - Extending computational infrastructure to NSLS-II covalent inhibitors of cysteine proteases

Funding and acknowledgements

- Everyone in the team (all ~300)
- Computing support:
  - ALCF, OLCF
  - TACC, SDSC, IU
  - HPC Consortium
- Funding acknowledgement:
  - DOE National Virtual Biotechnology Laboratory (NVBL)
  - Argonne internal funding (LDRD)
  - DOE Exascale computing project (Cancer Deep Learning Environment)

THANK YOU!!!

ramanathana@anl.gov