CompBioMed
A Centre of Excellence in Computational Biomedicine

Webinar series

Webinar #13

"**Are** we there yet?" - Addressing emerging drug discovery challenges for SARS-CoV-2 with artificial intelligence driven molecular dynamics

16 September 2020

The webinar will start at 4pm BST

Speaker: **Arvind Ramanathan** (Argonne National Lab)

Moderator: Katya Ahmad (UCL)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823712

The series is run in collaboration with: VPH Institute Building the Virtual Physiological Human

CompBioMed
A Centre of Excellence in Computational Biomedicine

Webinar series

Webinar #13

“**Are** we there yet?” - Addressing emerging drug discovery challenges for SARS-CoV-2 with artificial intelligence driven molecular dynamics

16 September 2020

Welcome!

Speaker: **Arvind Ramanathan** (Argonne National Lab)

Moderator: Katya Ahmad (UCL)

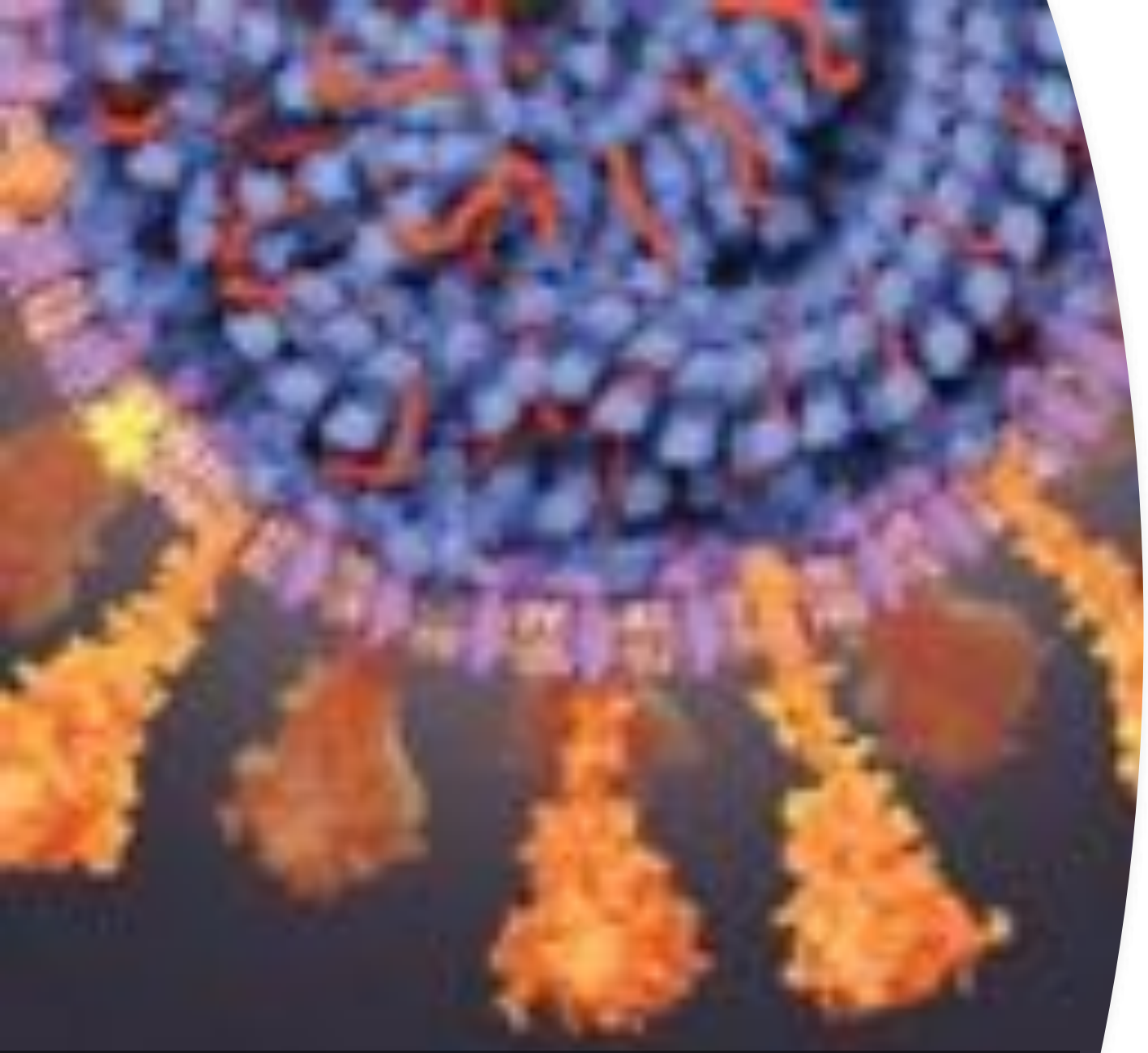
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 823712

The series is run in collaboration with: VPH Institute Building the Virtual Physiological Human

AI-integrated Drug Discovery for SARS-CoV2

Arvind Ramanathan

Argonne National Laboratory/
University of Chicago



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

COVID-19 and its impact on the US population

- Goals:
  - identify small molecules (or other therapeutic strategies) to treat COVID-19
  - develop AI/computing methods that fast-track therapeutic discoveries at scale

source: NY Times, Sep 10, 2020

National Virtual Biotechnology Lab (NVBL)

- Aid U.S. policymakers in responding to the COVID-19 pandemic with epidemiological information for decision making
- Accelerate production of critical medical supplies across the nation
- Supercomputing and artificial intelligence for design of targeted therapeutics
- Leverage chemical testing, analysis and biology within DOE to facilitate new ways for antigen and antibody testing

Argonne's strategy: 4 key areas

- **Computational drug screening**: can we design antiviral drugs by repurposing existing compounds against the entire viral proteome?
- **Epitope analysis and vaccine design**: what are the evolutionary hotspots or weaknesses in the virus that we can exploit to design immune system response?
- **Computational epidemiology**: can we model realistic scenarios of disease spread and outcomes?
- **Viral evolutionary analysis**: what are the evolutionary origins of the virus? Can we map the evolutionary trajectory of the virus?
- All areas leverage data science, AI, and HPC at Argonne and across multiple supercomputing facilities

SARS-CoV-2 target identification

Network-based methods for host pathogen interactions

PDB dataset

Candidate targets

AI-based loop modeling and structures

Workflow 0

Drug Databases
(ENAMIN, DRUGBANK, ZINC, etc.)

Featurization
(MOdred, graph/neural fingerprints)

Physics-based Docking

RL-Dock: AI approach

Active learning strategies for virtual screening (Rick Stevens)

10,000,000,000 compounds screened with AI models

Top 2.5%

250,000,000 poses docked

Top 2.5%

6,250,000 systems build and minimized

Top 2.5%

156,250 systems simulated

(that's about 12H on 1024 summit nodes)

Physics based models

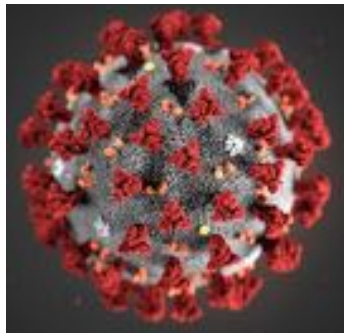
AI-based models

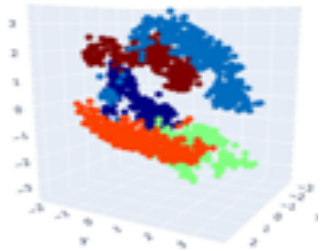
Workflow 1a, 1b

AI-based consensus ranking / scoring

Multi-scale, multi-resolution modeling of the SARS-CoV-2 viral proteins

Workflow 2b

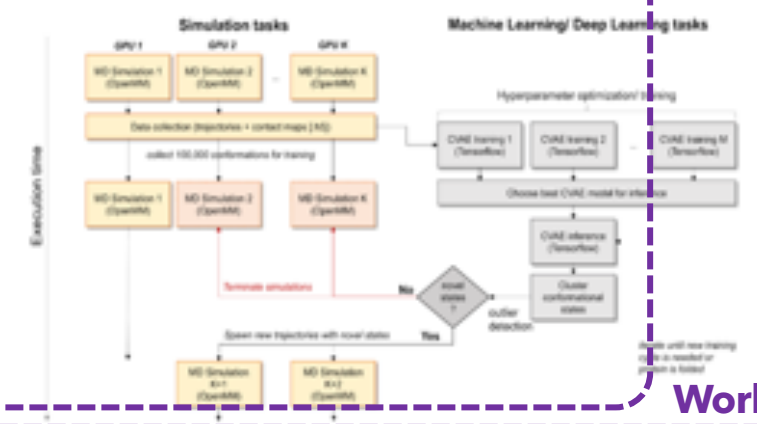




Rommie Amaro

Carlos Simmerling

DeepDriveMD: AI-driven adaptive sampling/ simulations

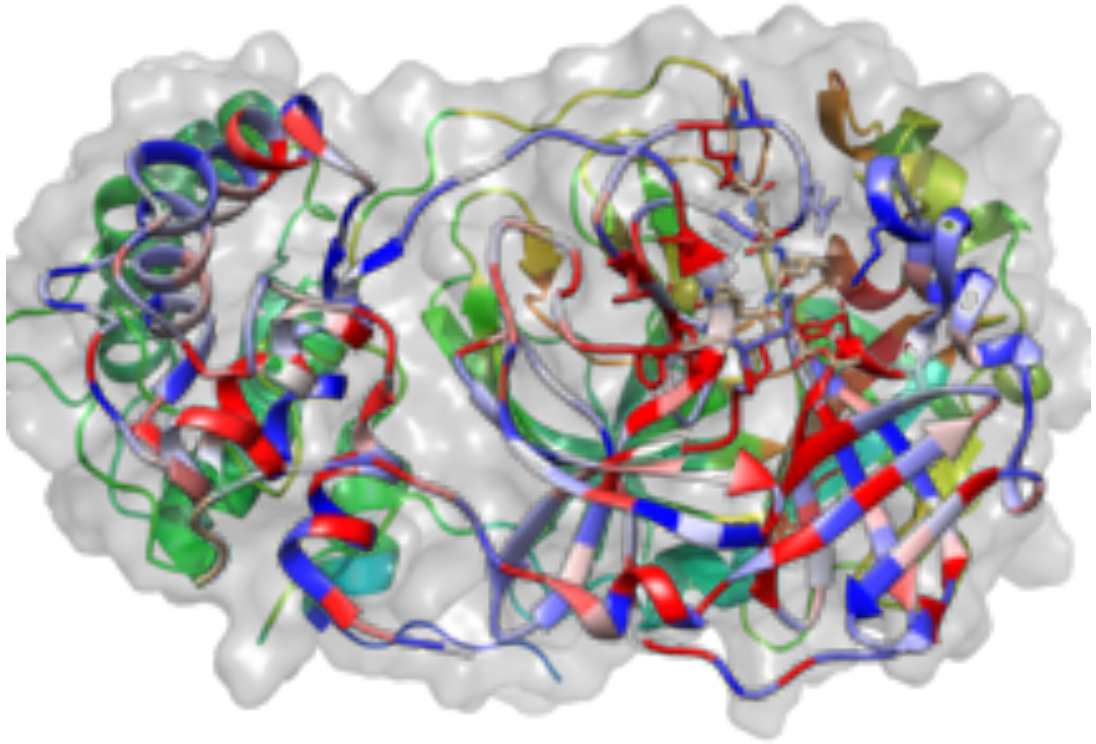


Workflow 2a

(Arvind Ramanathan, Shantenu Jha)

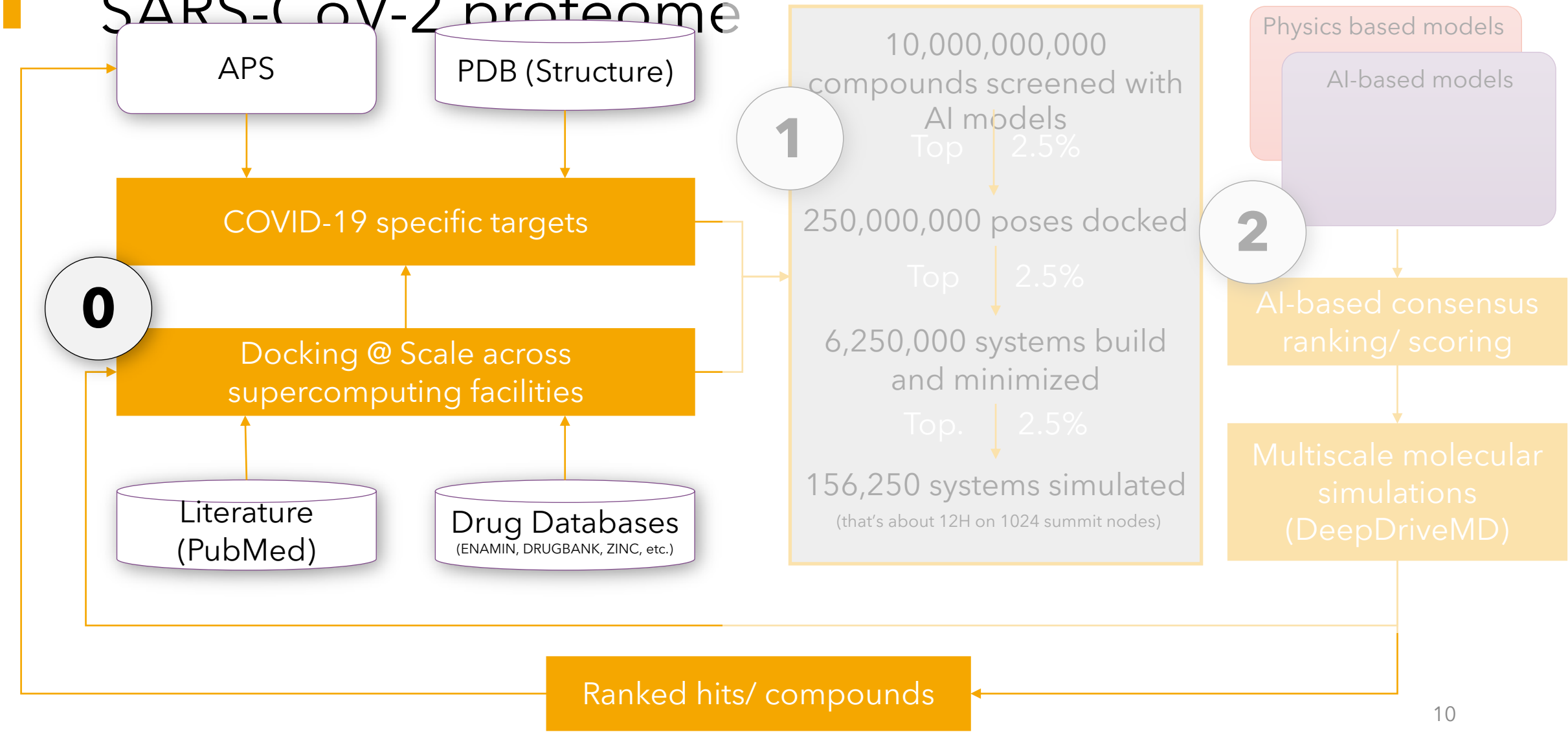
Nine target proteins from SARS-CoV-2

- Each with a distinct role in viral life-cycle:
  - Main protease (3CLPro)
  - Papain-like protease (PLPro)
  - Orf7a (replication)
  - RNA dependent polymerase
  - Spike protein
  - Nsp15
  - Nsp3 (ADRP)
  - Nsp9
  - Nsp10-Nsp16 complex



For each target, we have identified evolutionary coupling information leading to identification of novel binding sites. Collaboration w/ Sean McCorkle (BNL), Paul Adams (LBL), Peter Coveney (UCL, UK), Shozeb Haider (UCL, UK)

Using AI/ML to discover drugs that can target SARS-CoV-2 proteome



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

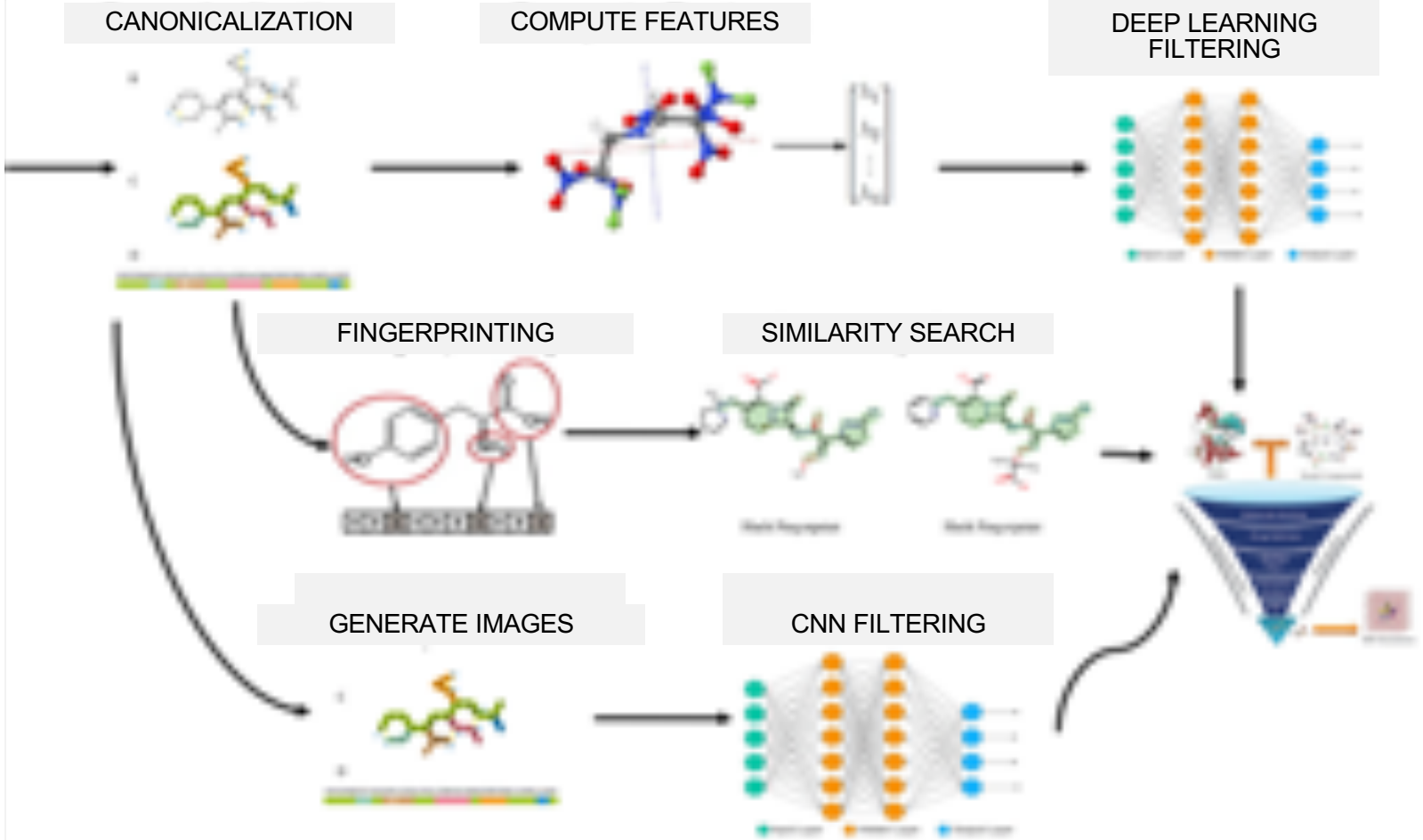
ZINC15

LINCS

SureChEMBL

PubChem

AND MORE

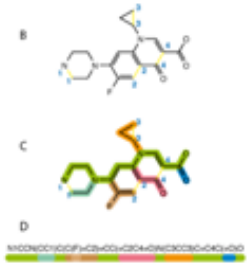


COMPUTING RESOURCES

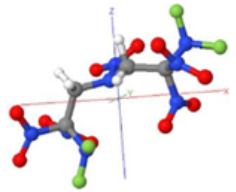
Yadu Babuji, Ben Blaiszik, Kyle Chard, Ryan Chard, Ian Foster, Logan Ward, Tom Brettin et al

First release of HPC-computed features for AI-based drug screening

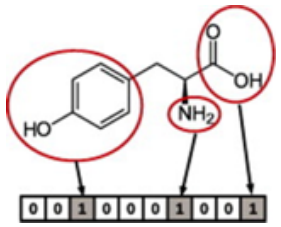
- **23 input datasets**, **4.2B molecules**, **60 TB** of molecular features and representations
- Data processing pipeline used ~2M core hours on ALCF Theta, TACC Frontera, OLCF Summit
  - Convert each molecule to a **canonical SMILES**
  - For each molecule, compute:
    - ~1800 2D and 3D **molecular descriptors** using Mordred
    - **Molecular fingerprints** encoding structure
    - **2D images** of the molecular structure
- Computed data provide **crucial input features to AI models** for predicting molecular properties such as docking scores and toxicity



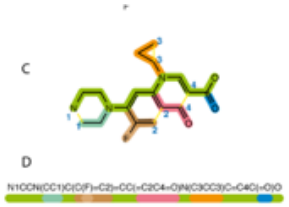
Canonical SMILES
23 CSV files with 4.2B molecules



Mordred Descriptors
420,130 CSV files, 48.70TB



Molecular Fingerprints
4,221 CSV files with base64 encoded fingerprints, 578.27GB



2D images
420,707 Pickle GZ files, 11.48 TB

https://2019-ncovgroup.github.io/data/

Ian Foster and team

Natural language processing: Dataset and Code

Manual Extraction:

- Engaged Argonne CELS admin staff to extract small molecules from key SARS/SARS-CoV-2/MERS papers
- Extracted >800 molecules, structures

Automated Extraction:

- Labeled relevant small molecules in their natural language context in CORD-19 papers
- Built named deep-learning entity recognition (NER) models to extract drug references from entire corpus (>24k full text articles)

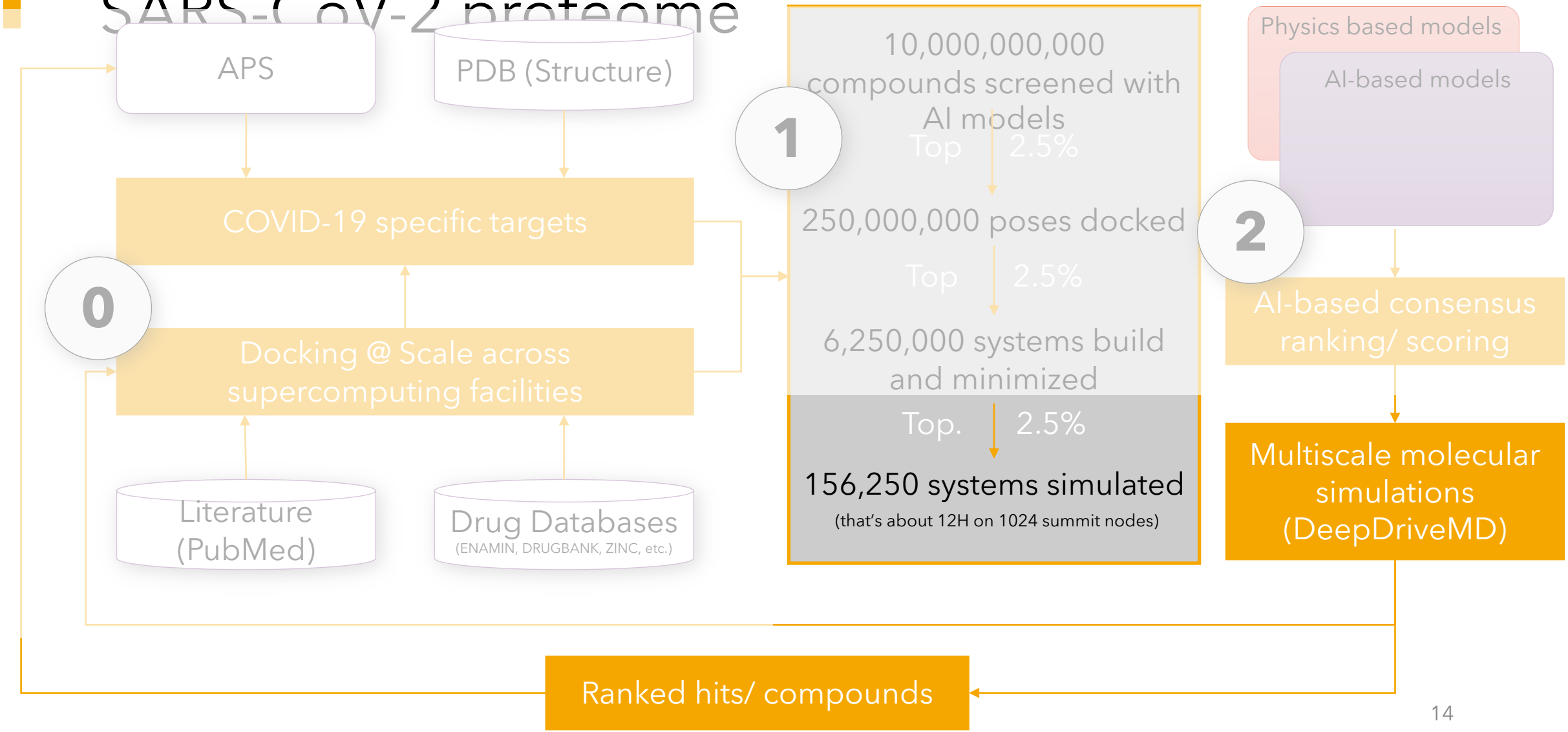
Lit - A Collection of Literature Extracted Small Molecules to Speed Identification of COVID-19 Therapeutics Dataset https://doi.org/10.26311/lit

Yadu Babuji, Ben Blaiszik, Kyle Chard, Ryan Chard, Ian Foster, India Gordon, Zhi Hong, Kasia Karbarz, Zhuozhao Li, Linda Novak, Susan Sarvey, Marcus Schwarting, Julie Smagacz, Logan Ward & Monica Orozco White

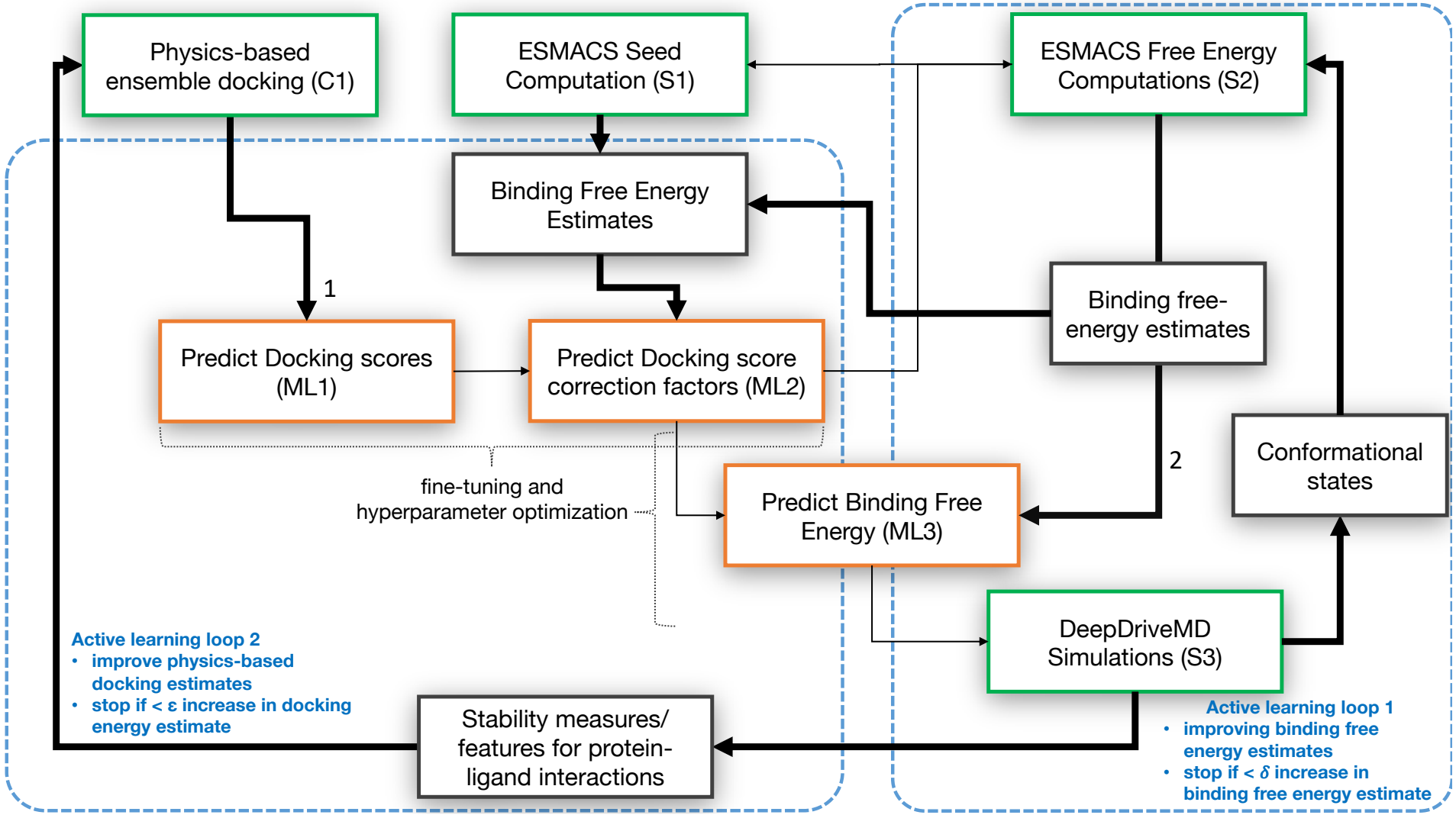
Dataset published 2020 via Materials Data Facility

Code, training data: https://github.com/globus-labs/covid-nlp

Using AI/ML to discover drugs that can target SARS-CoV-2 proteome

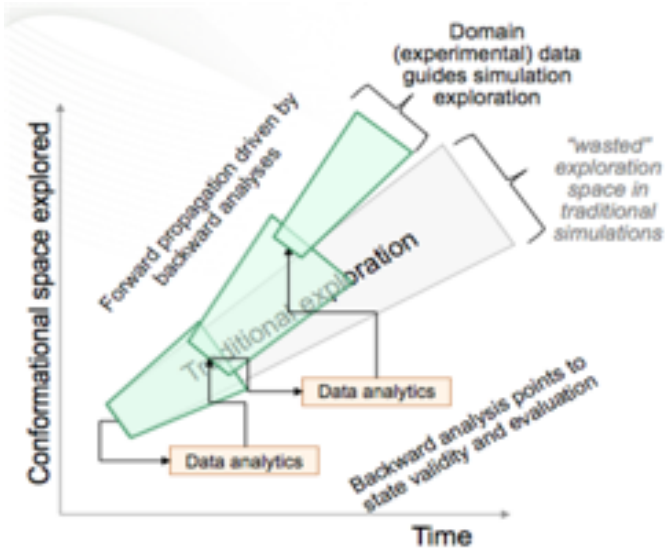


improving docking and finding better ligands that bind to SARS-COV-2 proteome

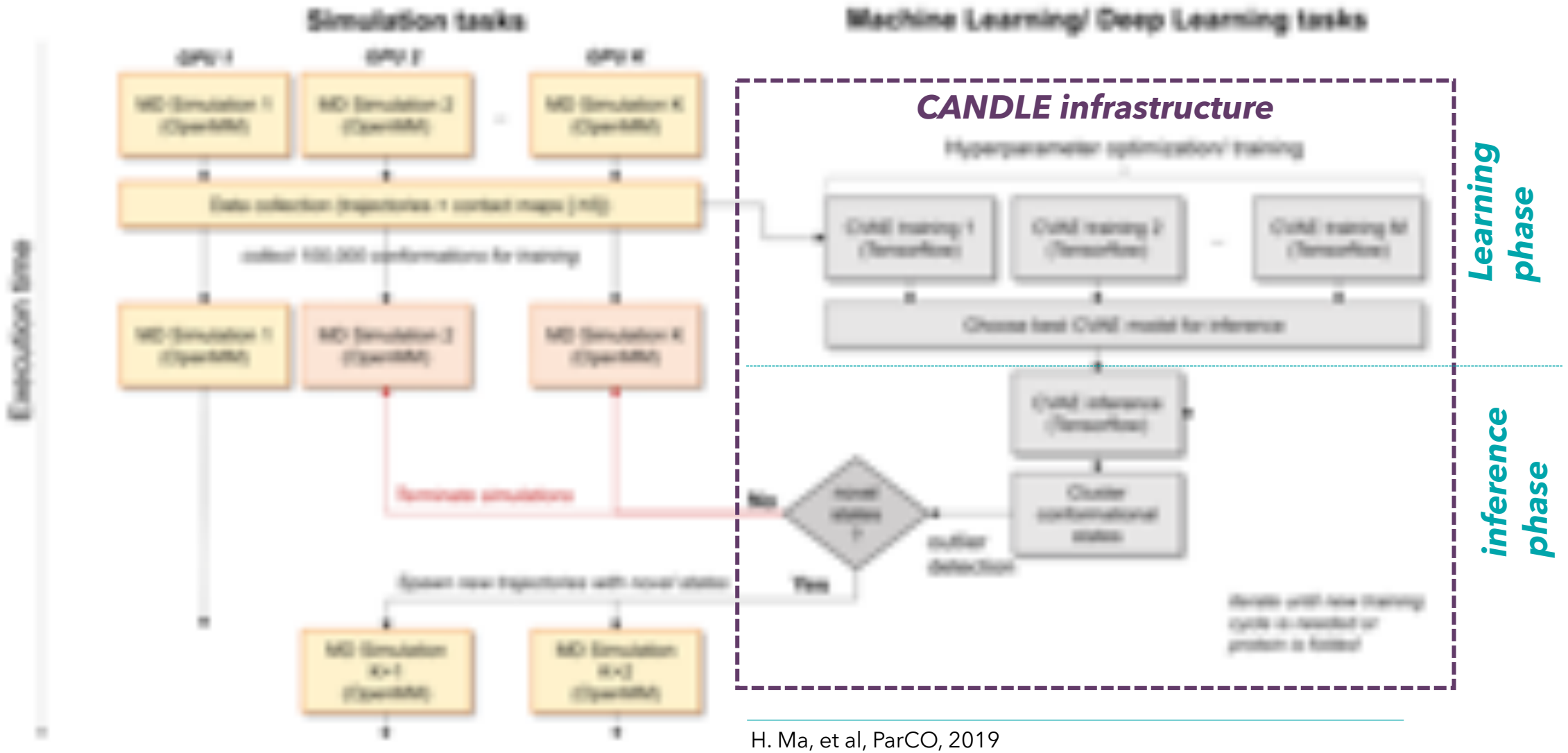


DeepDriveMD Overview: Interleave simulations and analytics adaptively for reducing computing overheads

- Generate ensemble of simulations in parallel as opposed to one realization of process
  - Statistical approach: **O(10^6 - 10^8)**!
- Ensemble methods necessary, not sufficient!
  - Adaptive Ensembles: **Intermediate data, determines next stages**
- Adaptivity: How, What
  - Internal data: Simulation generated data used to determine "optimal" adaptation



*Chodera, J.D., Noe, F., Curr. Opin. Struct. Biol. (2014)

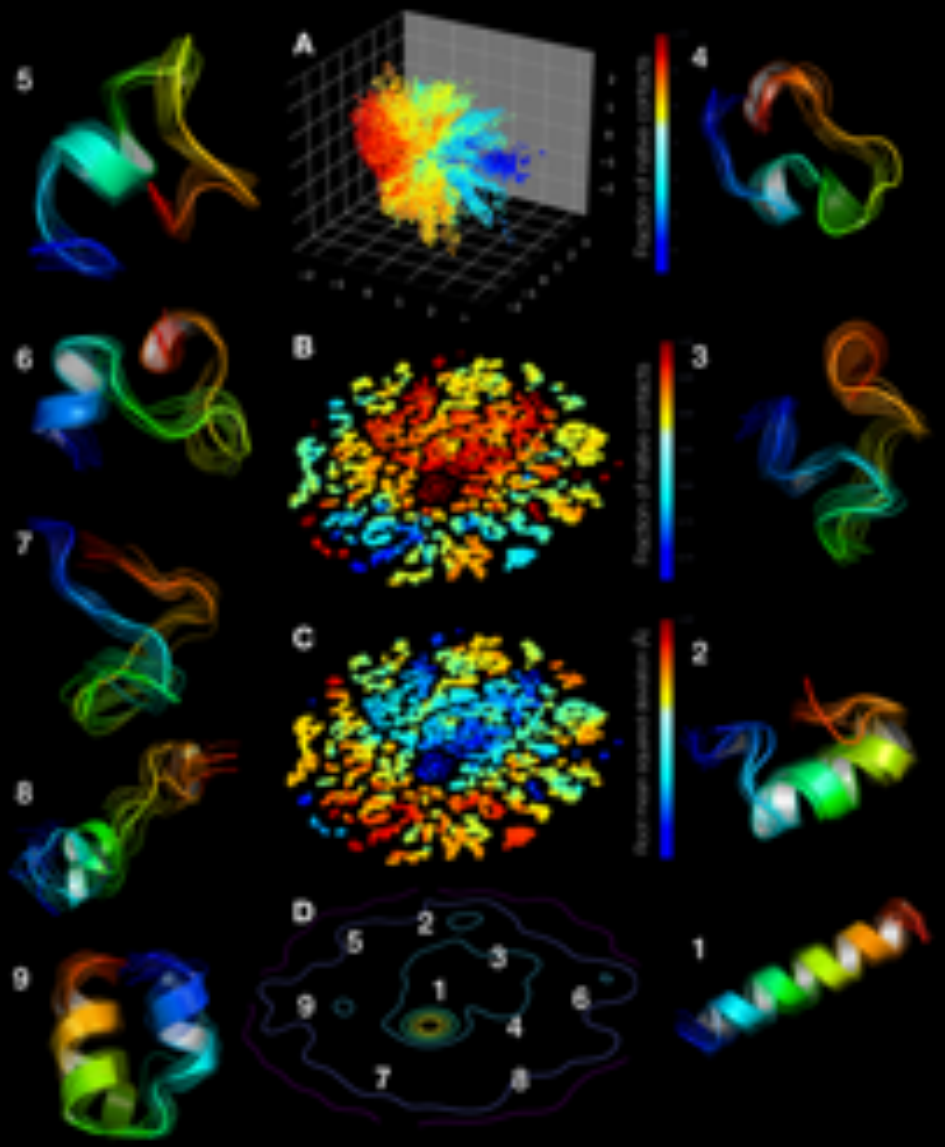


H. Ma, et al, ParCO, 2019
H. Ma, et al, Workshop on Deep Learning on Supercomputers, 2019

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

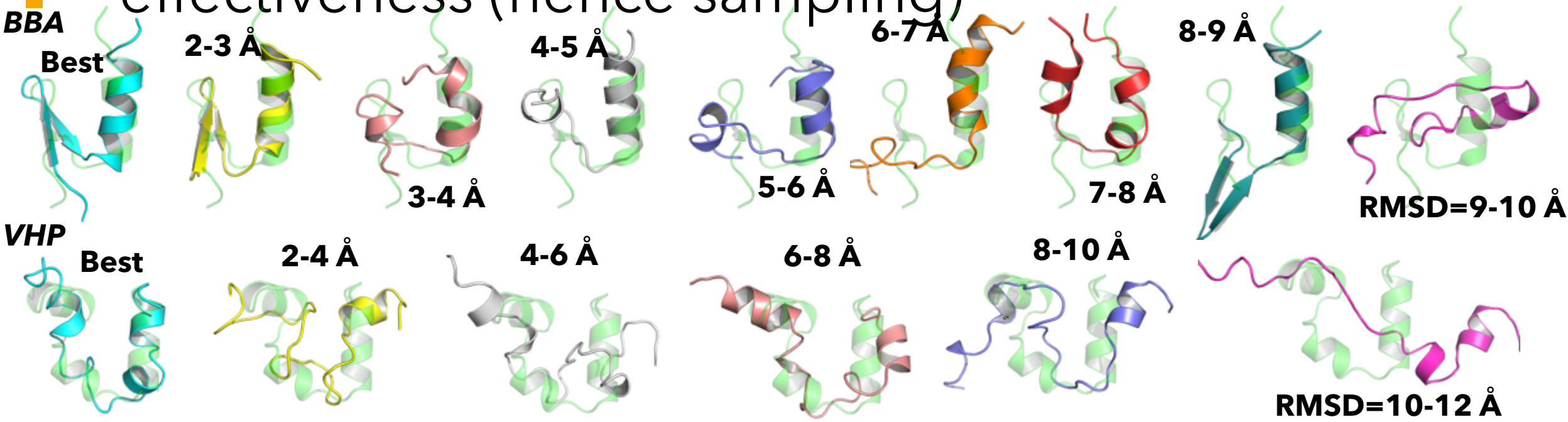
Deep clustering of protein folding simulations

- Convolutional Variational Auto Encoders (CVAE)
  - Low dimensional representations of states from simulation trajectories.
  - CVAE can transfer learned features to reveal novel states across simulations
- Integrating Bayesian learning to support uncertainty in sampling novel states
  - HPC Challenge (1): DL approaches to achieve near real-time training & prediction!
  - HPC Challenge (2): Hyperparameter optimization (while model is training)!



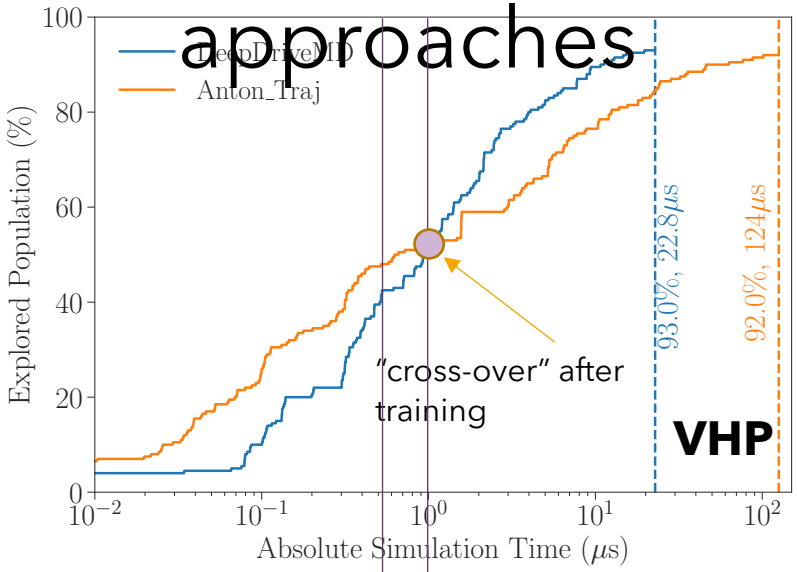
Bhowmik, D., et al, BMC Bioinformatics (2018).

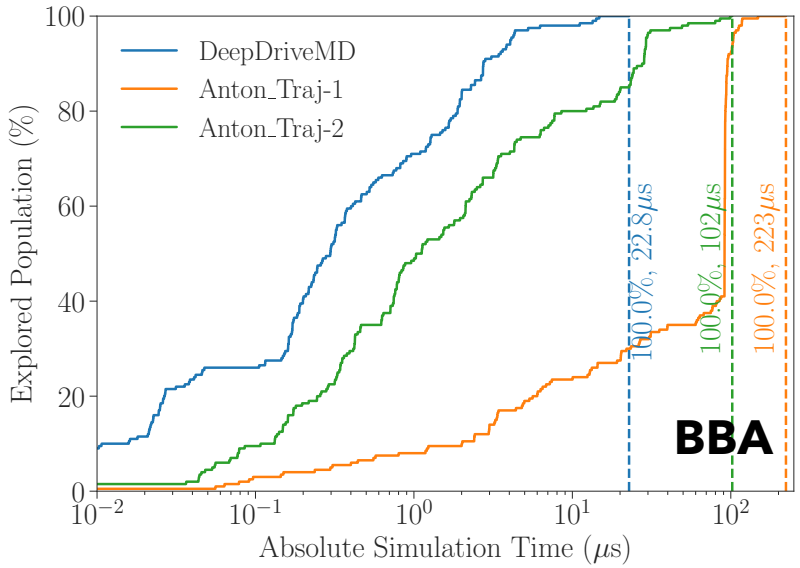
Larger number of simulations improves folding effectiveness (hence sampling)

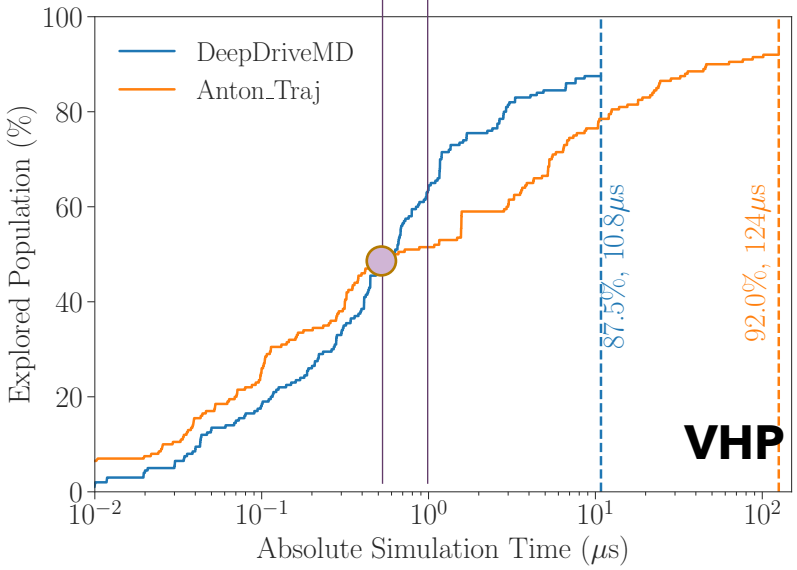


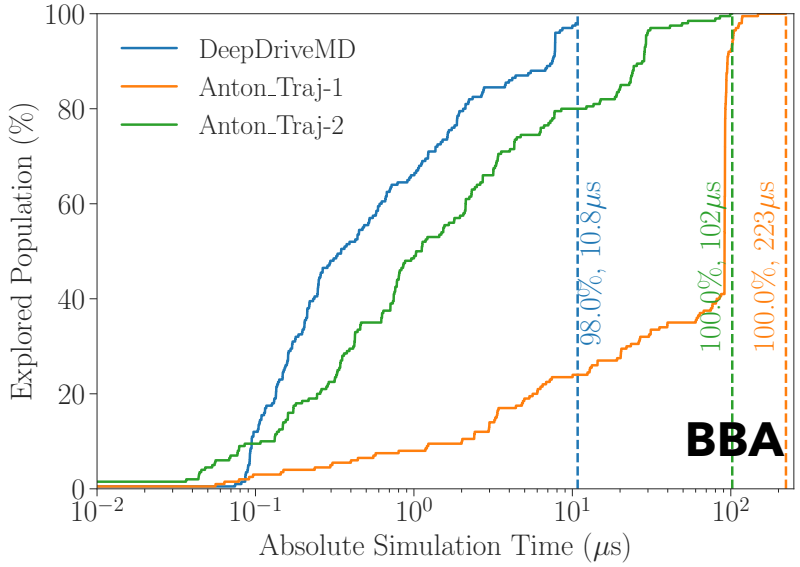
System	Total no. of simulations	Total simulation time (us)	First, subsequent simulations	Iterations	Min. RMSD
Fs-peptide	840	18.2	100, 10	7	0.29
BBA (FSD-EY)	1200	22.8	100, 10	10	1.8
VHP	1200	22.8	100, 10	10	3.83

DeepDriveMD shows at least an order of magnitude efficient sampling compared to traditional approaches







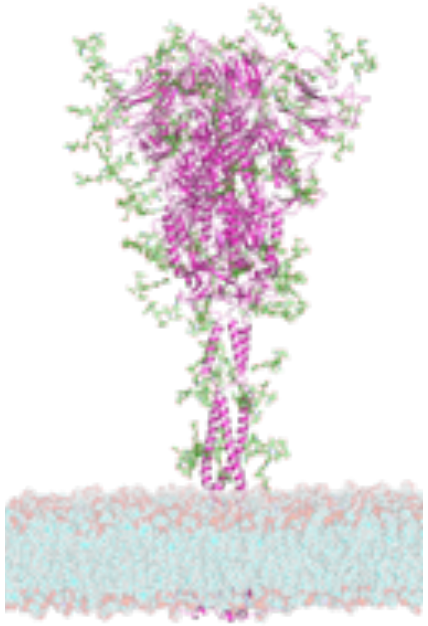


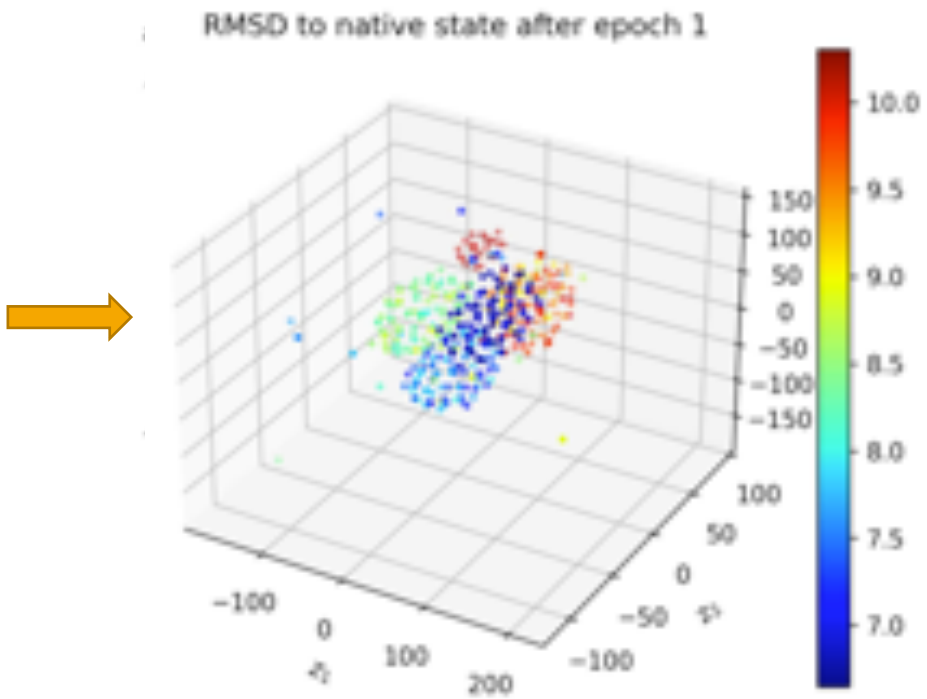
- **including the data from the "learning phase"**: one order of magnitude improvement in sampling:
  - Distinct "cross-over" after training where sampling is accelerated significantly after learning/ estimating the conformational states

Reference trajectories are from D.E. Shaw (Science, 2011)

- ***not including the data from "learning phase":*** At least two orders of magnitude improvement in sampling:
  - If Anton trajectories take O(microsecond) to sample a particular state, DeepDriveMD samples it in O(100 ns)
- For BBA, 98% sampled states are observed within 10 microseconds!

Using fully convolutional VAE to identify conformational states in Spike protein simulations

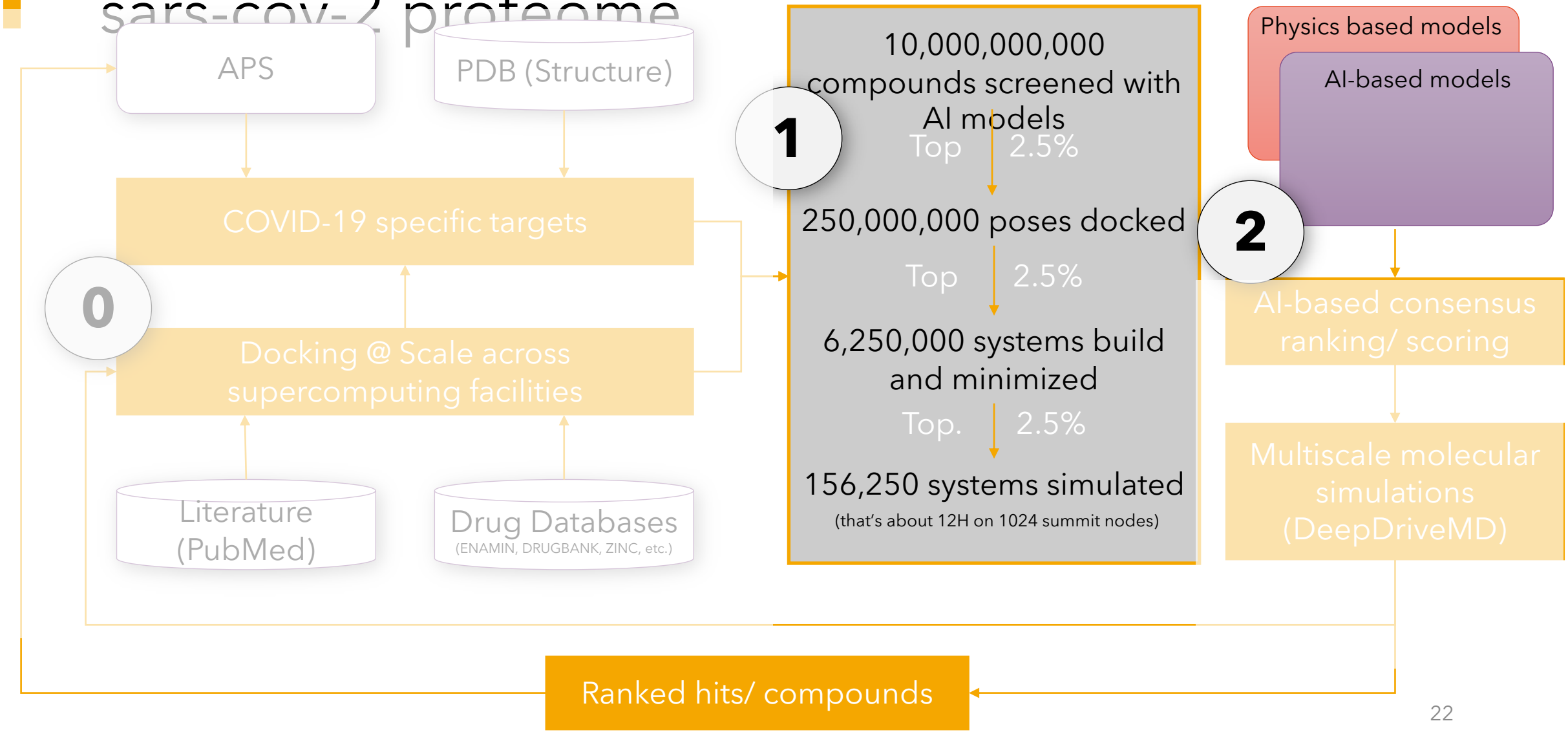




No. GPUs (V100)	Memory	Time per batch (8)
1	20213/32510 MiB	7.561
2	9947/32510 MiB (Encoder) 12987/32510 MiB (Decoder)	7.481

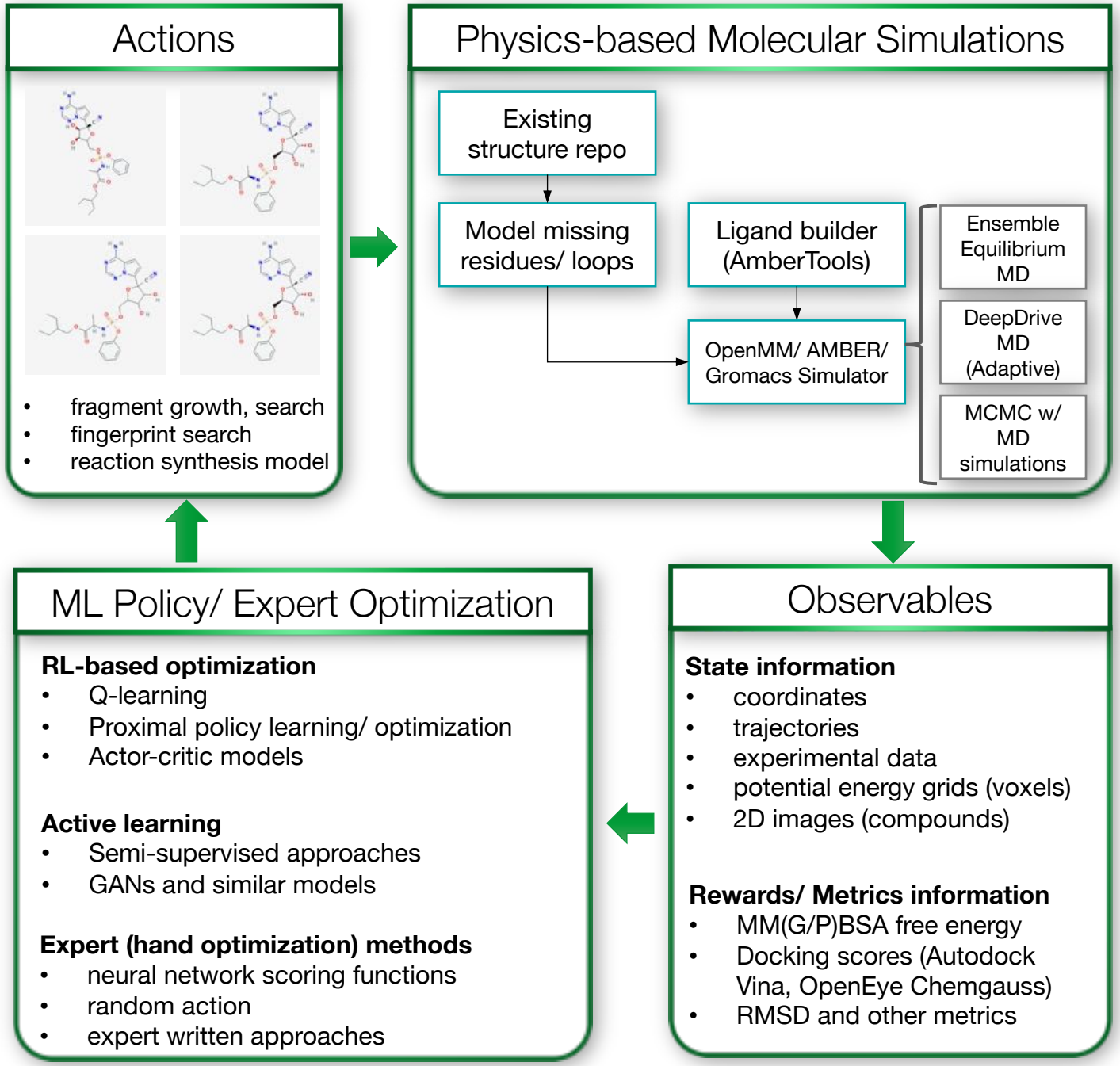
- Modification of the VAE architecture to accommodate larger systems (E.g. Spike protein – 1.5 million atoms)
- Model parallel example:
  - encoder and decoder on individual GPUs
  - implemented with Pytorch
- Can improve performance with layer-wise adaptive rescaling
- Joint work with Alex Brace (Argonne intern), Abe Stern (NVIDIA), Thorsten Kurth (NVIDIA), Anda Trifan (CSGF), Rommie Amaro (UCSD), Carlos Simmerling (Stony Brook University)
- Implementation with Cerebras (CS1) and Sambanova: Harry Yoo (Argonne), Jessica Liu (Cerebras), Vishal Subbiah (Cerebras), Arvind Sujeeth (Sambanova), Venkat Vishwanath (ALCF) and Murali Emani (ALCF)

using AI/ML to discover drugs that can target sars-cov-2 proteome



Reinforcement learning driven MD

- Motivation: physics-based models are guided by an action space determined by AI
- Can we expand the compound space explored using RL?
- For SARS-CoV-2 proteome:
  - relevant for specific mutations compared to other CoV proteins
  - suggest repurposing based on shape/structural complementarity

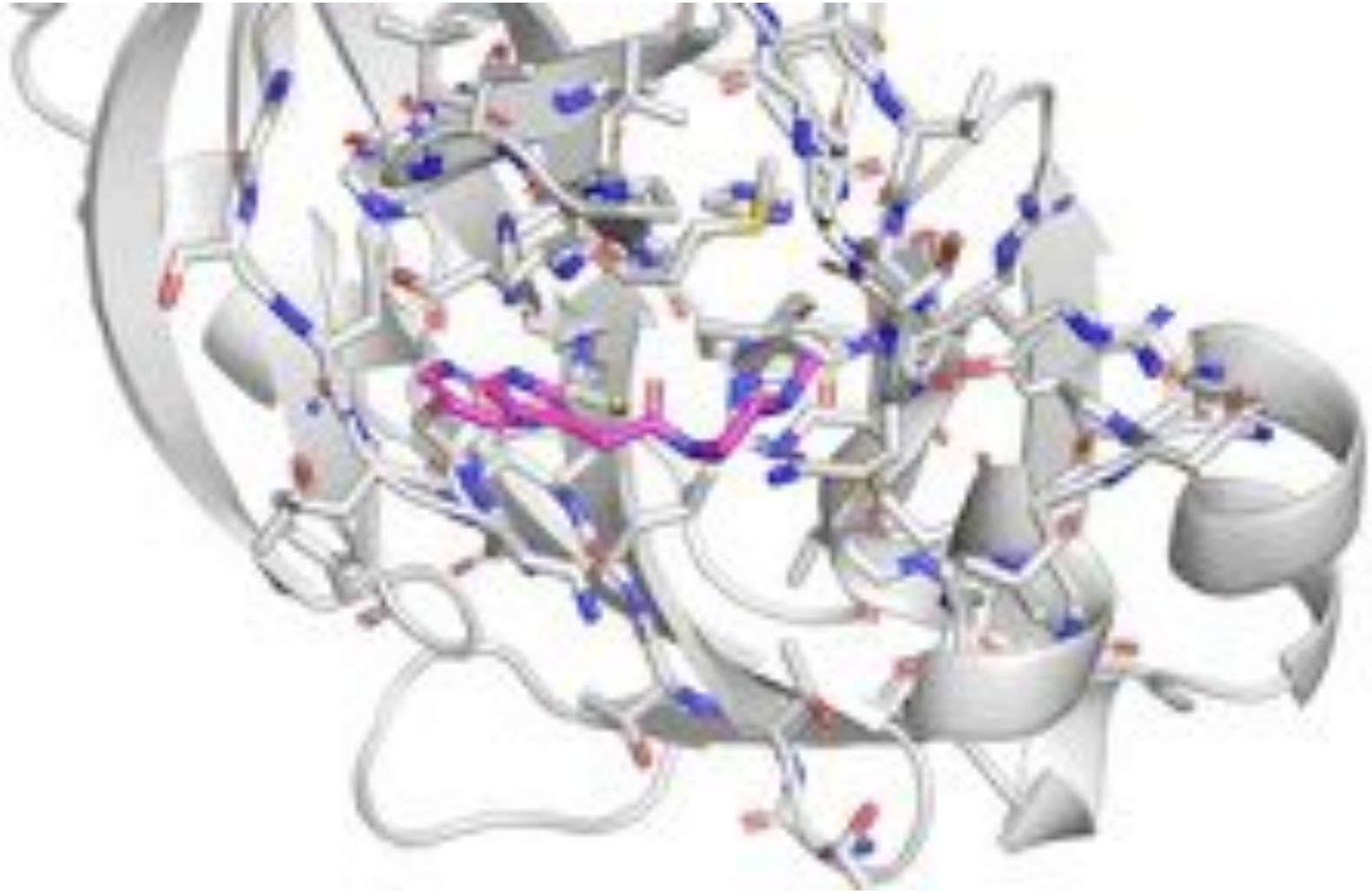


Austin Clyde, Arvind Ramanathan

Using expert guided fragment growth for JAK2 kinase

- **Data**
  - Set of known inhibitors with structures and experimental binding affinity
  - Set of known decoys
- **Tests**
  - Show from known inhibitors we can estimate experimental BF
  - Sample known pocket conformers including water positions
  - **Show when starting with a decoy we predict it's a decoy and optimize it towards a known inhibitor**

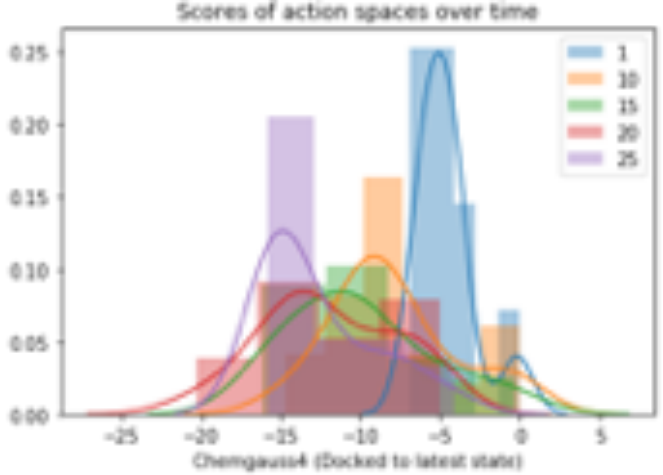
Fragment growth with an expert docking policy for SARS-CoV-2 3CLPRo

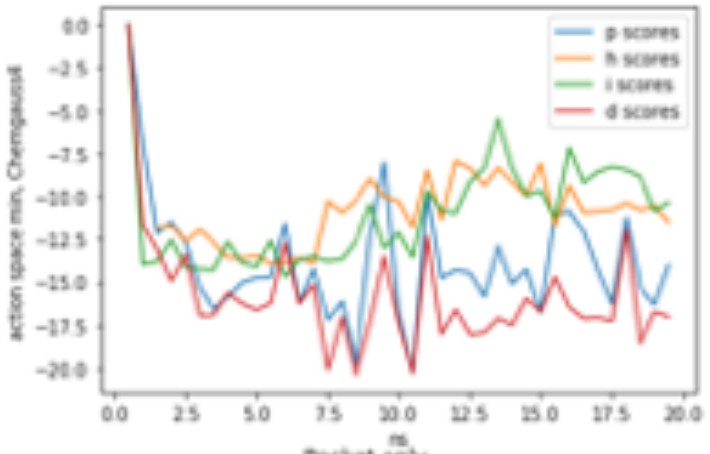


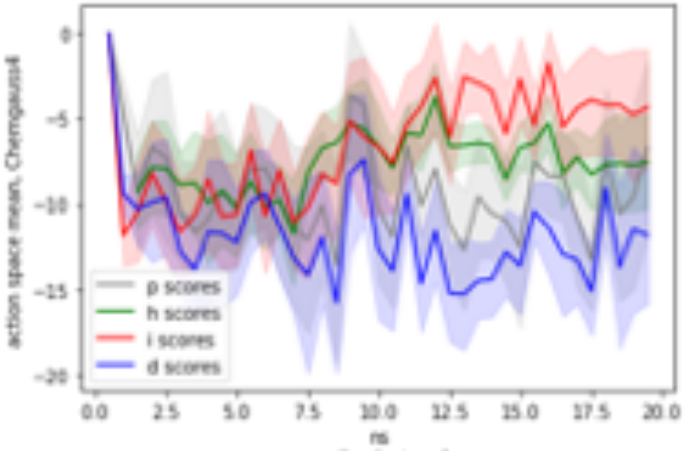
Action space

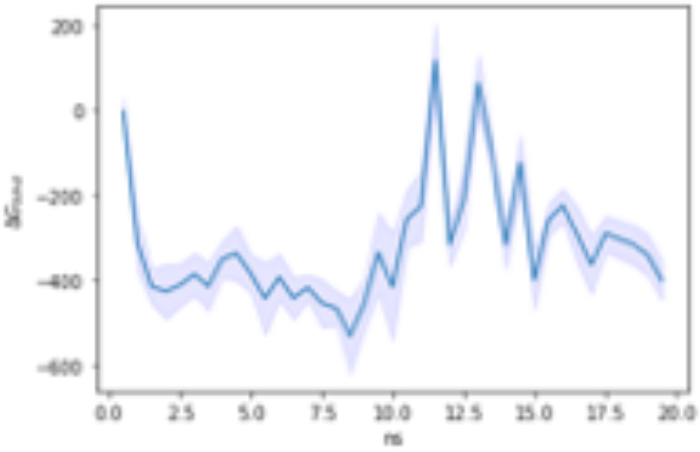
MMGBSA

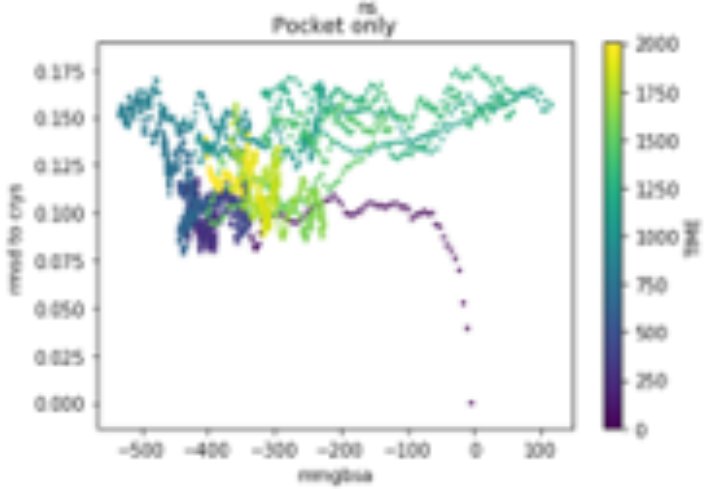
Sampling

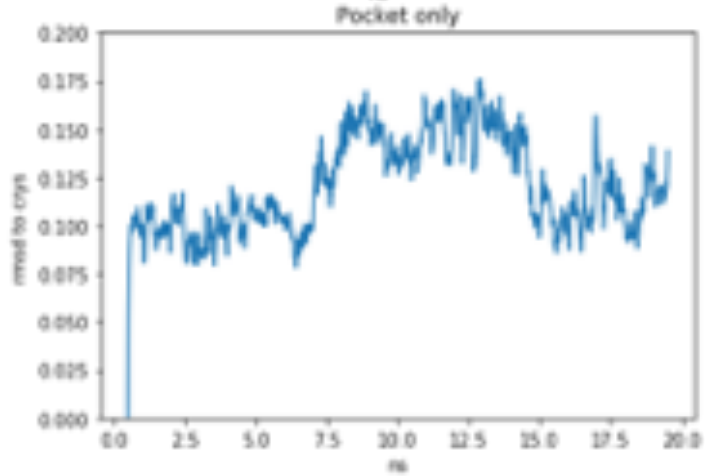


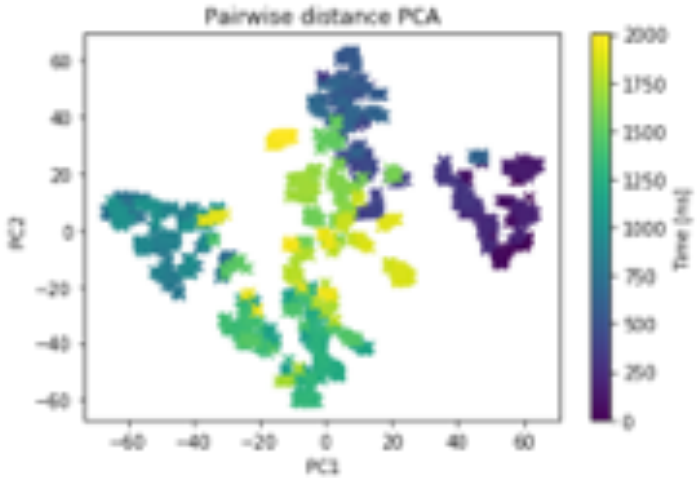


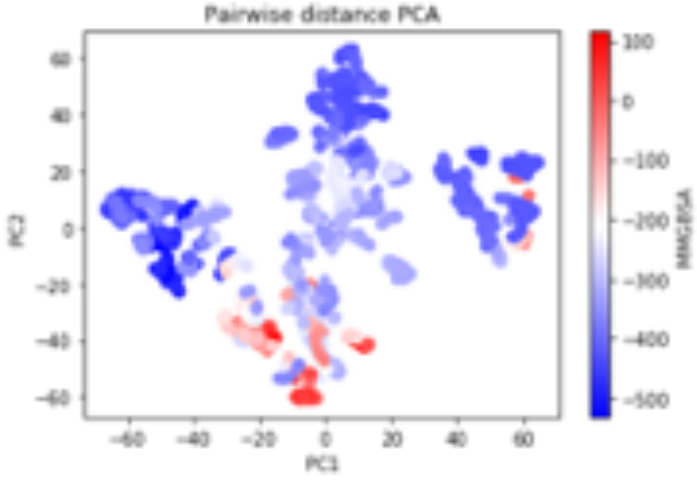


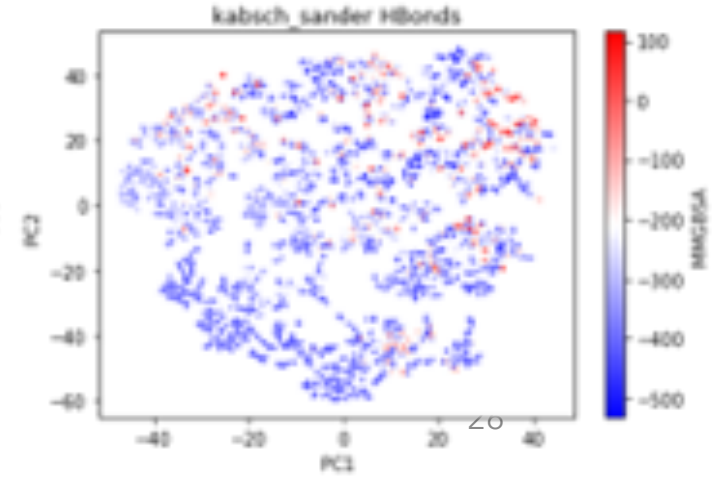






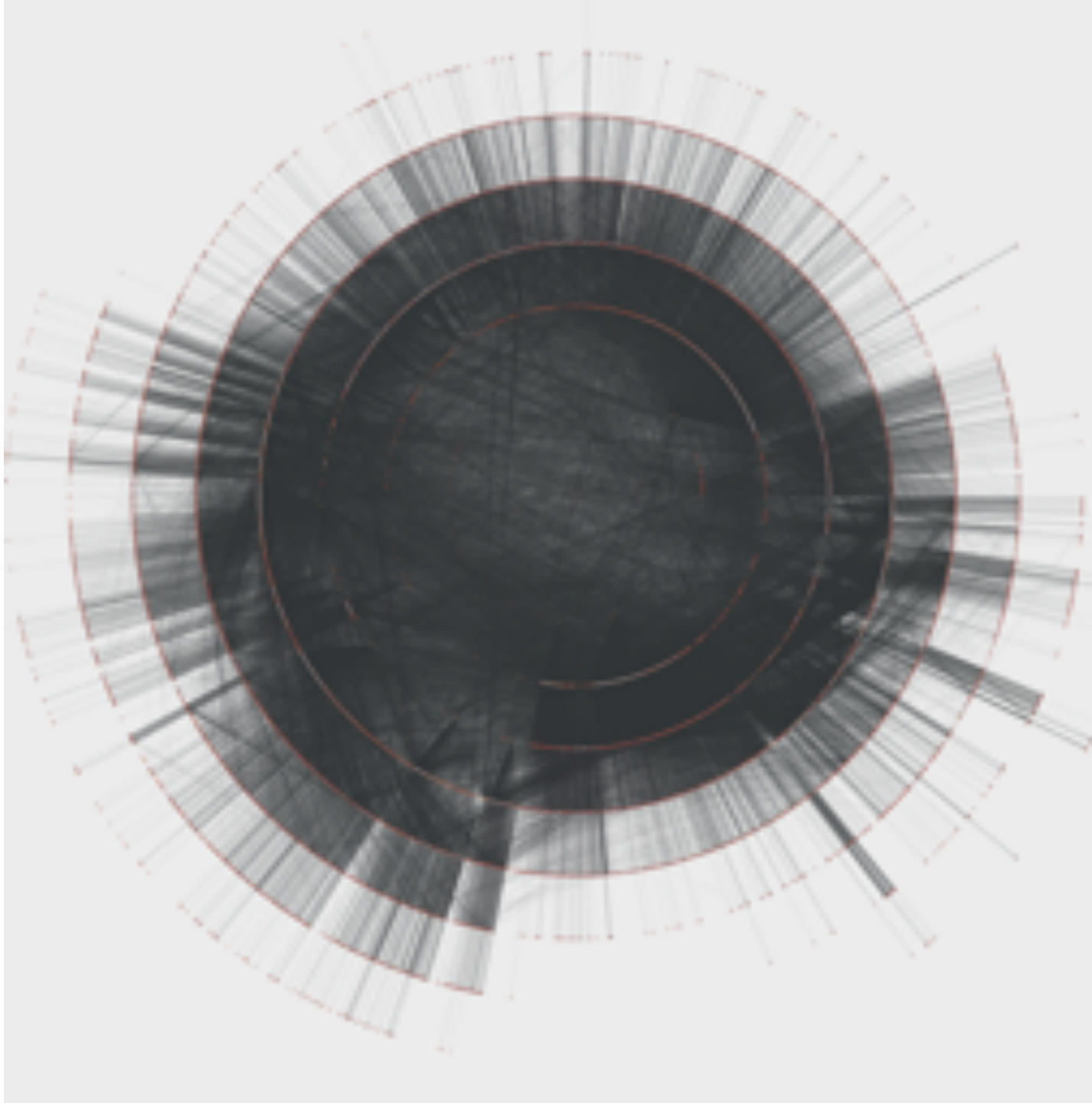






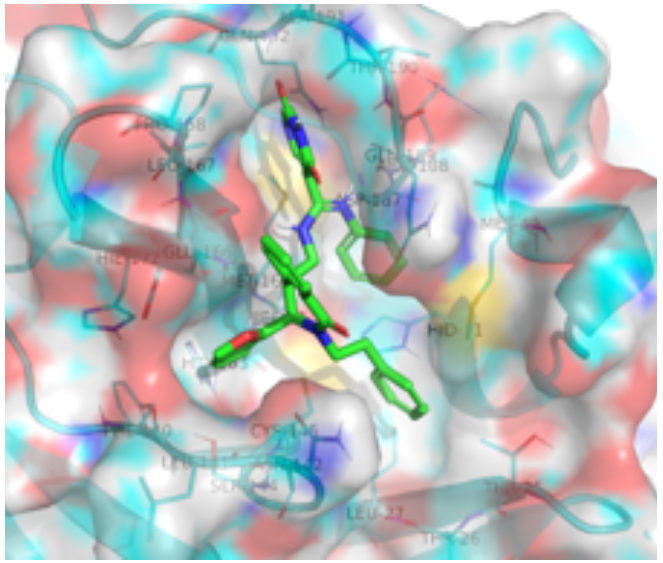
Large graph inference for optimizing ligands from fragments

- ~6 million compounds (MCULE) in a radial layout:
  - nodes → fragments or ligands
  - edges → connecting fragments
- Nodes at leaves give rise to tight binders
- PageRank and random walk theory approaches lead to dynamic pruning to quickly identify chemical spaces that are relevant for binding



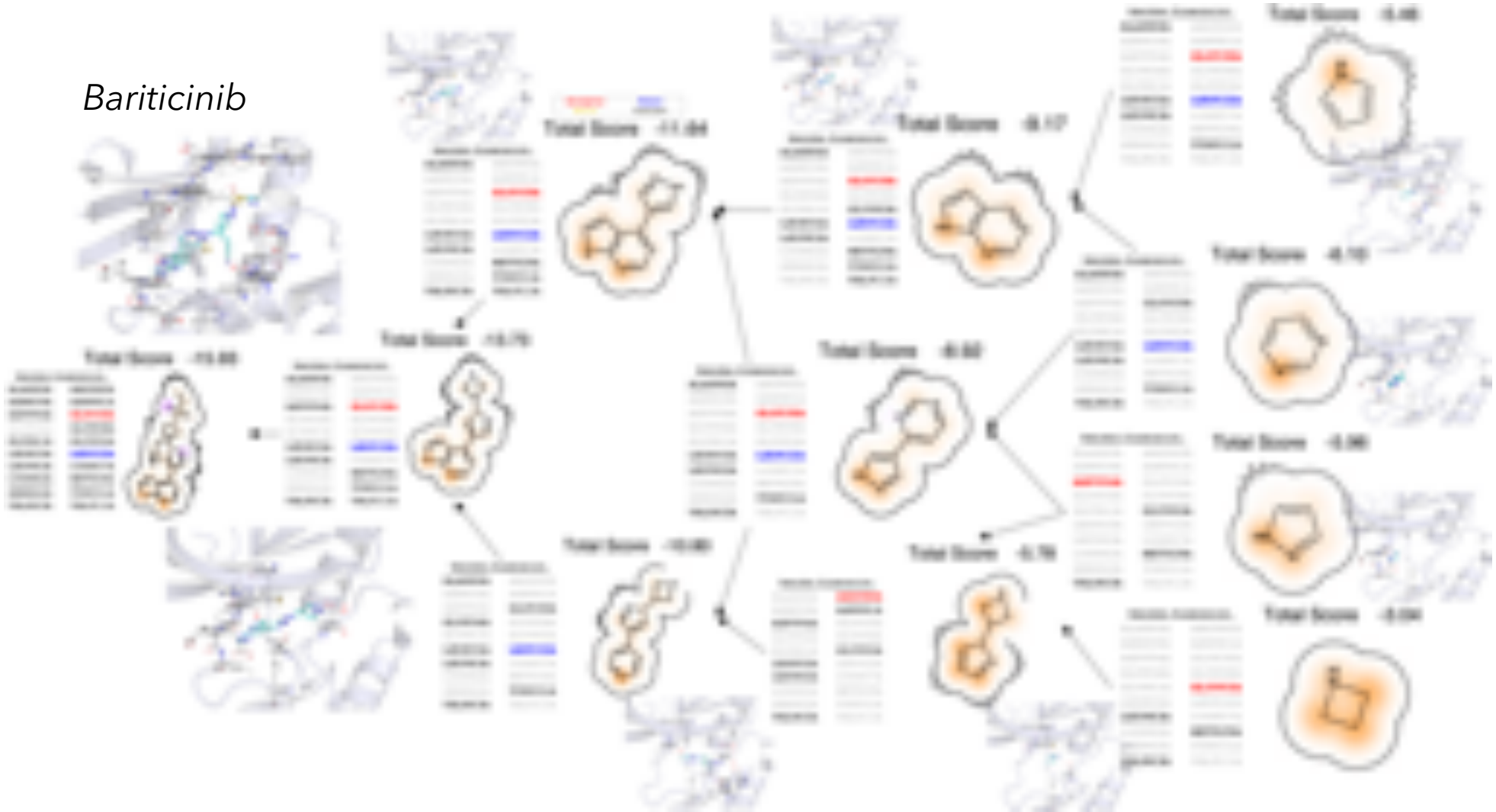
Page-rank based top ranked compounds lead to strong binding inhibitors for 3CLPro

Molecular Structure (Name)	PageRank Value (Rank)	Docking Score (Chemgauss4)
(No Name)	[illegible] (1)	[illegible]
(No Name)	[illegible] (2)	[illegible]
(No Name)	[illegible] (3)	[illegible]
(6-Thioinosine, 2-methylene-)	[illegible] (6,368,237)	-3.97
 (1,2,4-oxadiazole-5(4H)-thione)	[illegible] (6,368,218)	-4.98
 (methylene hexacyclocinone)	[illegible] (6,368,219)	-5.88

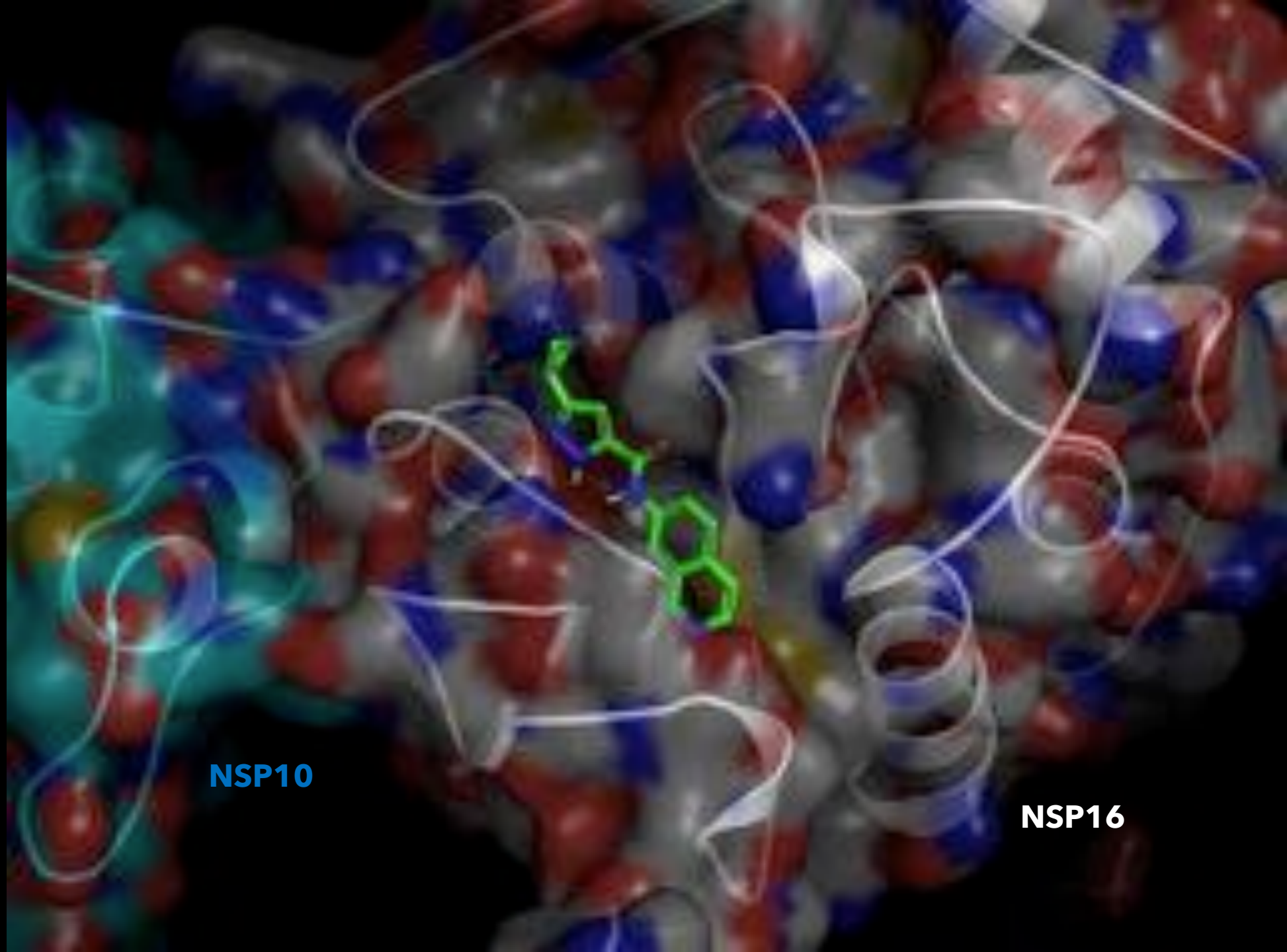


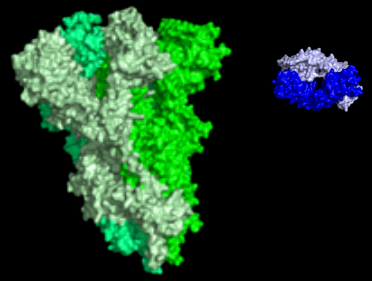
- Binding site complementarity covers all (P1-P4) sites of 3CLPro
- Stable hydrogen bond interactions with Gln192 and hydrophobic interactions

Graph traversal-based inference lead to optimal design of inhibitors for JAK2-kinase



- Covering multiple binding sites across NSP10-NSP16
- Challenging since the binding site including multiple pockets spread across the interface
- Strategy of using multiple pockets provides insights into intrinsic flexibility





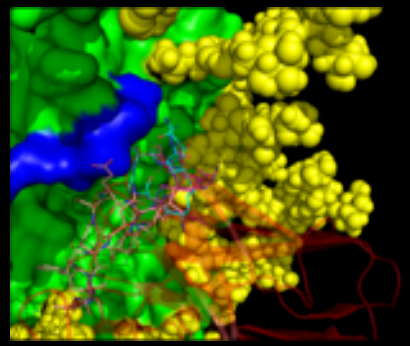
SABDB

PDB

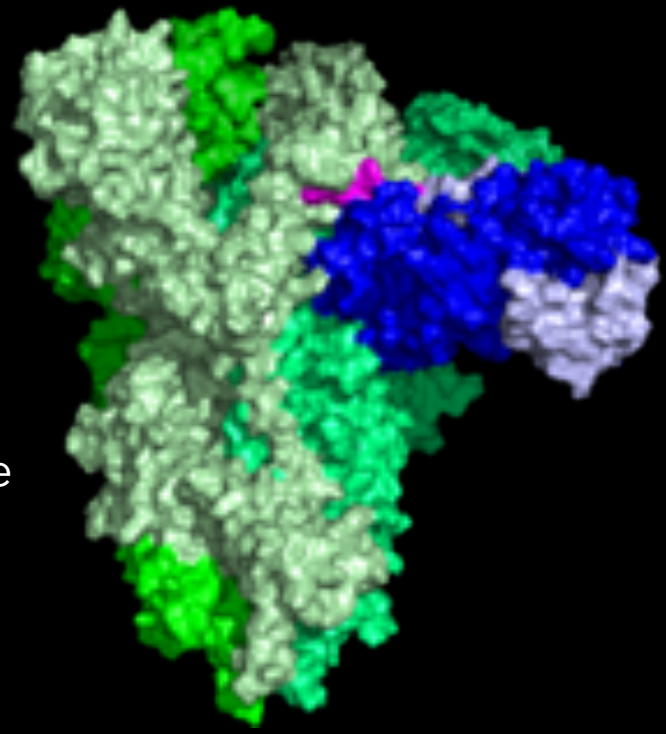
Rosetta Commons

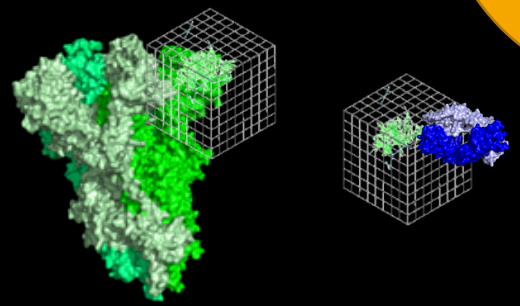
1. Seed antigen/antibody structures

2. Sample Targeted pool of antibody CDRs

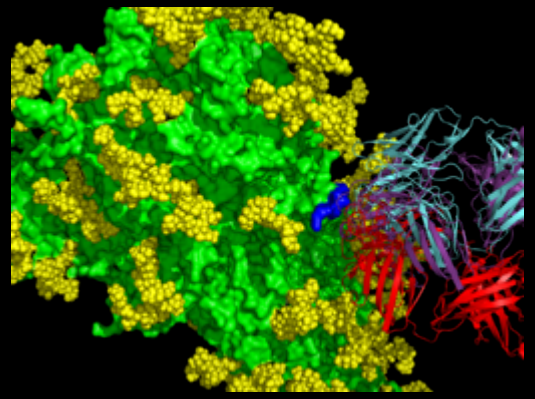


5. After X cycles, output new antibody structure/sequence



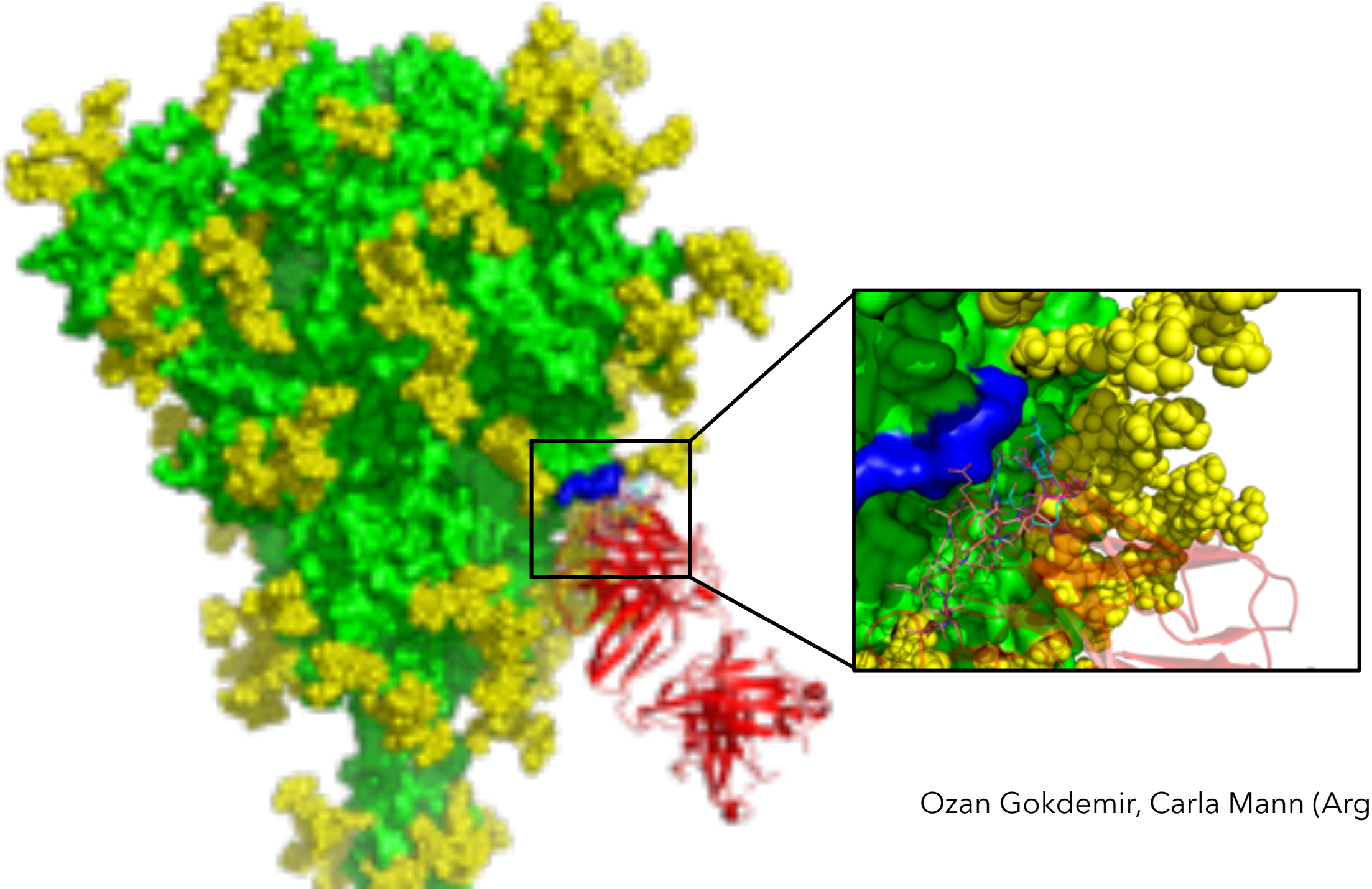


3. Reinforcement learning driven online optimization with physics-based mechanistic models



4. Sequence-based generative models optimize probability distributions for CDR binding

Ozan Gokdemir, Carla Mann (Argonne)



Ozan Gokdemir, Carla Mann (Argonne)

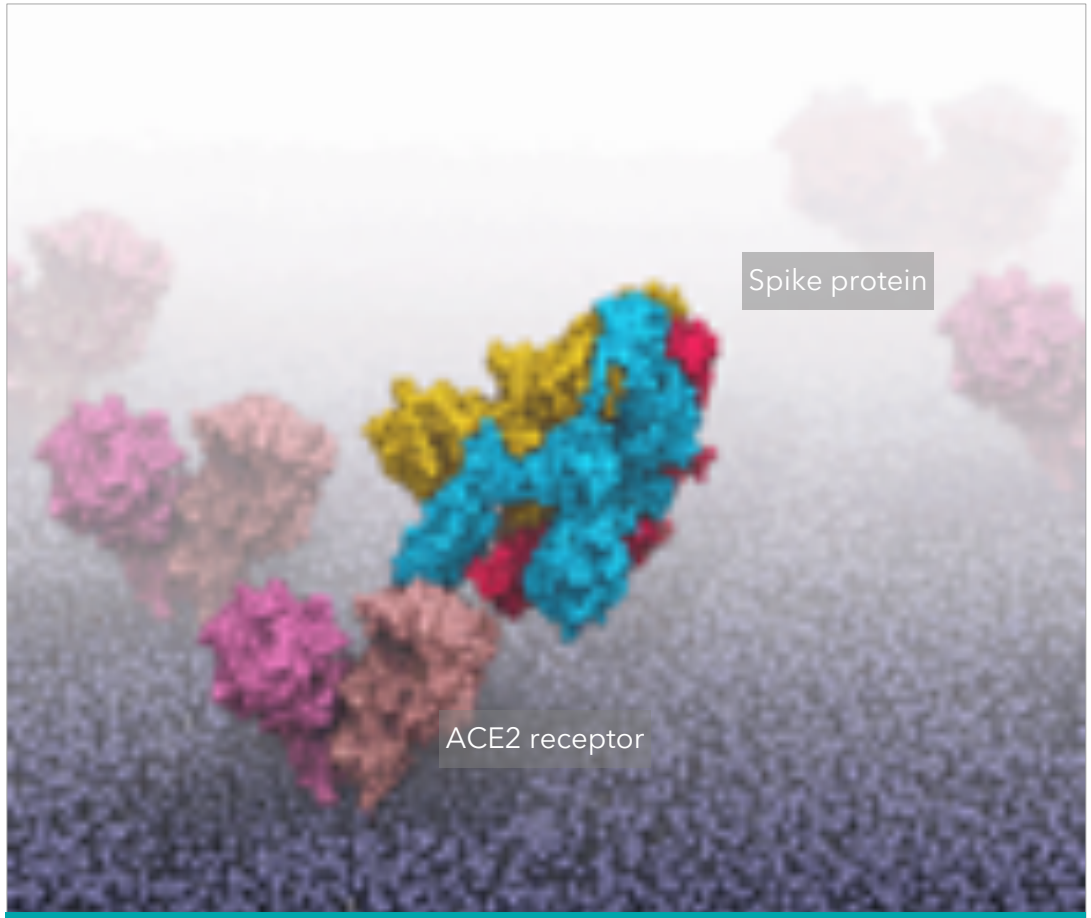
Future work / Outlook

- Significant scientific discoveries have been accelerated through AI-driven simulations and molecular design strategies:
  - nearly 30 compounds have been initially tested; additional assays and refinements with larger compound libraries are underway
  - DeepDriveMD and RLMM constitute new ways to integrate physics-based models with AI to accelerate time-to-solution

- Computational needs are much larger compared to traditional simulation-based approaches alone:
  - better utilization of resources
  - heterogeneous architectures provide new ways to exploit HPC + AI
  - need workflow management systems + runtime (for data transfer and I/O) to manage large jobs

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)



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

THANK YOU!

(RAMANATHANA@ANL.GOV)

Q&A

To pose a question, you can write your question in the “Questions” tab

Thank you for participating!

…don't forget to fill in our feedback questionnaire…

Visit the CompBioMed website (www.compbiomed.eu/training)
for a full recording of this and other webinars,
to download the slides
and to keep updated on our upcoming trainings.
For any other questions:
ramanathana@anl.gov

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823712

The series is run in collaboration with: VPH Institute Building the Virtual Physiological Human