Scalable HPC & AI Infrastructure for COVID19 Therapeutics

Shantenu Jha

Computation and Data-Driven Discovery, Brookhaven National Laboratory
RADICAL Lab, Rutgers University

Advancing Medical Care through Discovery in the Physical Sciences

https://indico.jlab.org/event/447/

BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

@BrookhavenLab

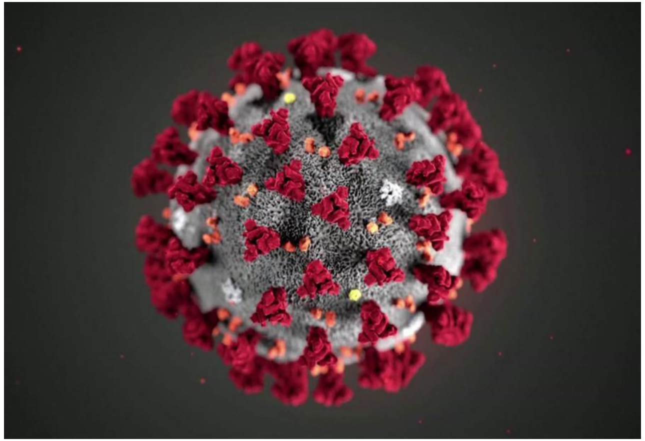
National Virtual Biotechnology Lab (NVBL)

- National Virtual Biotechnology Lab (NVBL)
    - https://science.osti.gov/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 to facilitate new antigen and antibody testing

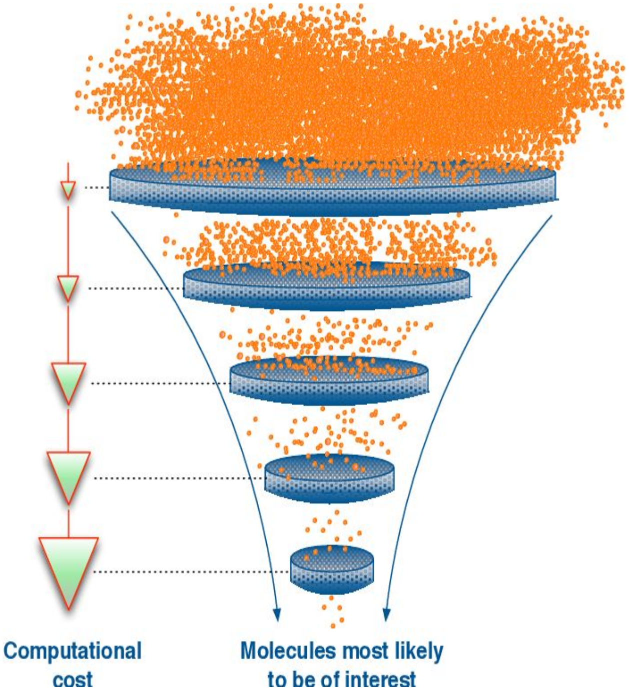
NVBL given US Secretary of Energy Honour Award (2021)

BROOKHAVEN
NATIONAL LABORATORY



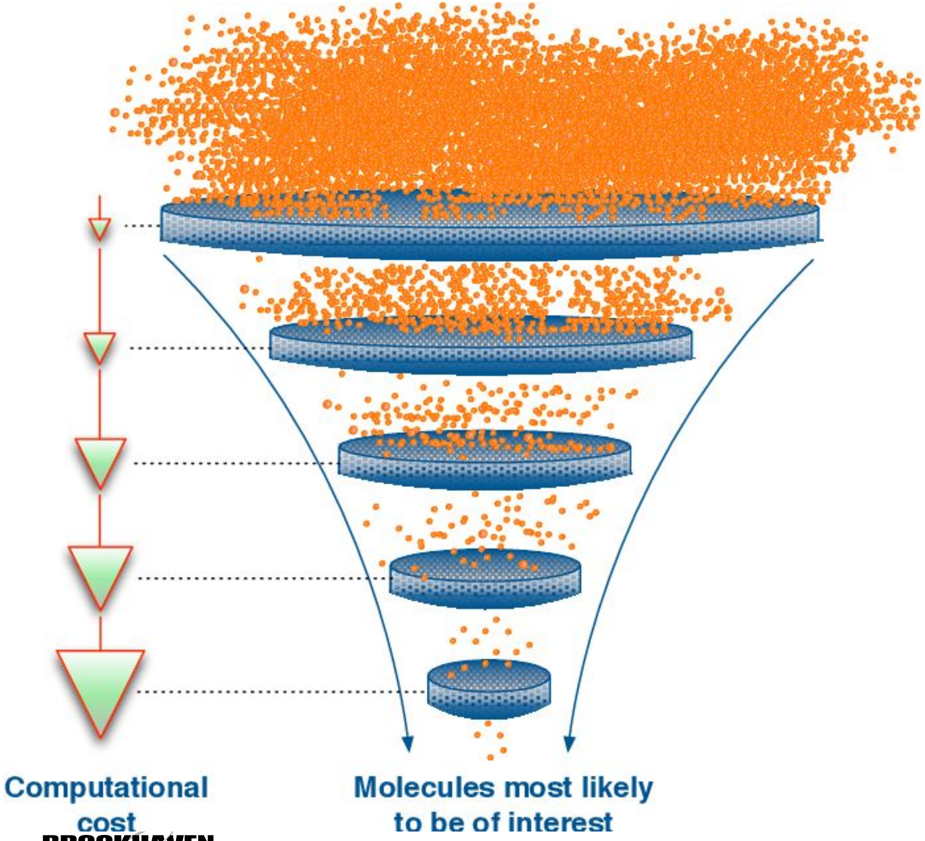
Overview

- Drug Discovery & Design is a complex, expensive
  - O(10) years; O(10^9) $; O($10^{68}$) candidates
- Scale-Accuracy trade-off:
  - AI-driven HPC methods 1000 x *effective performance* of traditional HPC simulations
- AI-driven HPC methods will be formulated as heterogeneous and adaptive workflows:
  - Systems software evolve in response



Ref. Aspuru-Guzik

High-Throughput Virtual Scaling



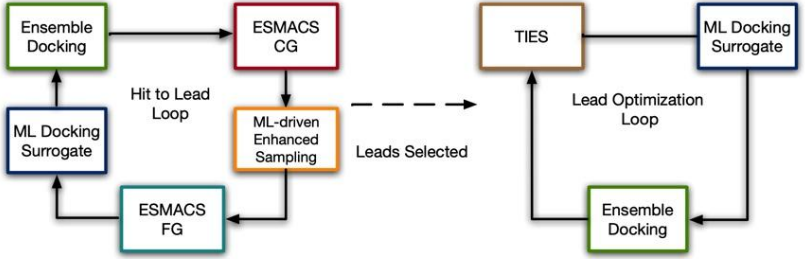
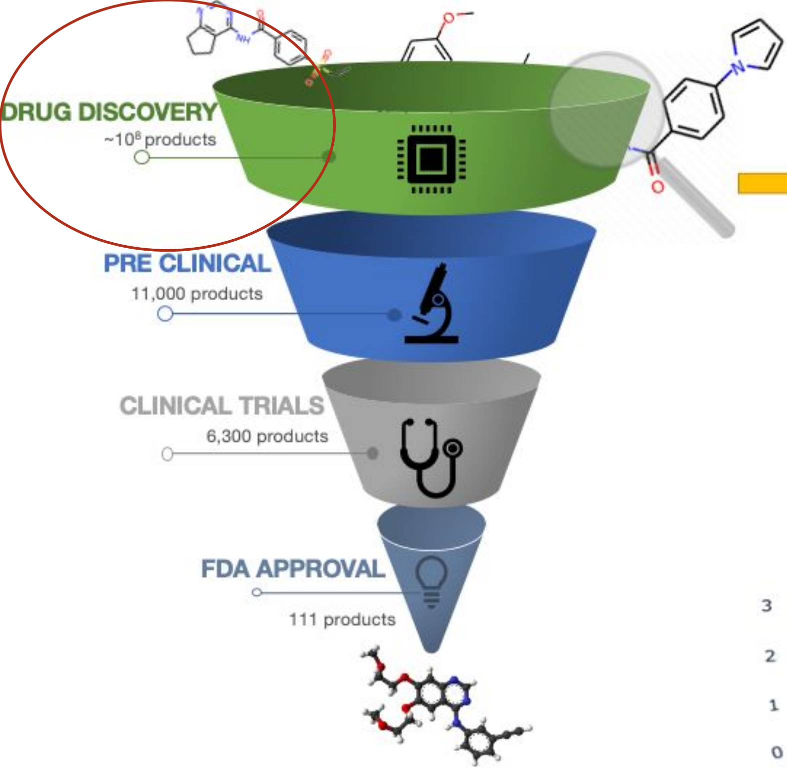


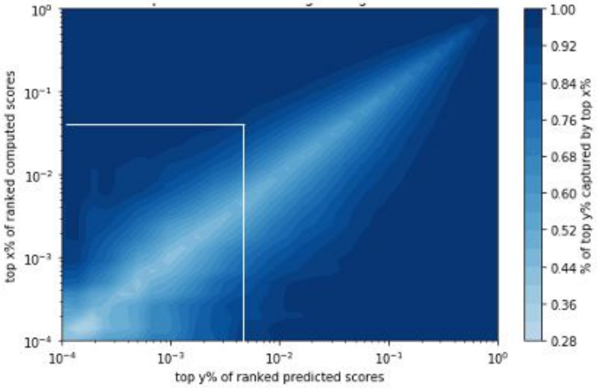
Figure 1: The computational campaign to advance COVID-19 therapeutics has two coupled loops: drug candidates go through four stages in the Hit-to-Lead loop; a small set of drugs are selected for the Lead Optimization loop. The following methods and protocols are implemented as distinct workflows (WF): Ensemble Docking (WF1), ML-driven Enhanced Sampling (WF2), both coarse-grained (CG) and fine-grained (FG) ESMACS (WF3), and TIES (WF4).

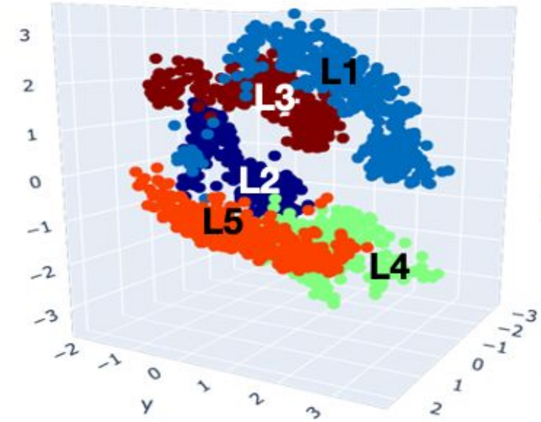
Ref. Aspuru-Guzik



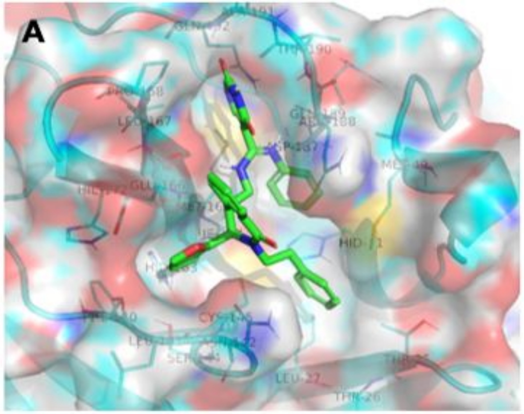
10^{68} estimated drug-like compounds

ML-based models (filter 10^{12} compounds)





AI-driven adaptive conformational sampling (DeepDriveMD)



ESMACS based free-energy estimates

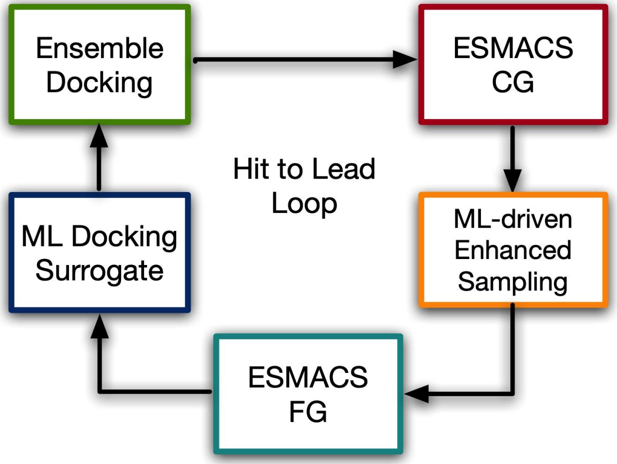
- Improve virtual screening throughput by integrating AI + physics-based models

Campaign: Hit-to-Lead Loop

Multi-stage campaign employed to select promising drug candidates:

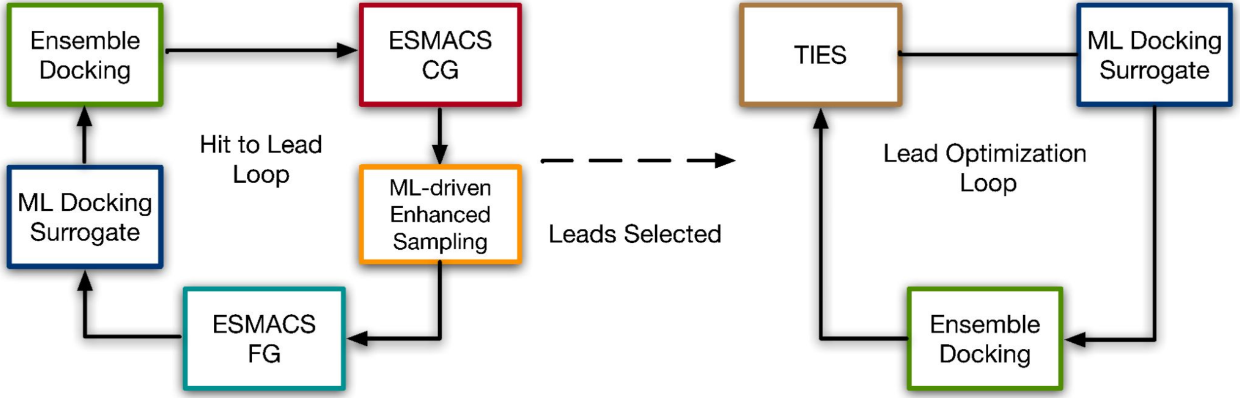
- **WF1**: High-throughput ensemble docking to identify small molecules (“hits”)
- **WF2**: AI-driven Molecular Dynamics for modeling specific binding regions and understanding mechanistic changes involving drugs
- **WF3**: Binding Free Energy calculations of promising leads (“Hit-to-Lead”)

https://arxiv.org/abs/2010.06574



Campaign: Lead Optimization

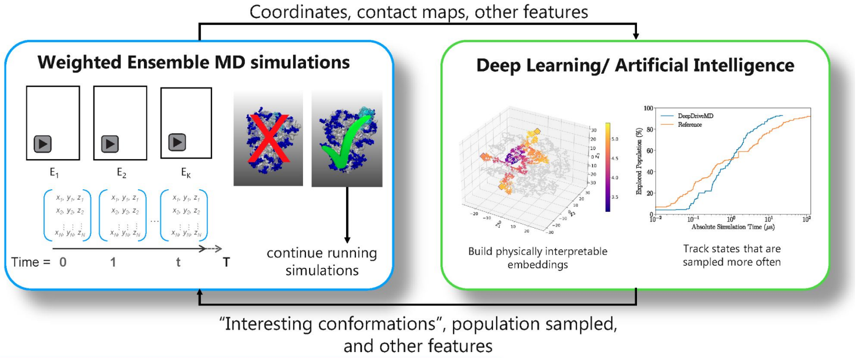
Multi-stage campaign employed to select promising drug candidates:

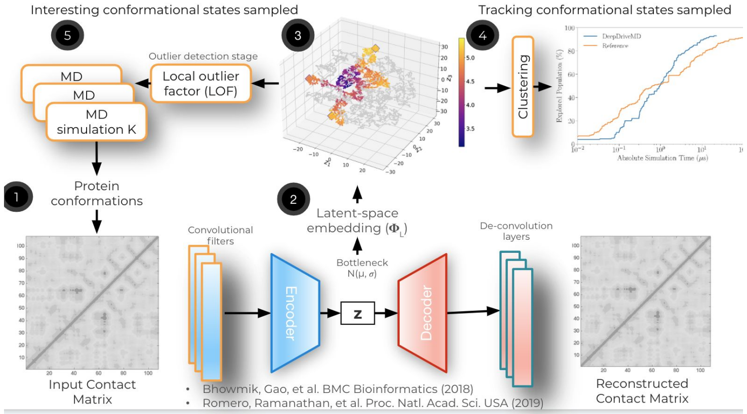


- **WF4**: TIES -- Alchemical Binding Free Energy calculations of promising leads (Lead Optimization)

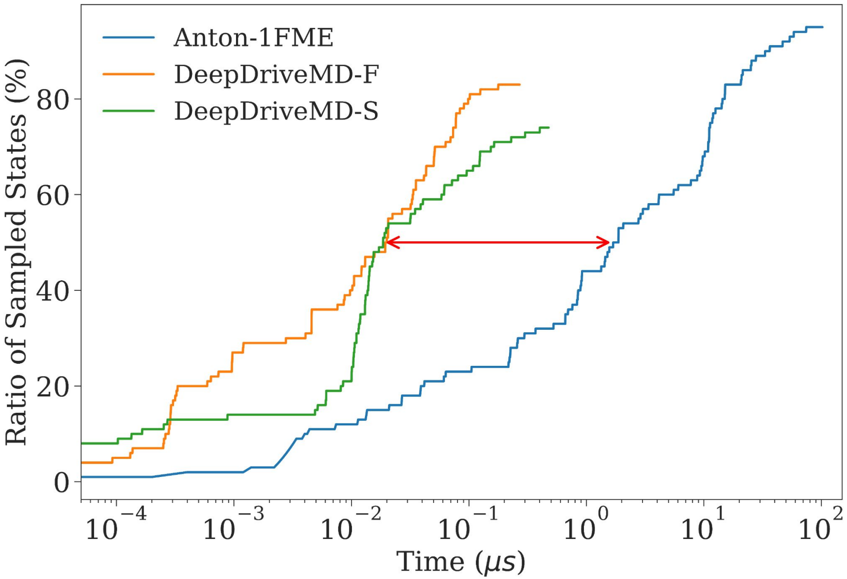
ML-driven Ensemble (WF2): 10-100x Protein Folding

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

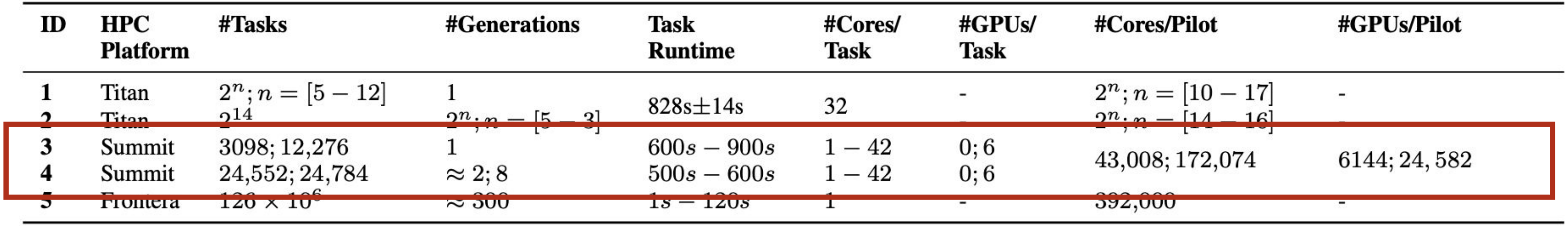

Coordinates, contact maps, other features
Weighted Ensemble MD simulations
E_1
E_2
E_K
Time = 0
1
t
T
continue running simulations
Deep Learning/ Artificial Intelligence
DeepDriveMD
Reference
Explored Population (%)
Absolute Simulation Time (μs)
Build physically interpretable embeddings
Track states that are sampled more often
"Interesting conformations", population sampled, and other features



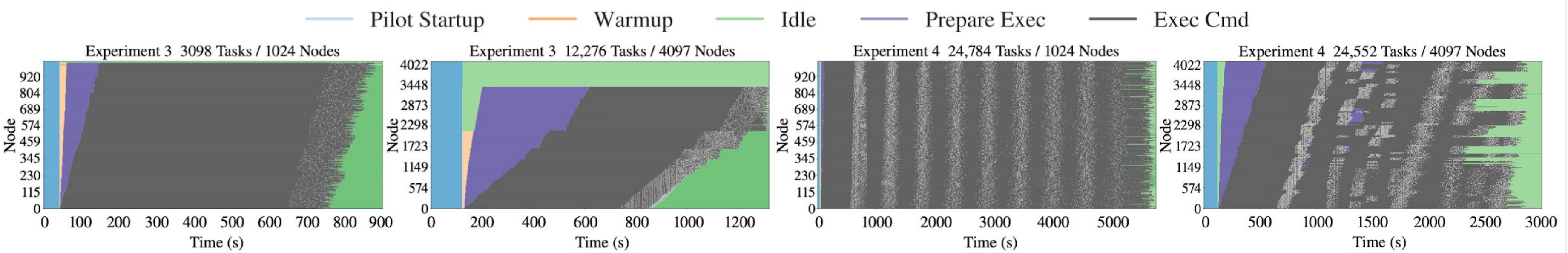
- Bhowmik, Gao, et al. BMC Bioinformatics (2018)
- Romero, Ramanathan, et al. Proc. Natl. Acad. Sci. USA (2019)



Characterizing RP on Leadership Platforms



ID	HPC Platform	#Tasks	#Generations	Task Runtime	#Cores/ Task	#GPUs/ Task	#Cores/Pilot	#GPUs/Pilot
1	Titan	$2^n; n = [5 - 12]$	1	828s±14s	32	-	$2^n; n = [10 - 17]$	-
2	Titan	2^{14}	$2^n; n = [5 - 3]$			-	$2^n; n = [14 - 16]$	-
3	Summit	3098; 12,276	1	$600s - 900s$	$1 - 42$	0; 6	43,008; 172,074	6144; 24,582
4	Summit	24,552; 24,784	≈ 2; 8	$500s - 600s$	$1 - 42$	0; 6		
5	Frontera	126×10^6	≈ 300	$1s - 120s$	1	-	392,000	-

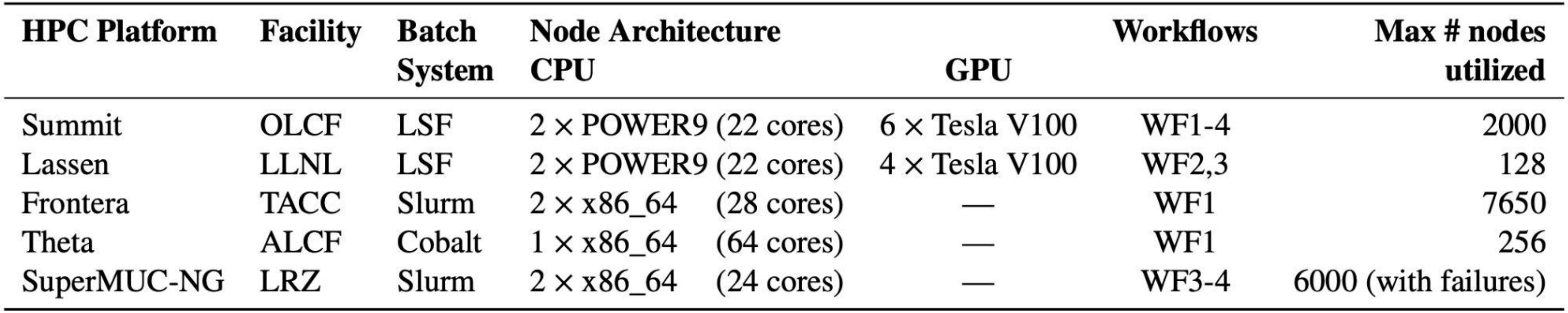


(a) (b) (c) (d)

https://arxiv.org/abs/2103.00091

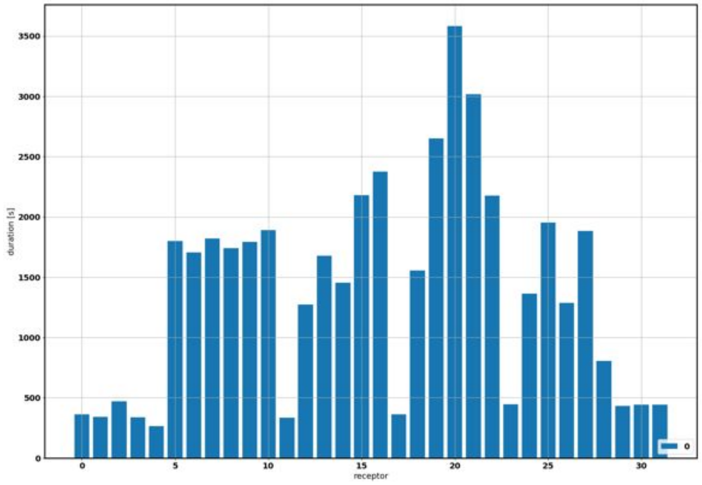
Computational Challenges: Heterogeneity

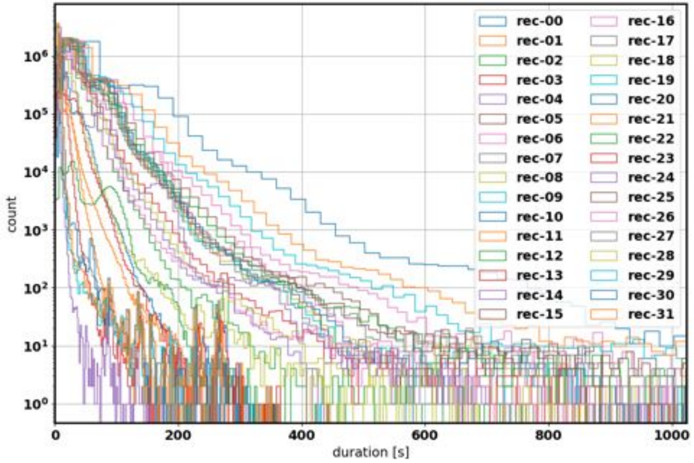
- **Heterogeneity** of different types and at multiple levels
    - Coupled AI-HPC (WF2)
    - High-throughput function calls (WF1)
    - Ensembles of MPI tasks (WF3/4)
- Spatio-temporal variation within and across WF1

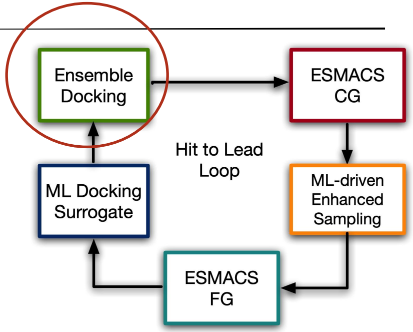


HPC Platform	Facility	Batch	Node Architecture		Workflows	Max # nodes
		System	CPU	GPU		utilized
Summit	OLCF	LSF	2 × POWER9 (22 cores)	6 × Tesla V100	WF1-4	2000
Lassen	LLNL	LSF	2 × POWER9 (22 cores)	4 × Tesla V100	WF2,3	128
Frontera	TACC	Slurm	2 × x86_64 (28 cores)	—	WF1	7650
Theta	ALCF	Cobalt	1 × x86_64 (64 cores)	—	WF1	256
SuperMUC-NG	LRZ	Slurm	2 × x86_64 (24 cores)	—	WF3-4	6000 (with failures)

Ensemble Docking: (WF1)

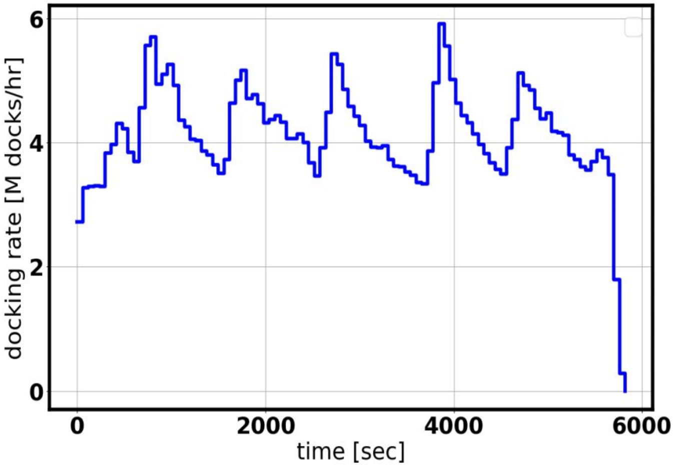




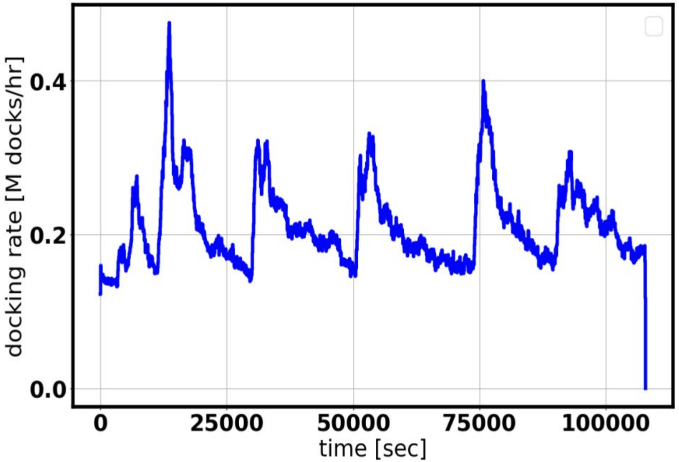


- Docking: OpenEye; Library (ORD): 6.25M ligands (drug candidate); 32 targets/receptors
  - Fluctuations in docking execution time library (ORD) for different receptors
  - Long-tailed Tx for different ligands for a given target (receptor)
  - Many work items (function calls) need to be distributed
  - Call duration varies two order of magnitudes (1-100s). Mean duration 8s.

Ensemble Docking: (WF1)



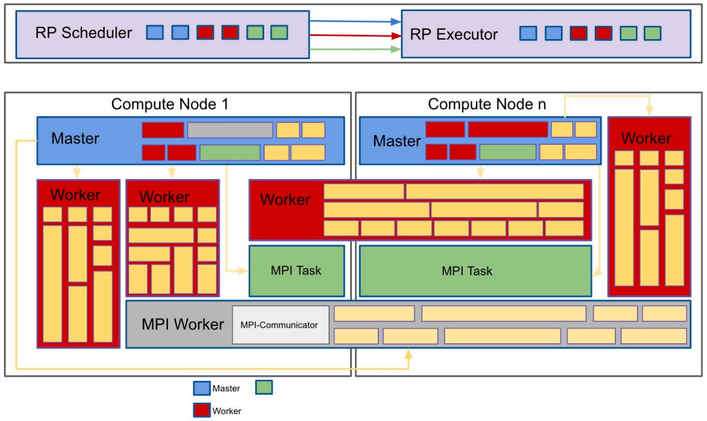
(a)

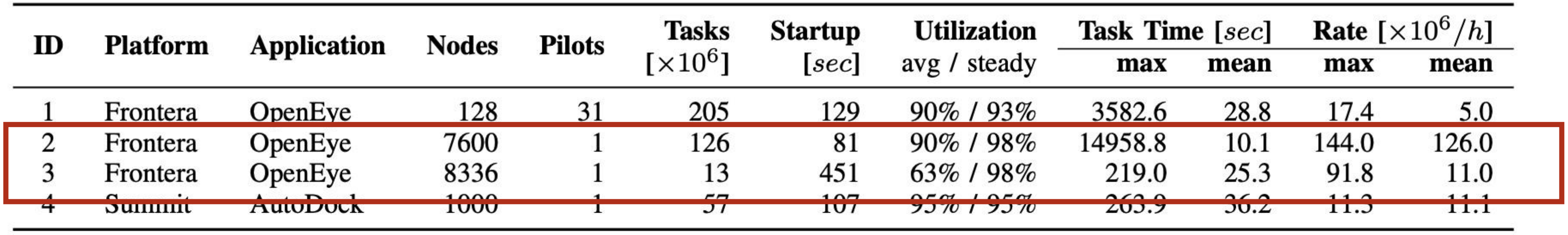


(b)

- Docking: OpenEye; Library (ORD): 6.25M ligands (drug candidate); 32 targets/receptors
  - Fluctuations in docking execution time library (ORD) for different receptors
  - Long-tailed Tx for different ligands for a given target (receptor)
  - Many work items (function calls) need to be distributed
  - Call duration varies two order of magnitudes (1-100s). Mean duration 8s.

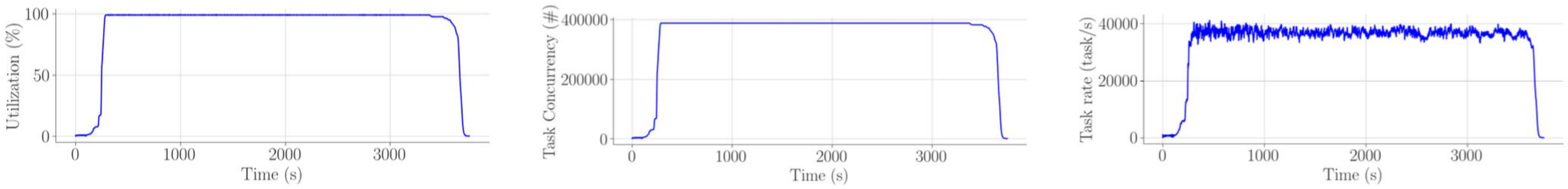
Ensemble Docking (WF1) with RAPTOR



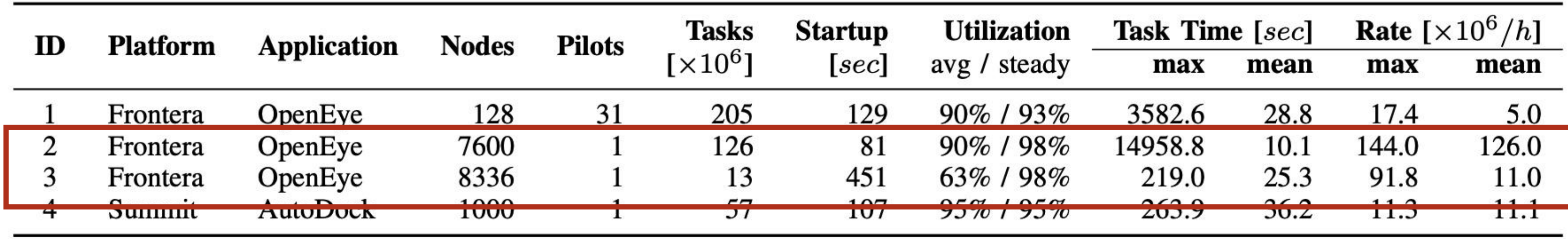


ID	Platform	Application	Nodes	Pilots	Tasks $[\times 10^6]$	Startup $[sec]$	Utilization avg / steady	Task Time $[sec]$ max	Task Time $[sec]$ mean	Rate $[\times 10^6/h]$ max	Rate $[\times 10^6/h]$ mean
1	Frontera	OpenEye	128	31	205	129	90% / 93%	3582.6	28.8	17.4	5.0
2	Frontera	OpenEye	7600	1	126	81	90% / 98%	14958.8	10.1	144.0	126.0
3	Frontera	OpenEye	8336	1	13	451	63% / 98%	219.0	25.3	91.8	11.0
4	Summit	AutoDock	1000	1	57	107	95% / 95%	263.9	36.2	11.3	11.1

RADICAL-Pilot (RP) with RAPTOR : Performance



(a) (b) (c)



ID	Platform	Application	Nodes	Pilots	Tasks $[\times 10^6]$	Startup $[sec]$	Utilization avg / steady	Task Time $[sec]$		Rate $[\times 10^6/h]$	
								max	mean	max	mean
1	Frontera	OpenEye	128	31	205	129	90% / 93%	3582.6	28.8	17.4	5.0
2	Frontera	OpenEye	7600	1	126	81	90% / 98%	14958.8	10.1	144.0	126.0
3	Frontera	OpenEye	8336	1	13	451	63% / 98%	219.0	25.3	91.8	11.0
4	Summit	AutoDock	1000	1	57	107	95% / 95%	263.9	36.2	11.3	11.1

Impacting SARS-CoV-2 Medical Therapeutics

radical

- **Scale of Operation:**
  - **~10^{11}** Docking calculations
  - ~10^{3} ML-driven MD calculations
  - ~5 x 10^{4} Binding Free Energy Calculations
  - ~2.5 x 10^{6} node-hours (~30 days, all Summit)
- Peak Performance
  - ~ **8000** nodes (Frontera, April. 2021)
  - ~ **4000** nodes on Summit
- Extensible Computational Infrastructure and Capabilities
  - Beyond COVID-19 ?

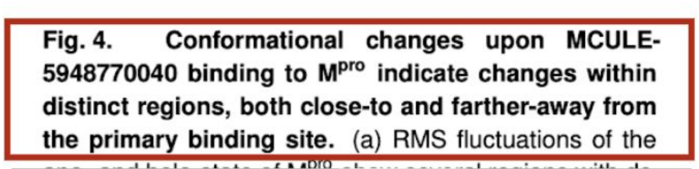
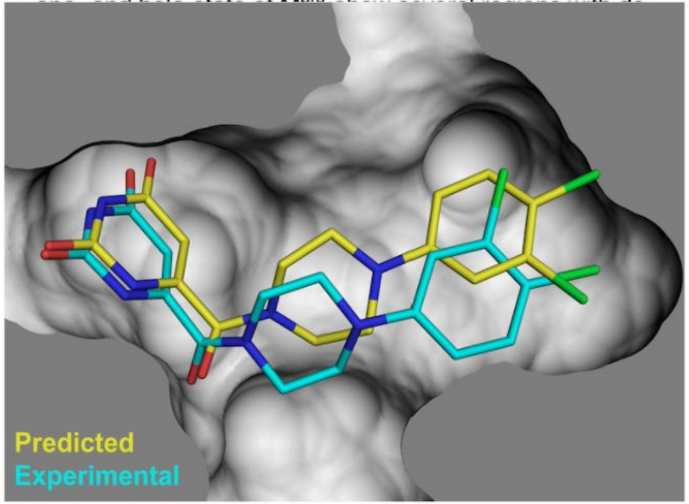


Fig. 4. Conformational changes upon MCULE-5948770040 binding to M^{pro} indicate changes within distinct regions, both close-to and farther-away from the primary binding site. (a) RMS fluctuations of the



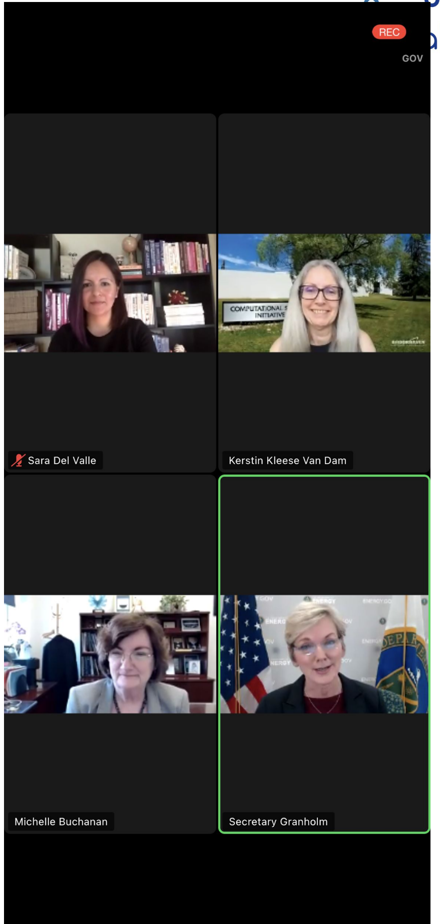
… under review PNAS

Therapeutics: Needle in multiple Haystacks?

- **Scale of Operation:**
  - **~10^{11}** Docking calculations
  - ~10^{3} ML-driven MD calculations
  - ~5 x 10^{4} Binding Free Energy Calculations
  - ~2.5 x 10^{6} node-hours (~30 days, all Summit)
- Peak Performance
  - ~ **8000** nodes (Frontera, April. 2021)
  - ~ **4000** nodes on Summit
- Extensible Computational Infrastructure and Capabilities
  - Beyond COVID-19 ?

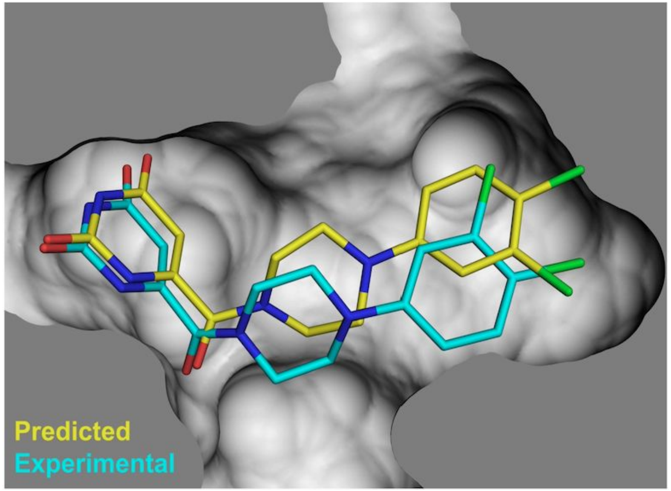
… under review PNAS

BROOKHAVEN
NATIONAL LABORATORY

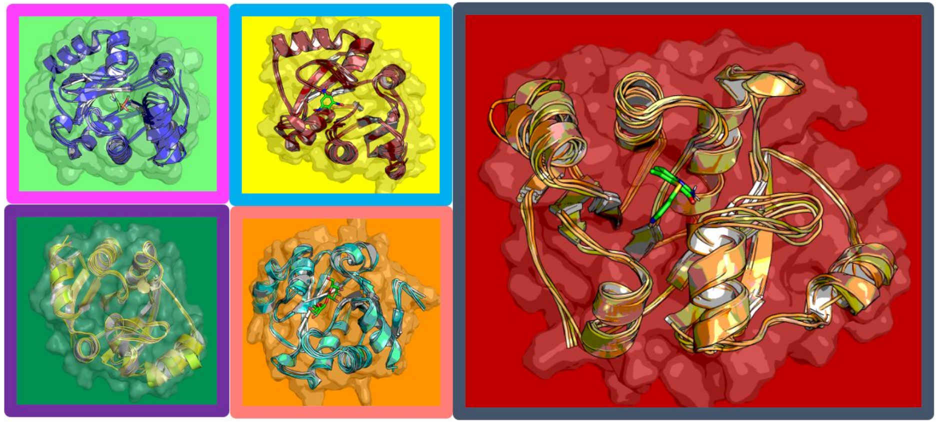


Summary

- ML enhances the **effective performance**
  - ML “improve” performance of simulations
  - “ …. *simulations are mere generators of data for powerful ML models” !*
- Exascale computing on petascale platforms!
  - Developed 1st gen of AI-HPC infrastructure
  - Sophistication of AI-HPC methods will grow
- Rethink systems software ecosystem
  - Collective perf. of heterogeneous workflows; not just single tasks
  - Advances in adaptive runtime systems for such workflows



Thank you!



Funding acknowledgement:

- DOE National Virtual Biotechnology Laboratory
- DOE CANDLE ECP
- ECP ExaWorks and ECP ExaLearn
- ASCR Surrogates Benchmarking Initiative
- NSF RADICAL-Cybertools