Coupling streaming AI and HPC ensembles to achieve 100-1000× faster biomolecular simulations

Alexander Brace\textsuperscript{1,3*}, Igor Yakushin\textsuperscript{1*}, Heng Ma\textsuperscript{1*}, Anda Trifan\textsuperscript{1,4}, Todd Munson\textsuperscript{2+}, Ian Foster\textsuperscript{1,3+}, Arvind Ramanathan\textsuperscript{1++}, Hyungro Lee\textsuperscript{5*}, Matteo Turilli\textsuperscript{5,6}, Shantenu Jha\textsuperscript{5,6++}

\textsuperscript{1}Data Science and Learning Division, Argonne National Laboratory, \textsuperscript{2}Mathematics and Computer Science Division, Argonne National Laboratory, \textsuperscript{3}University of Chicago, \textsuperscript{4}University of Illinois Urbana-Champaign, \textsuperscript{5}Rutgers University, \textsuperscript{6}Computational Science Initiative, Brookhaven National Laboratory; \textsuperscript{*}Senior authors \textsuperscript{+}First authors
Science motivation: Studying complex biological systems and processes is important

Protein folding provides insights into how their 3D structure enables function e.g. FSD-EY ($\beta\beta\alpha$)

Protein ligand complexes evaluate drug binding to discover novel inhibitors e.g. Papain-like protease (PLPro)

Many-Atom Multiscale Systems uncover large-scale dynamics necessary to understand systems such as SARS-CoV-2

Casalino, L. et al, AI-driven multiscale simulations illuminate mechanisms of SARS-CoV-2 spike dynamics, https://doi.org/10.1177/10943420211006452 (SC’20 Gordon Bell Award for HPC in COVID-19 Research)
What is molecular dynamics and why do we do it?

- Molecular dynamics (MD) can act as a **computational microscope** which can reveal biomolecular detail at greater spatial and temporal scales than is possible with experiment.

- Integrates Newton’s second law of motion to advance each atom in the system forward in time along a potential energy gradient.

\[
\vec{F}_i = ma = m_i \frac{d^2 \vec{r}_i}{dt^2} = -\nabla U(\vec{R})
\]

Membrane simulation
Multiscale simulations impose new requirements on emerging hardware/software

Traditional Analytics + Simulation

- Job scheduler
- Simulation(s)
- Data storage (Disks)
- Analytics
- Visualization

Job scheduler

Simulation(s)

Data storage (Disks)

Analytics

Visualization

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

- Large simulations generate > O(100TB) of data
- Humanly impossible to peek into “biologically” interesting events!
- Traditional method is unsustainable at exascale
AI-enabled MD simulations with DeepDriveMD

Coordinates, contact maps, other features

Ensemble MD simulations

Deep Learning/Artificial Intelligence

Build physically interpretable embeddings

Track states that are sampled more often

“Interesting conformations”, population sampled

Learning Everywhere

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

- **Execution:** Each task is executed as an independent function call via Parsl, as orchestrated by the Colmena “Thinker”.

- **Parallelism:** Decoupled design of components allows for high degrees of parallelism (e.g., 100s of parallel simulation tasks).

- **Communication:** Each component writes outputs to disk (e.g., simulation data, model weights, etc) and a path to the data is returned via Proxystore. However, other communication protocols can be used without change to the framework.

Ward, L. et al., *Colmena: Scalable Machine-Learning-Based Steering of Ensemble Simulations for High Performance Computing*, https://arxiv.org/abs/2110.02827
DeepDriveMD enables $10^4 \times$ 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 more than 1000X faster than Anton-1 simulations.

**Note:** Uncertainty from 10 trials is shown in light red.

Brace, A. et al., *Achieving 100X faster simulations of complex biological phenomena by coupling ML to HPC ensembles.*, https://arxiv.org/abs/2104.04797
Modular Simulate-Train-Infer Framework

**Interface:** An abstraction layer over the particular simulation / training / inference tasks allows users to customize DeepDriveMD to any simulation engine or ML method.
The correct compute settings are automatically determined via Pydantic type checking by specifying the “name” field in YAML along with conditional args.

Code: https://github.com/ramanathanlab/deepdrivemd/tree/main
Acknowledgements

Logan Ward & Kyle Chard for supporting the Parsl/Colmena-based implementation of DeepDriveMD

Rick Stevens (Argonne/University of Chicago)
Rommie Amaro (University of California San Diego)
John Stone (University of Illinois Urbana-Champaign)
Lillian Chong (University of Pittsburgh)
Tom Gibbs (NVIDIA)
~300 collaborators on COVID-19 research

Funding:
-- DOE ASCR Co-design of AI Approaches for Multi-modal datasets
-- Exascale Computing Project (ECP): Cancer Deep Learning Environment (CANDLE), ECP Co-design for Online Data Analysis and Reduction CODAR, ExaWorks
-- Department of Energy’s Advanced Scientific Computing Research Program
-- DOE National Virtual Biotechnology Laboratory

Computing:
-- Argonne/Oak Ridge Leadership Computing Facilities
-- Livermore Computing at the Lawrence Livermore National Laboratory
-- HPC Consortium for COVID-19 research
THANK YOU!!

QUESTIONS/ COMMENTS:
RAMANATHANA@ANL.GOV