OAK RIDGE National Laboratory

Artificial Intelligence Enabled Multiscale Molecular Simulations

Arvind Ramanathan

Team Lead, Integrative Systems Biology, Computational Sciences and Engineering Division, Health Data Sciences Institute, Oak Ridge National Laboratory, Oak Ridge

ramanathana@ornl.gov http://ramanathanlab.org

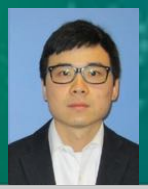
Srikanth Yoginath
CSMD

Christopher B. Stanley
CSED

Dmitry I. Liakh
NCCS

Ramakrishnan Kannan
CSMD

Debsindhu Bhowmik
CSED



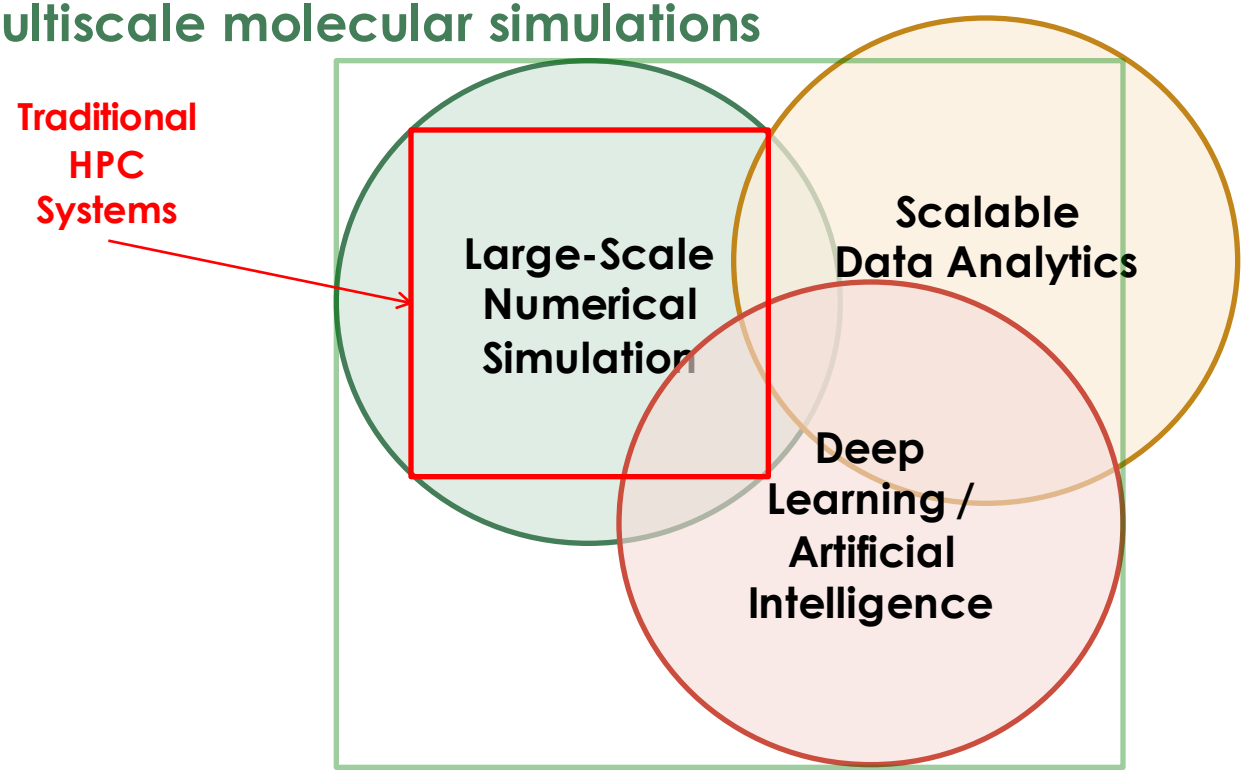
Heng Ma
CSED

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

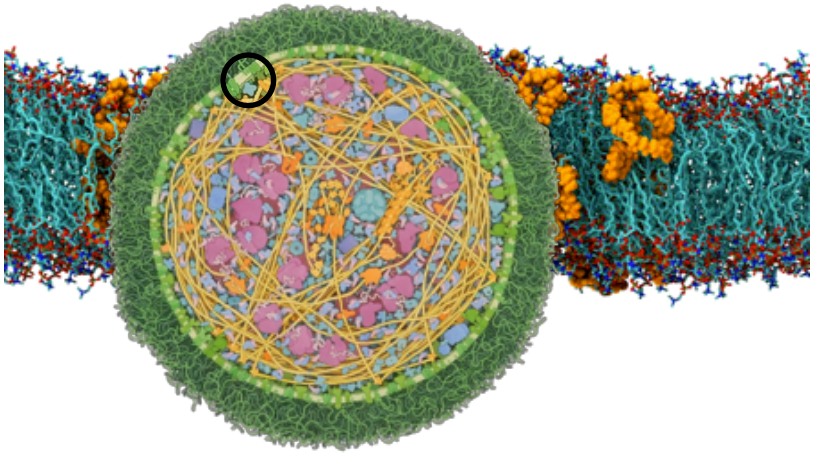
U.S. DEPARTMENT OF ENERGY

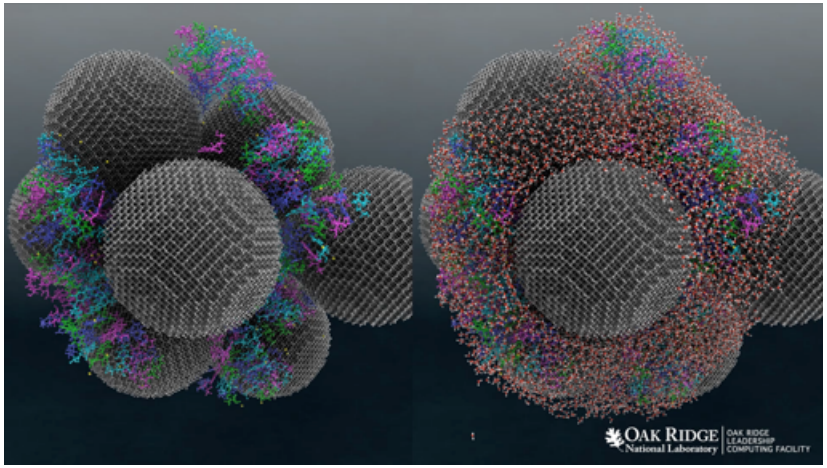
Drive Integration of Simulation, Data Analytics and Machine Learning

Molecules Library: AI-driven multiscale molecular simulations

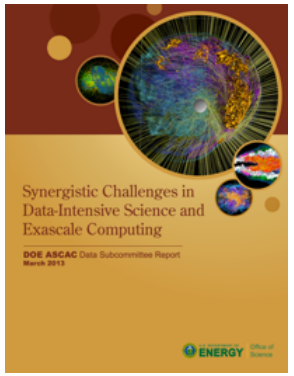


Multi-scale phenomena in biological systems pose challenges for modeling/ simulations @ Exascale



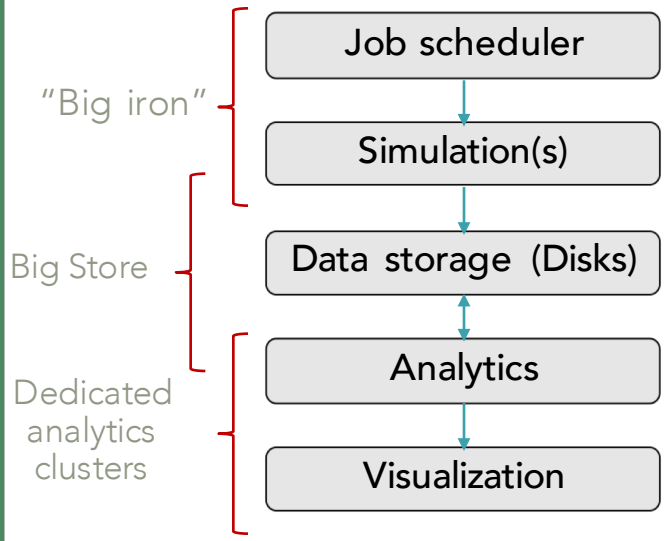


- Simulations of physical phenomena take 45-60% of supercomputing time
  - Coupled to experimental data such as Spallation Neutron Source, X-ray scattering/ diffraction facilities, etc.
- "Exascale simulations will require some analyses… be performed while data is still resident in memory…"



Towards AI-driven simulations: Interleaving data analytics + simulations

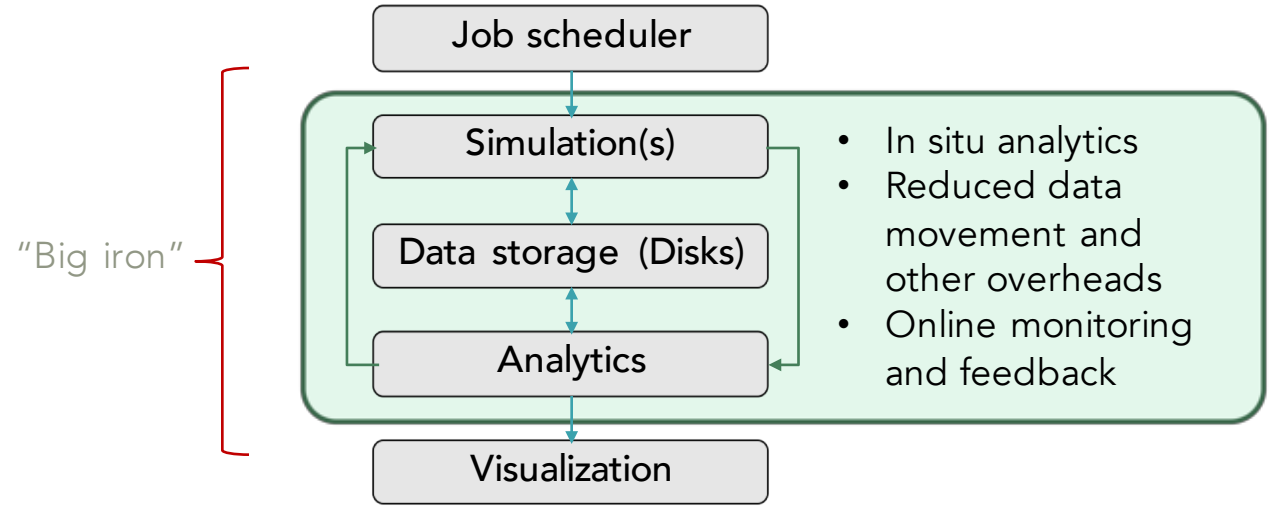
Traditional Compute + Simulations



Unsustainable at Exascale

- Data movement bottlenecks
- Parallel analytics bottlenecks

Interleaving Analytics + Simulations

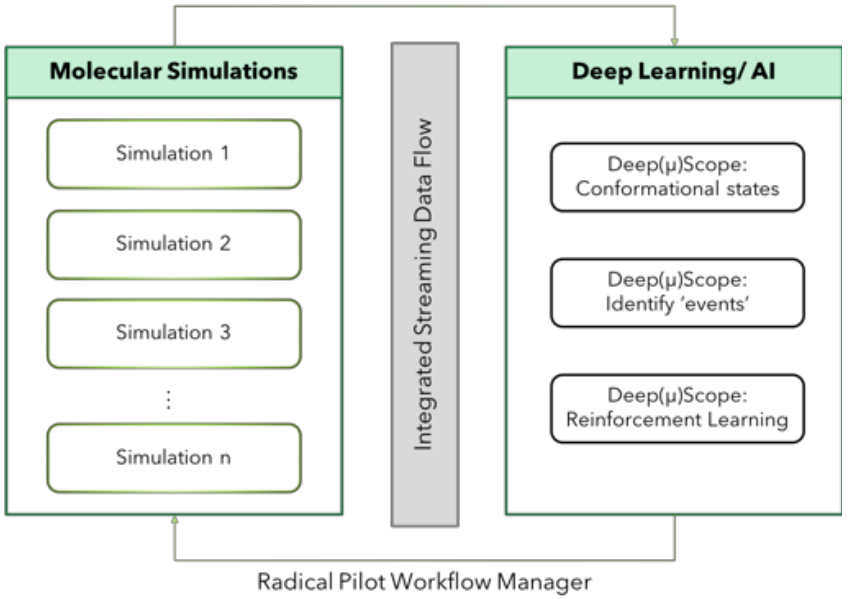


Iterative forward/backward loop

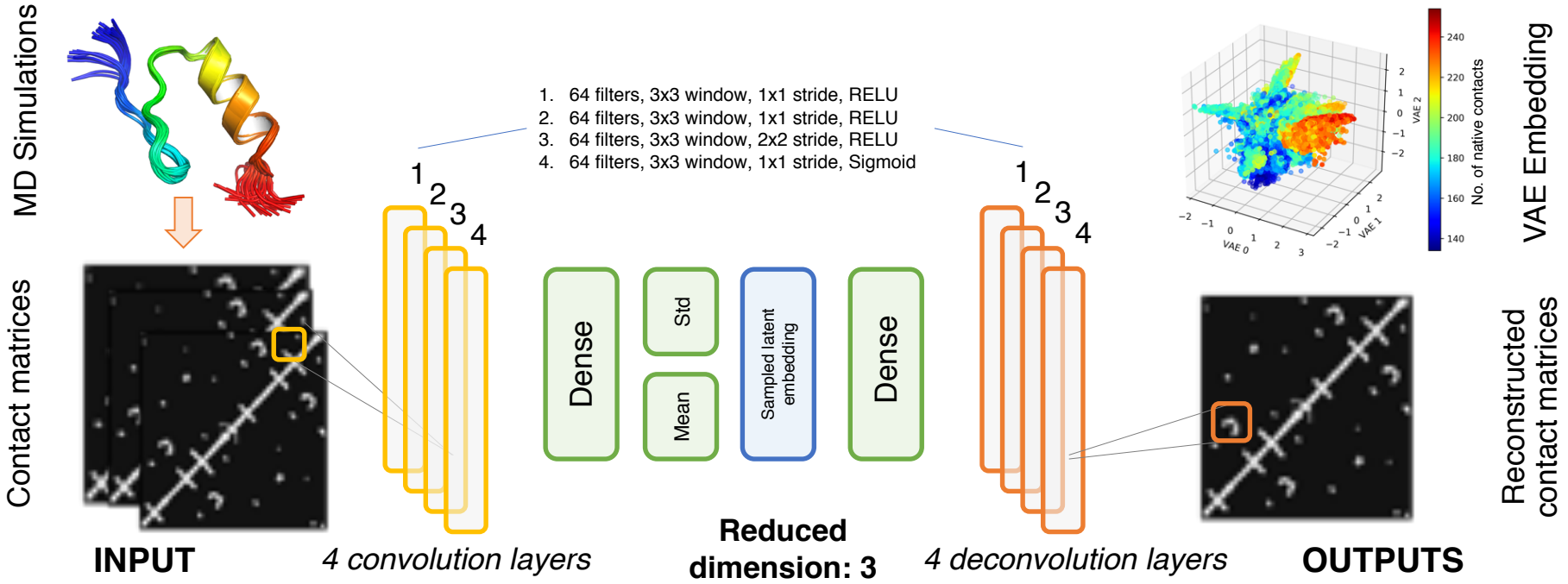
- **High performance framework** to monitor & analyze simulations as they are running with little/ no modification to simulation software
- **Demonstrate on molecular dynamics simulations**, but generalize framework for broad applicability

Outline: Can artificial intelligence (AI) techniques be leveraged for accelerating molecular simulations?

- Building effective (low-dimensional) latent representations of simulation datasets:
  - Using deep learning approaches for molecular dynamics (MD) data
  - Scaling convolutional variational autoencoder for MD
- Predicting where we should go next in MD simulations:
  - Building a recurrent autoencoder to predict future steps
- Preliminary work on a reinforcement learning approach for protein folding/ docking



A variational approach to encode protein folding with convolutional auto-encoders

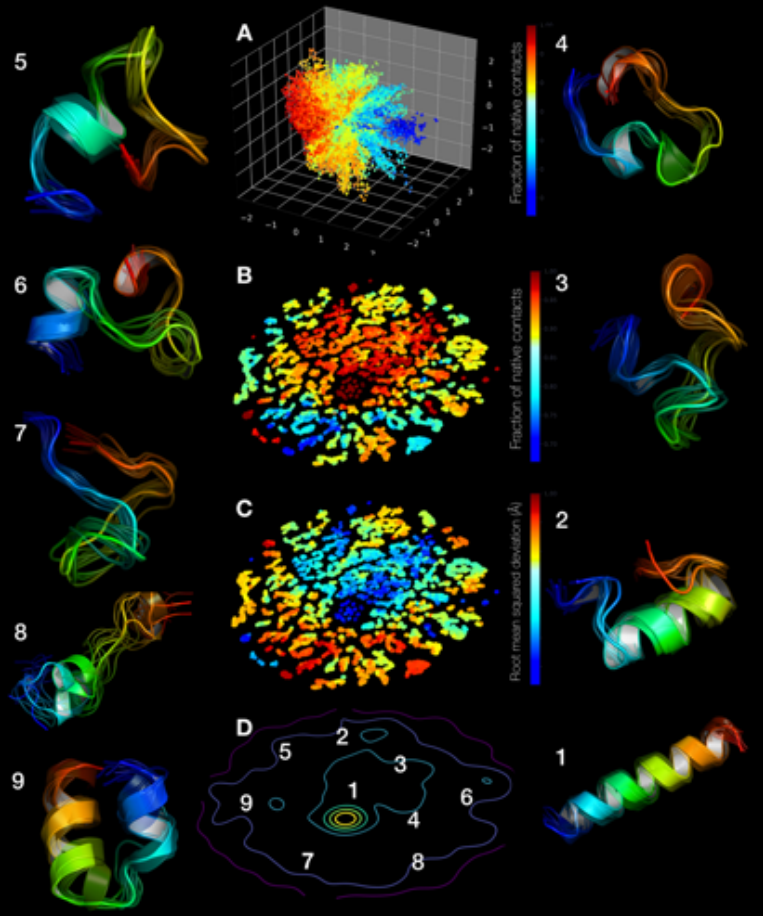


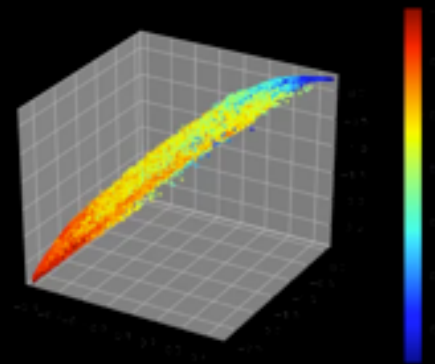
Related work:
Hernandez 17 arXiv,
Doerr 17 arXiv

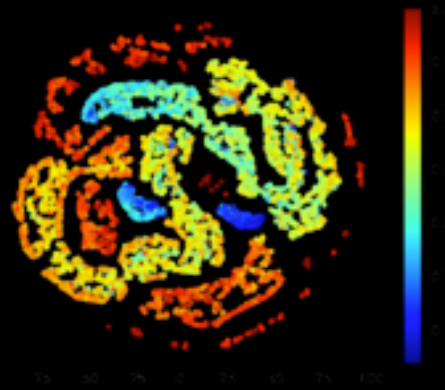
OAK RIDGE National Laboratory

D. Bhowmik, M.T. Young, S. Gao, A. Ramanathan, BMC Bioinformatics (accepted)

Deep Learning reveals "metastable states" in protein folding…



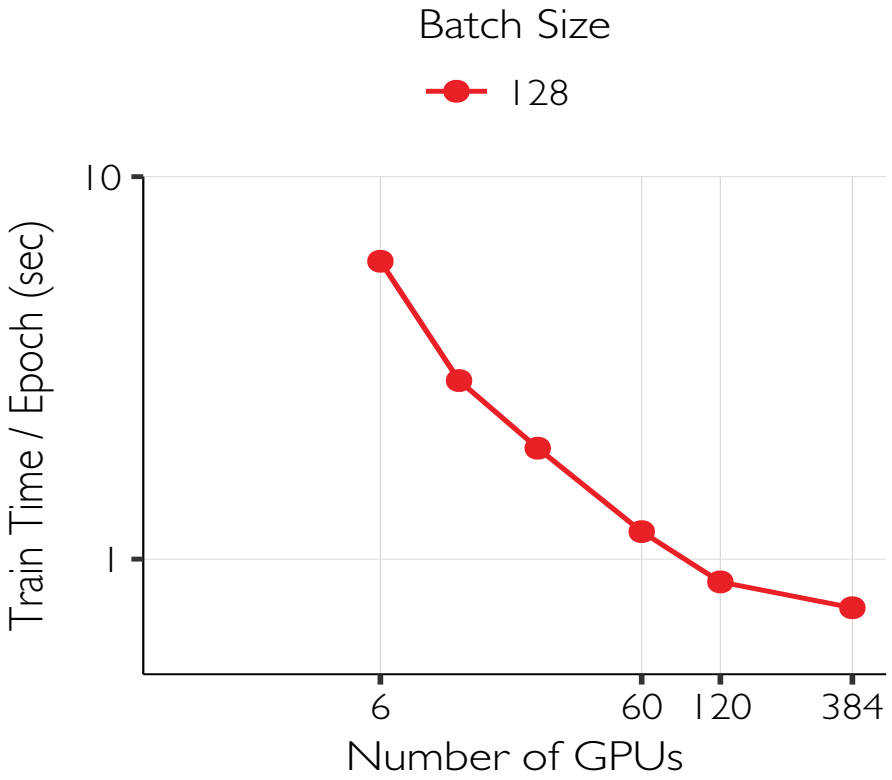




MSM Builder Datasets, Pande group

Scaling & Performance: Enabling DL approaches to achieve near real-time training/prediction

- Scaling deep learning on Summit to facilitate online training
  - DeepEx: a custom-built deep learning stack for Summit
- Exploiting low rank structure of scientific data:
  - Accelerate training
  - Scale to larger datasets
- Performance on Resnet like convolutional nets



S. Yoginath , M. Alam, K. Perumalla, R. Kannan, D. Bhowmik, A. Ramanathan, ORNL Tech Report

Current platforms for hyperparameter optimization rely on sequential optimization techniques

- Bayesian optimization, Bandit
  usually sequential search proc
- Exponential scaling:
  - The number of samples required
    optimization procedure is scales
    search dimensions, as in 2^D, wher
  - Forgotten in the recent exciteme

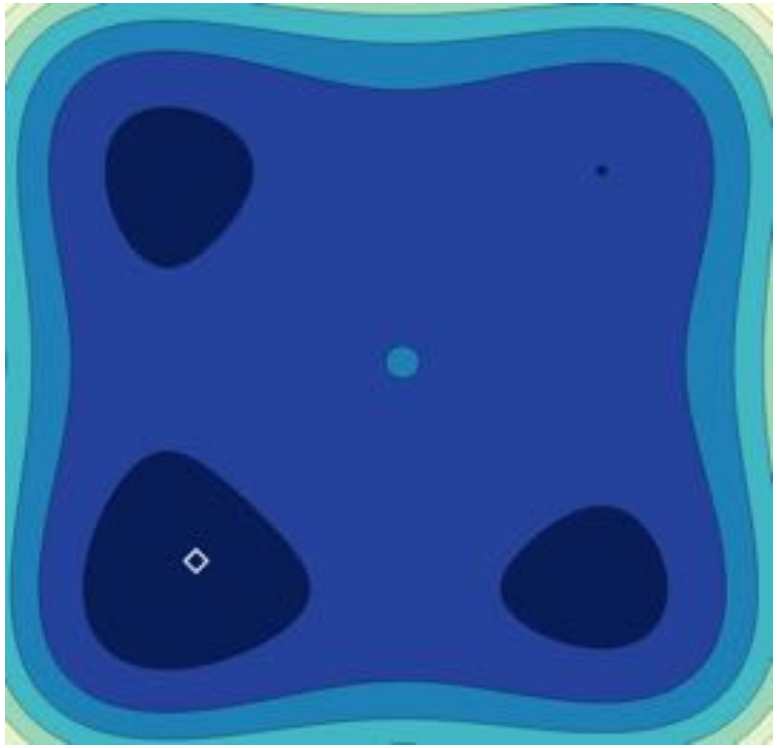
HyperSpace: Distributed Parallel Bayesian Optimization

- Hyperspace, instead seeks to focus on the search space:
  - Parallelism to exploit the statistical structure of the search space
  - Reveal partial dependencies across parameter spaces
- Build many surrogate functions in parallel
- {Prayer}!

N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger. Gaussian pr
abs/0912.3995, 2009.

S. Grunewalder, J.-Y. Audibert, M. Opper, and J. Shawe-Taylor. Regre
Thirteenth International Conference on Artificial Intelligence and Statis
2010.

HyperSpace: Parallel exploration of large search spaces



https://github.com/yngtodd/hyperspace

1. Define the bounds of each hyperparameter search space.

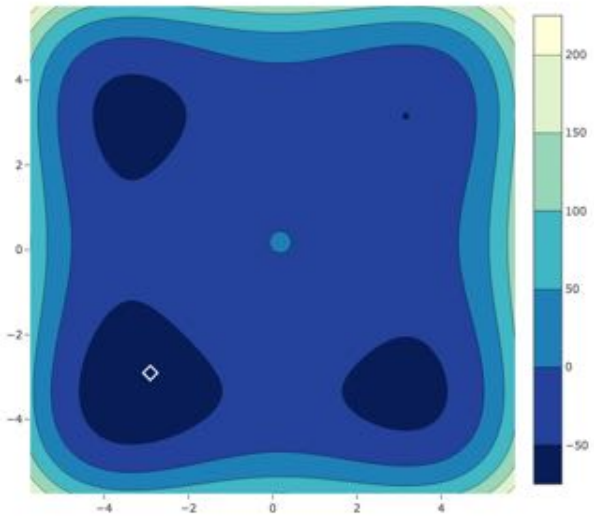
2. Divide each search space bound into two nearly equal sub-bounds with overlap ϕ, where $\{\phi \in \mathbb{R} \mid 0 \leq \phi \leq 1\}$.

3. Create all possible combinations of hyperparameter sub-bounds to form 2^D search spaces (hyperspaces) where D is the number of model hyperparameters.

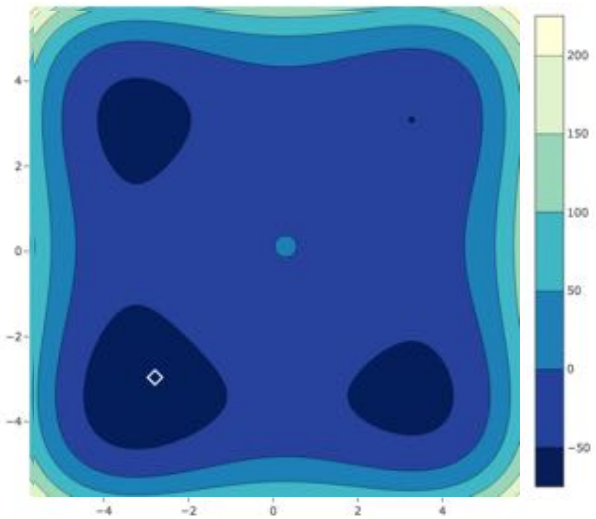
4. Run Bayesian optimization over each hyperspace in parallel

M. Todd Young, J. D. Hinkle, R. Kannan, A. Ramanathan, HyperSpace: Massively Parallel Bayesian Optimization, Workshop on High Performance Machine Learning, 2018, Lyon, France

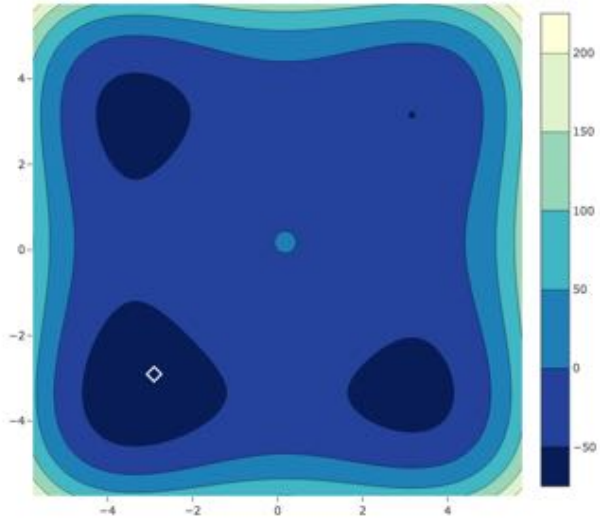
Parallel exploration of large search spaces works better than random/ sequential based optimization



HyperSpace



Random search



SMBO

- Exploiting statistical dependencies in the hyperparameter dimensions leads to better set of parameters for ML models

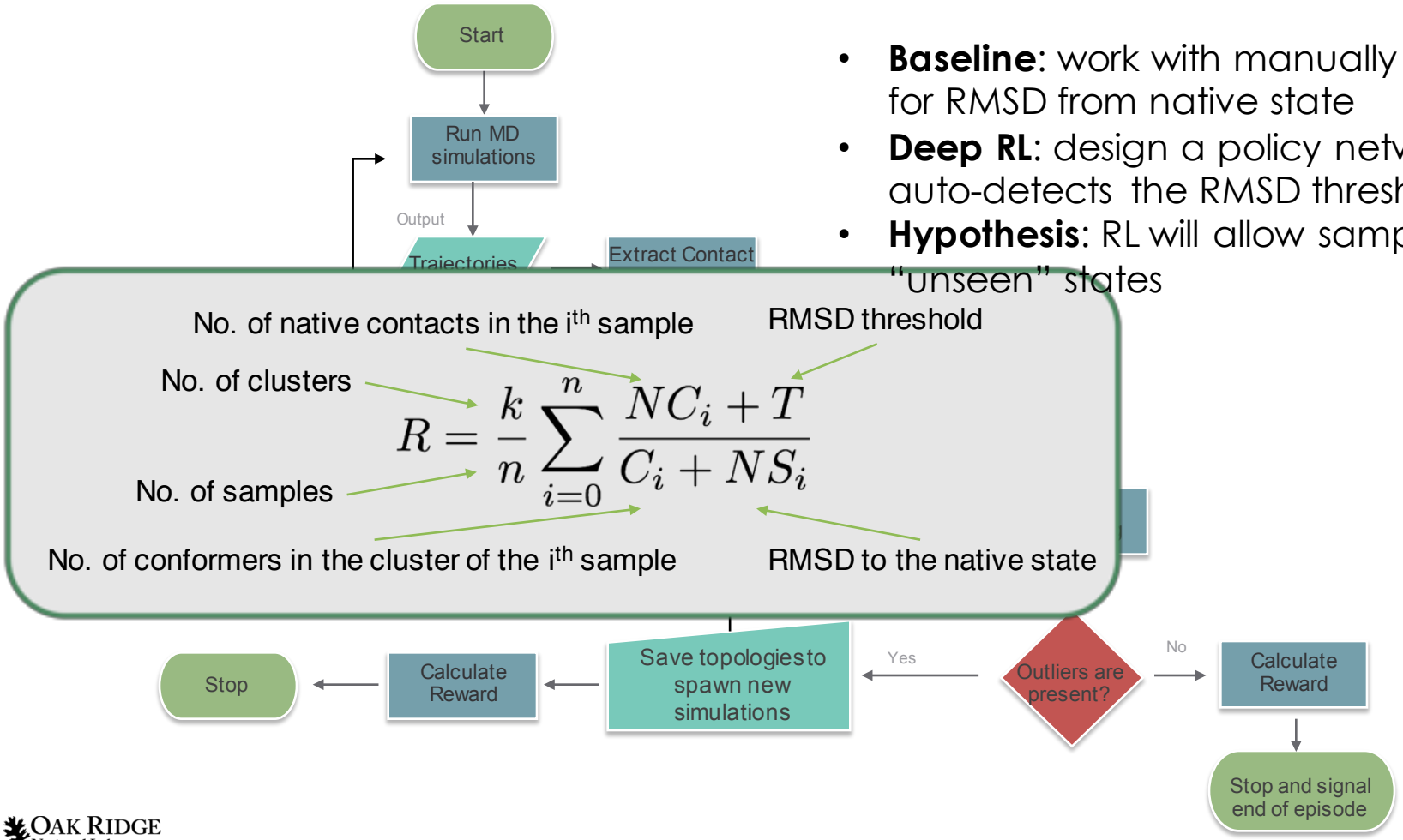
OAK RIDGE National Laboratory

https://hpml2018.github.io/HPML2018_7/index.html#/

Outline: Can AI techniques be leveraged for biological experimental design?

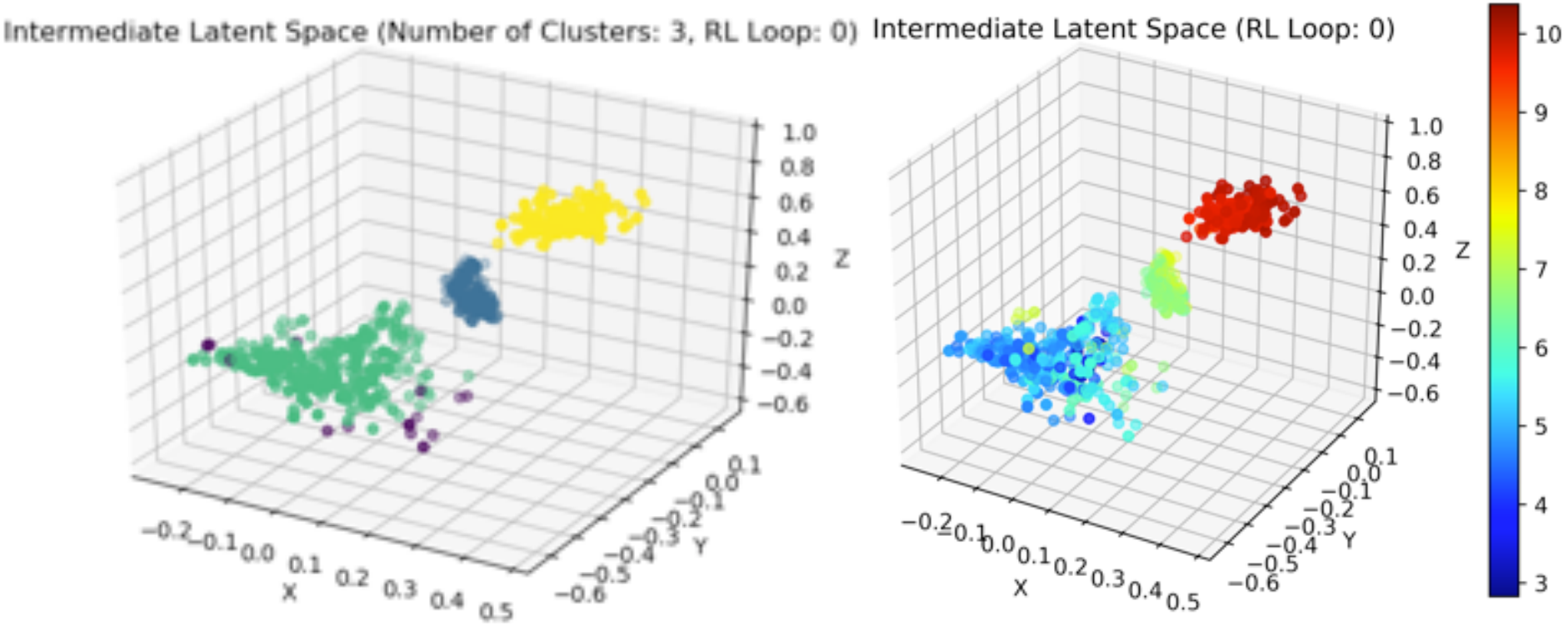
- Building effective (low-dimensional) latent representations of simulation datasets:
  - Using deep learning approaches for molecular dynamics (MD) data
  - Scaling convolutional variational autoencoder for MD
- Predicting where we should go next in MD simulations:
  - Building a recurrent autoencoder to predict future steps
- Preliminary work on a reinforcement learning approach for protein folding/ docking

RL-Fold: a naïve design based on native structure



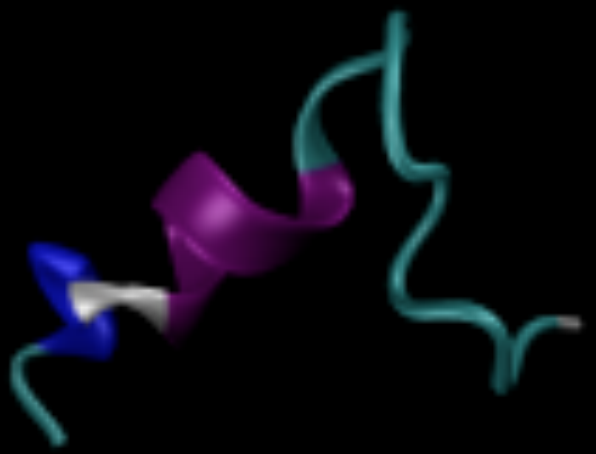
- **Baseline**: work with manually set threshold for RMSD from native state
- **Deep RL**: design a policy network that auto-detects the RMSD threshold
- **Hypothesis**: RL will allow sampling of "unseen" states

Pre-trained deep learning model allows RL explores possible states in protein folding



OAK RIDGE National Laboratory

How does the folding look?



- Within 3-4 iterations, RL reaches near native state RMSD
- Further cycles explore misfolded states:
  - Unfold within a few steps of MD simulations
  - Sampling allows exploration of more intermediate states
- Builds on all-atom simulations + RL in a loop

Summary

- *Deep learning / AI techniques show promise*: learning biophysical characteristics that can be used to guide simulations
- *Reinforcement learning:* Preliminary evidence suggests the approach is feasible to speed up protein folding simulations!
  - How to integrate with physics-based models?
  - How to build scalable approaches beyond RL?
  - How to integrate with sparse experimental observables?
- Enabling iterative, active, and optimal experimental design
- *Extensible library:* Molecules to enable analysis of MD simulations at scale with Deep(μ)scope supporting AI-driven MD simulations

Some emerging challenges in HPC for multi-scale simulations…

- Design of coupled data analytic and simulation workflows on OLCF - Summit and ALCF – A21/Theta
  - In situ analytics approaches are required
  - Streaming applications of ML are different from post-processing of data

- Scaling DL/ AI approaches for biomolecular simulations
  - Faster and more efficient training for deep learning / AI approaches
  - Tensor based approaches to build deep learning algorithms

OAK RIDGE National Laboratory

THANK YOU !!!

- ORNL LDRD – Exascale computing initiative
- DOE-NCI Joint Design of Advanced Computing Solutions for Cancer (JDAS4C)
- DOE Exascale Computing Project Cancer Deep Learning Environment (CANDLE)
- OLCF Early Science Access (Summit)

Questions/ Comments: ramanathana@ornl.gov