Argonne NATIONAL LABORATORY

GENERATIVE MODELING OF PROTEIN FOLDING TRANSITIONS WITH RECURRENT AUTO-ENCODERS

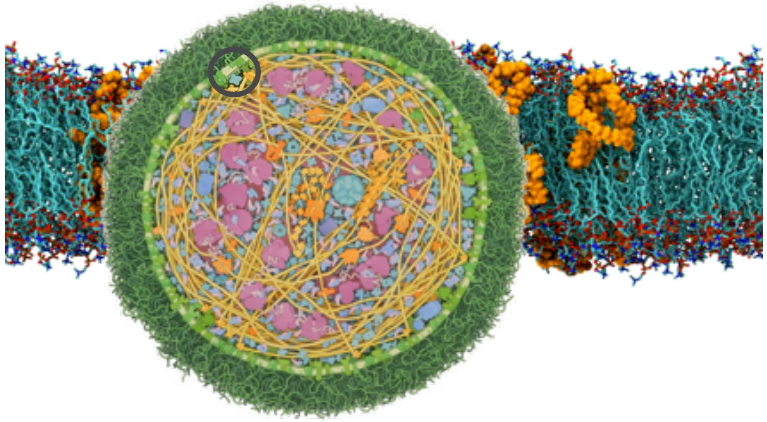
DEBSINDHU BHOWMIK,[1] MICHAEL T. YOUNG,[1] CHRISTOPHER B. STANLEY,[1] ARVIND RAMANATHAN[2]

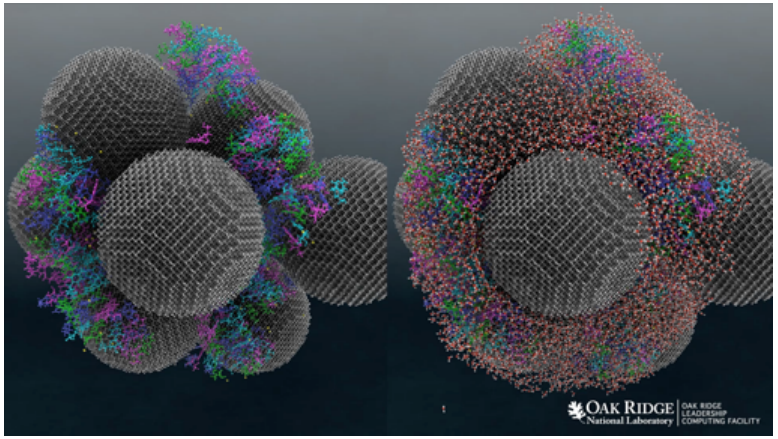
[1]Computational Science & Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830
[2]Data Science & Learning Division, Argonne National Laboratory, Lemont, IL 60439

Email: ramanathana@anl.gov

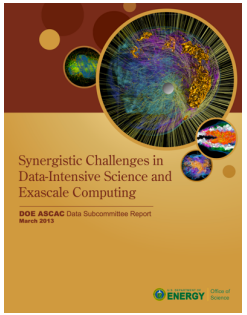
Deep Learning for Science Workshop ISC 2019

MOTIVATION & NEED: INTEGRATION OF AI + MOLECULAR DYNAMICS (MD) SIMULATIONS

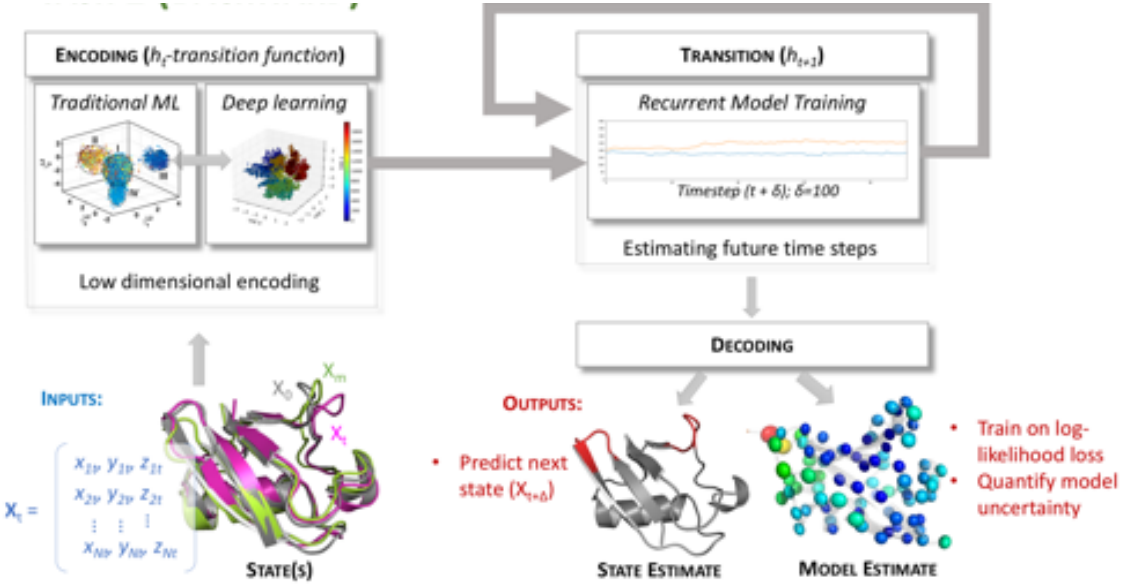




- Simulations of physical phenomena take 45-60% of supercomputing time
  - Coupled to experimental data
- “Exascale simulations will require some analyses… be performed while data is still resident in memory…”



KEY CONTRIBUTIONS: ALGORITHMS FOR AI-DRIVEN MD SIMULATIONS



- Data analytic framework without need to modify underlying simulation software:
  - Online (in situ) analytics for feature extraction and evaluation
  - Simulation scaling can be carried out independently

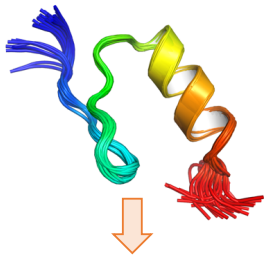
- Propagation framework in lower dimensions allows for better scheduling, improving throughput
  - Significant reduction in model evaluation through integration of equations
  - Simultaneously quantifies how good "current" state and model estimates are

OUTLINE: ALGORITHMS FOR AI-DRIVEN MD SIMULATIONS

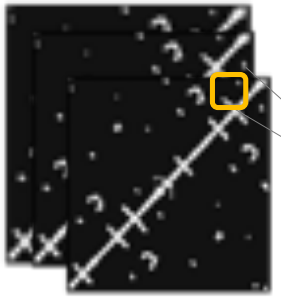
- Building biophysically meaningful, low-dimensional latent representations of simulation data:
  - Deep learning for MD data
  - Convolutional variational autoencoder
- Predicting where we should go next in MD simulations:
  - Building a recurrent autoencoder to predict future step
- Preliminary work on using reinforcement learning to fold proteins

A VARIATIONAL APPROACH TO ENCODE PROTEIN FOLDING WITH CONVOLUTIONAL AUTO-ENCODERS

MD Simulations



Contact matrices



INPUT

$$\mathcal{L} = \mathcal{E}_r + \mathcal{E}_l$$

Error of Reconstruction

How good is my pixel-by-pixel reconstruction?

$$\mathcal{E}_r = -\frac{1}{n}\sum_{i=1}^{n} X_i \log(f(z_i))$$

Latent loss

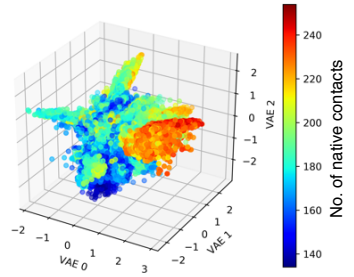
Let's also not get too dragged by outliers…

$$\mathcal{E}_l = \mathrm{KL}(z||Normal(0,1))$$

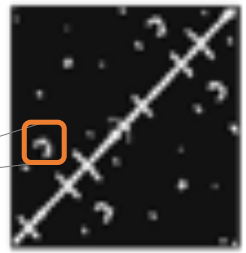
4 convolution layers

dimension: 3

4 deconvolution layers



VAE Embedding



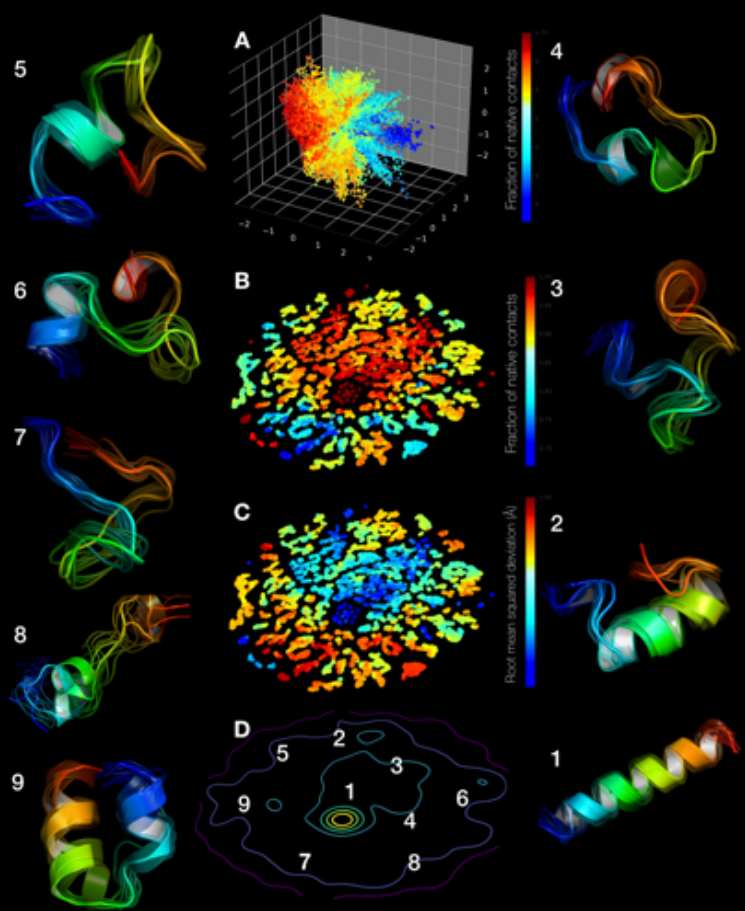
Reconstructed contact matrices

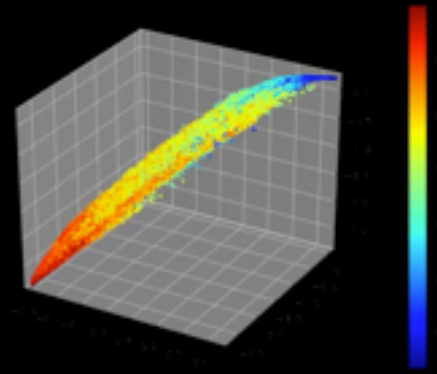
OUTPUTS

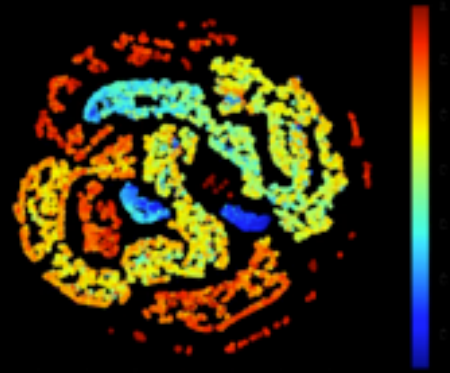
D. Bhowmik, M.T. Young, S. Gao, A. Ramanathan, BMC Bioinformatics (2019)

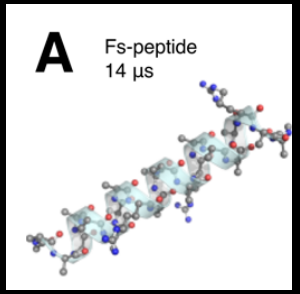
Related work:
Hernandez 17 arXiv,
Noe Nat. Comm. (2018)
Doerr 17 arXiv

CVAE REVEALS METASTABLE STATES IN PROTEIN FOLDING…









MSM Builder Datasets, Pande group

OUTLINE: ALGORITHMS FOR AI-DRIVEN MD SIMULATIONS

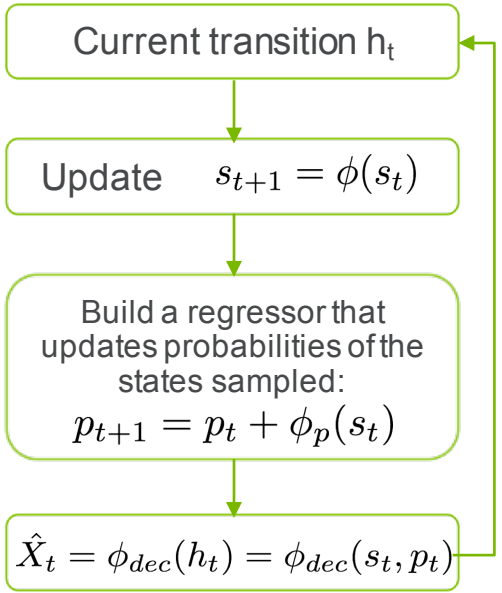
- Building biophysically meaningful, low-dimensional latent representations of simulation data:
  - Deep learning for MD data
  - Convolutional variational autoencoder
- Predicting where we should go next in MD simulations:
  - Building a recurrent autoencoder to predict future step
- Preliminary work on using reinforcement learning to fold proteins

DATA-DRIVEN PROPAGATION AND ESTIMATION FOR SIMULATIONS (1)

- Model the state update as an extrapolation process:
  - Learn the encoding such that there is an update s_t that corresponds to the current transition h_t
  - Multiple options for feature representations including linear, non-linear & hybrid models

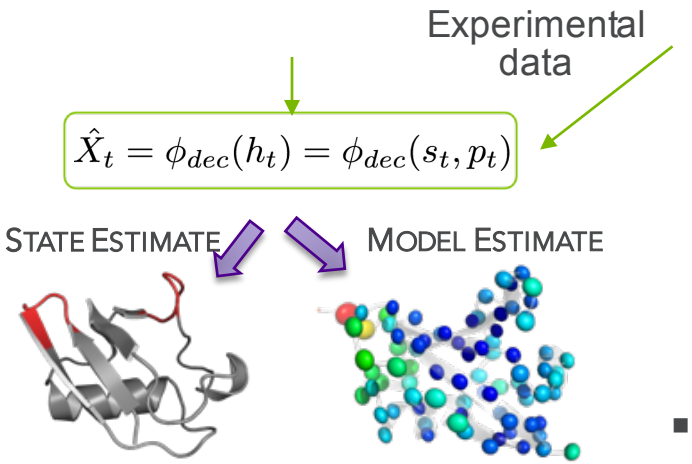
- Evolve the system in the feature space (s_t) using a single layer perceptron regressor
  - Efficient for training and running at local scale

- Successful applications of machine learning based simulations exist for smaller systems[1-3]



[1]S. Ehrhardt, A. Monszpart, N. Mitra, A. Vedaldi, arXiv: 1706.02179 (2017)
[2]S. Ehrhardt, A. Monszpart, N. Mitra, A. Vedaldi, arXiv: 1703.00247 (2017)
[3]J. Thompson, K. Schlachter, P. Sprechmann, K. Perlin, arXiv: 1607.03597v3 (2016)

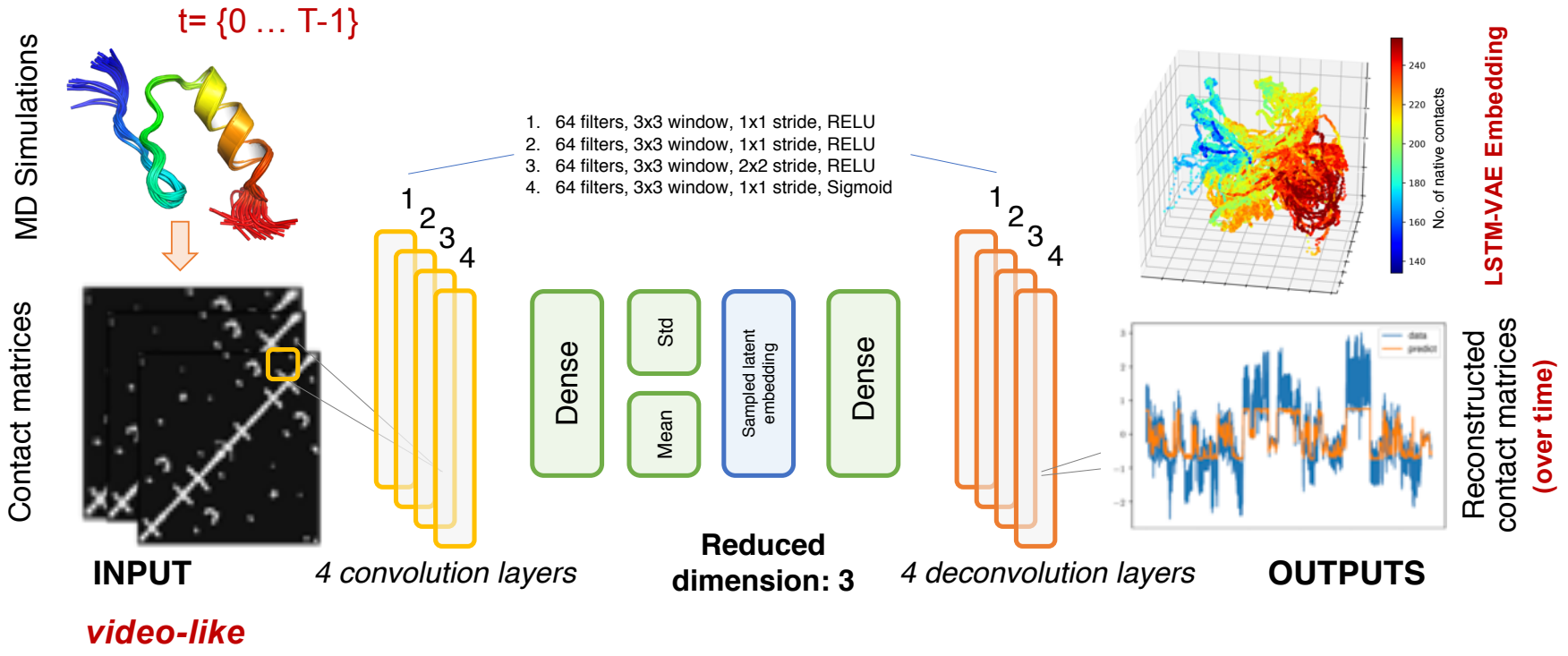
DATA-DRIVEN PROPAGATION AND ESTIMATION FOR SIMULATIONS (2)



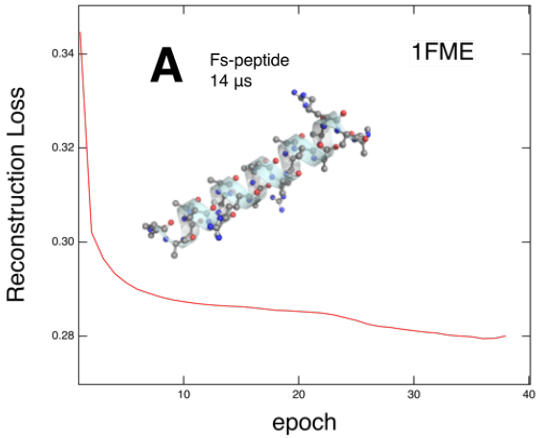
- The decoder provides :
  - **State estimate**: update positions, velocities, and other state variables
    - Current error in state is dictated by what the features have learned from the encoding step
  - **Model estimate**: how "far away" from actual system evaluation is the current state
    - L2 norm from extrapolation
- Decoding step can include experimental observations to prune states

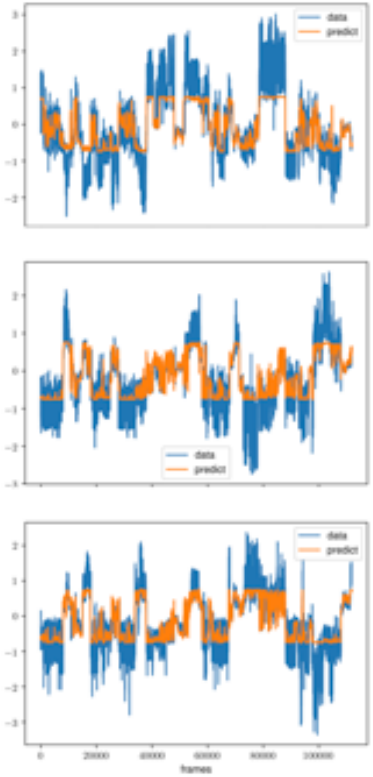
DOES THIS WORK?

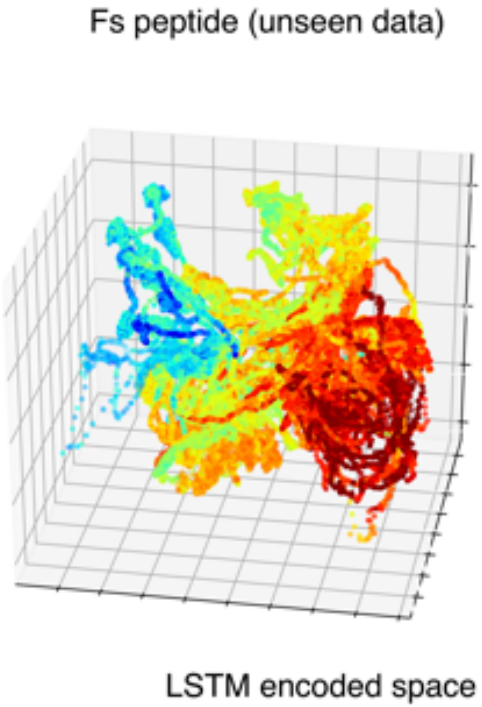
CAPTURING TEMPORAL EVOLUTION IN MD SIMULATIONS…



LSTM-AUGMENTED VAE CAPTURES FS-PEPTIDE TIME-DEPENDENT CHANGES ALONG FOLDING PATHWAYS





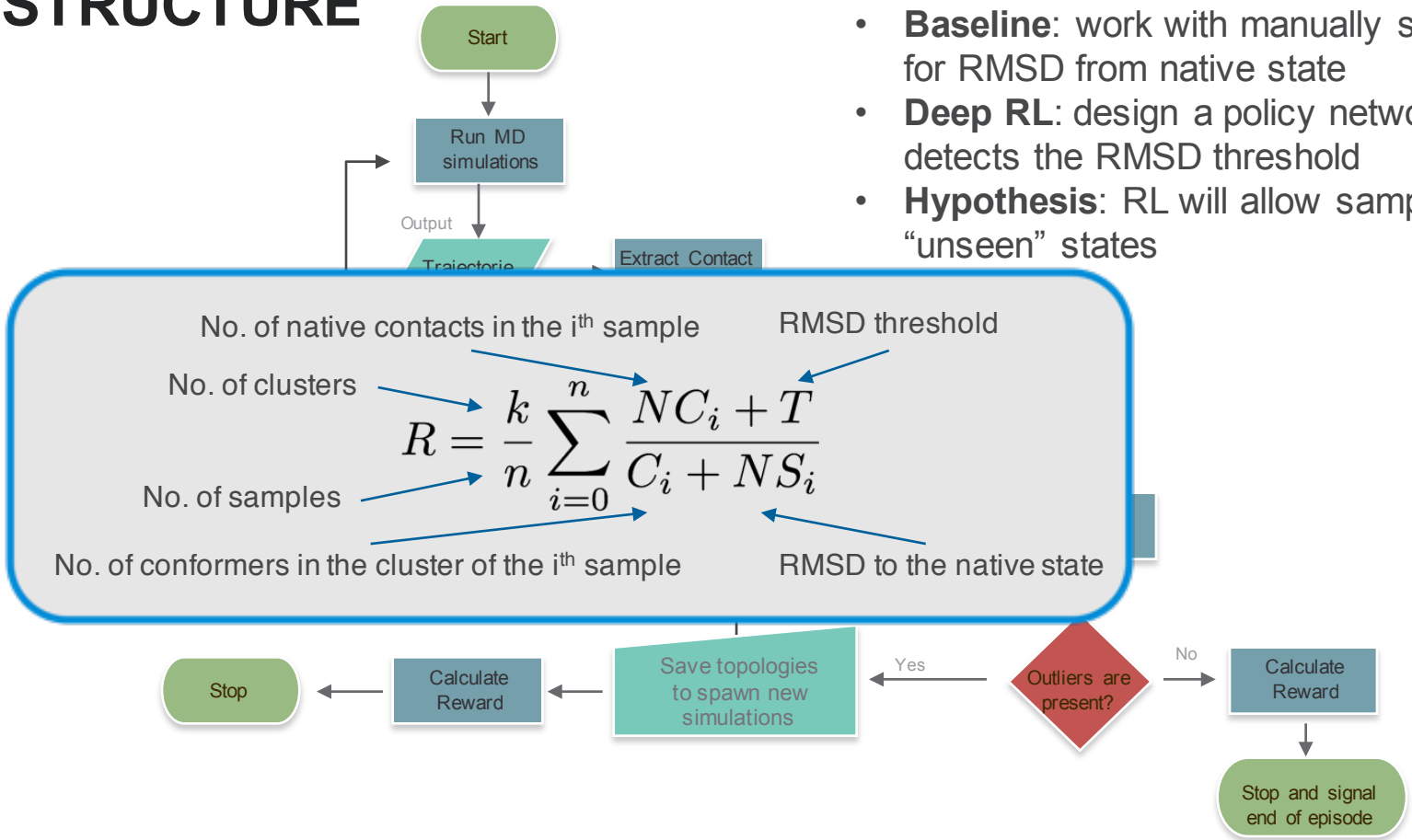


OUTLINE: ALGORITHMS FOR AI-DRIVEN MD SIMULATIONS

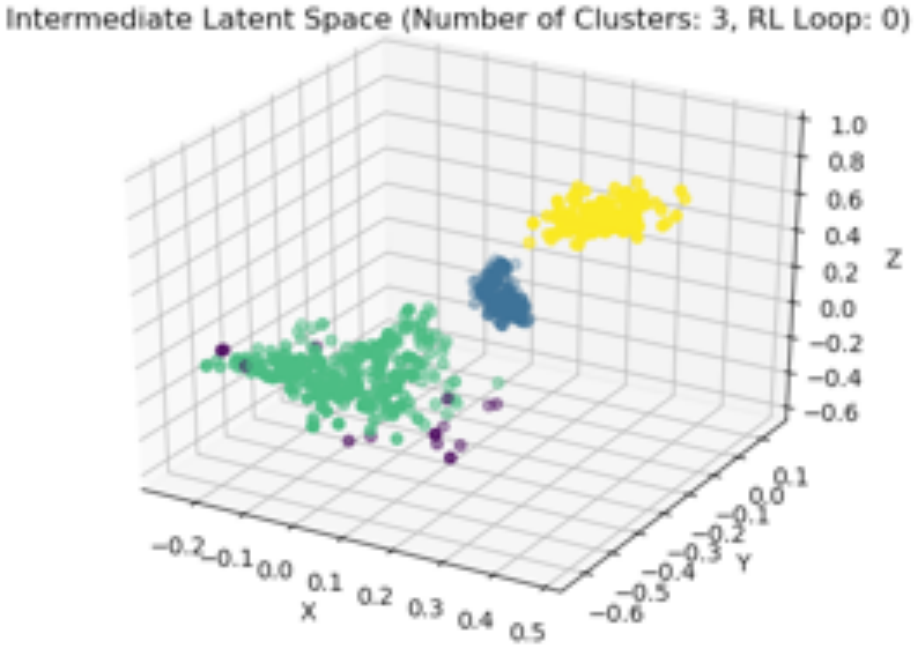
- Building biophysically meaningful, low-dimensional latent representations of simulation data:
  - Deep learning for MD data
  - Convolutional variational autoencoder
- Predicting where we should go next in MD simulations:
  - Recurrent autoencoders for predict future step
- Preliminary work on using reinforcement learning to fold proteins

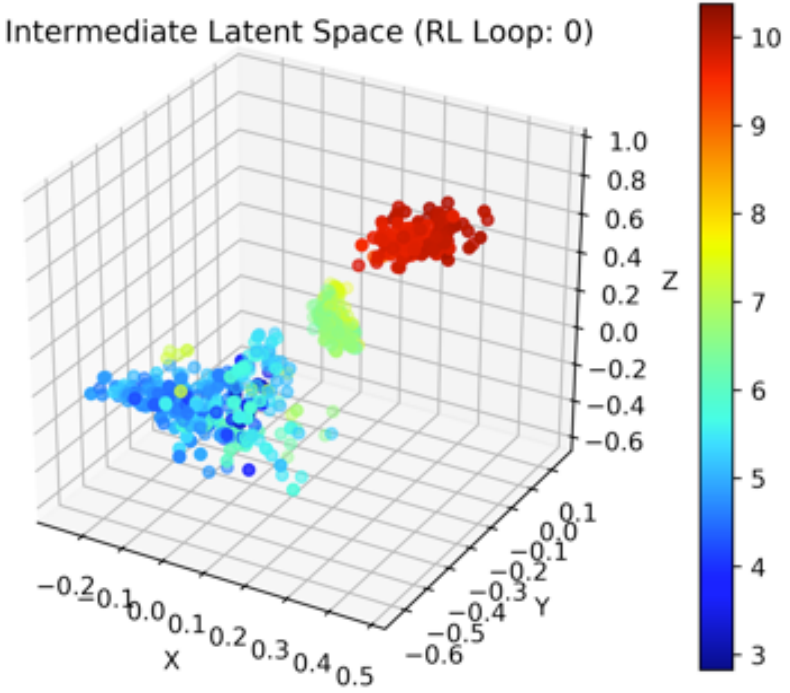
RL-FOLD: A NAÏVE DESIGN BASED ON NATIVE STRUCTURE

- **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

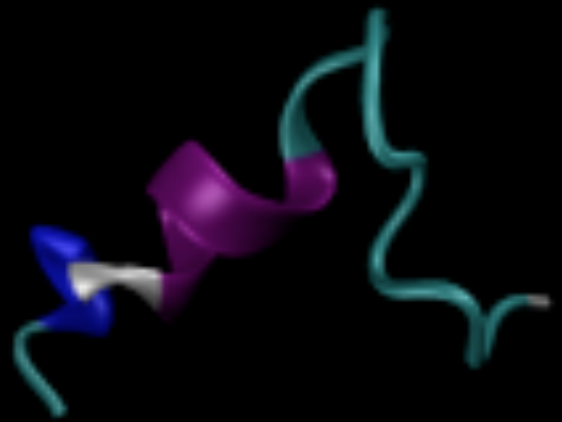


PRE-TRAINED DEEP LEARNING MODEL ALLOWS RL EXPLORES POSSIBLE STATES IN PROTEIN FOLDING





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 tha
- *Re* ic is feasible to spe
  - ls
  - R
  - o
- En e
- *Extensible library:* Molecules to enable a scale with Deep(μ)scope supporting AI-driven MD s

Team

- Debsindhu Bhowmik (ORNL)
- Heng Ma (ORNL)
- Shantenu Jha & the RADICAL team (Rutgers)
- Venkat Vishwanath (Argonne)
- and many summer interns…

Computing Time

- OLCF Early Access on Summit (OLCF)
- ALCF (for testing)
- ALCC Computing Allocation

Funding

- DOE-NCI JDACS4C
- DOE Exascale Computing Project Cancer Deep Learning Environment (CANDLE)

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

THANK YOU!!

ramanathana@anl.gov