DDMD: AI-ACCELERATED ENHANCE SAMPLING FOR MOLECULAR DYNAMICS SIMULATIONS

Heng Ma, Alex Brace, Igor Yakushin, Hyungro Lee, Ian Foster, Shantenu Jha, Arvind Ramanathan

Data Science and Learning Division, Argonne National Laboratory, Lemont, Illinois
Department of Computer Science, University of Chicago, Chicago, IL, Illinois
RADICAL-Lab, Rutgers University, New Brunswick, New Jersey
Computational Science Initiative, Brookhaven National Laboratory, Upton, New York
Bond vibrations
Side-chain rotations/motions
Helix/coil transitions
Higher-order secondary structural transitions
Disorder-to-order transitions
Multivalent interactions
Liquid-liquid phase separation

$10^{-15}$ $10^{-12}$ $10^{-9}$ $10^{-6}$ $10^{-3}$ $10^{0}$ $10^{2}$ Time (s)

Experimental Techniques
NMR
Solution scattering techniques (X-ray/ neutron)
Single-molecule Forster Resonance Energy Transfer (FRET)
X-ray-crystallography and Cryo-electron microscopy
Cryo-TEM

All-atom molecular dynamics (MD)
Ensemble methods and enhanced sampling methods
Coarse-grained MD and other sampling methods
Continuum methods, Multiscale modeling, Theory

Computational Methods
STATISTICAL INFEERENCE: GLUE INFORMATION ACROSS SCALES

Event detection
Dimensionality reduction and clustering
Quantifying conformational transitions

ML and deep learning approaches

Molecular & Macromolecular
Subcellular
Cellular

Å
nm – μm
0.1mm - mm

fs - μs
μs - ms
ms - s

Spatial and Temporal Scales

Standard simulations
Enhanced sampling workflows

AI/ML-driven workflows?
HOW CAN AI/ML HELP?

- Automatic AI/ML inferencing MD workflows
  - Learning the conformational states

- AI/ML surrogate models
  - Learning the dynamics/propagation
ADAPTIVE SIMULATION: ON-THE-FLY ANALYSIS AND DECISION MAKING

- Generate ensemble of simulations in parallel as opposed to one realization of process
  - Strength in numbers

- 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

DDMD - DeepDriveMD

- Simulation – Molecular Dynamics
  - OpenMM/NAMD on GPUs
  - Implicit/explicit solvent

- Aggregation
  - Adios Stream

- Training
  - Convolutional Variational AE
  - Keras/TensorFlow

- Inference
  - Outlier detection, sklearn
  - MDAnalysis

![Diagram](DDMDDiagram.png)
A VARIATIONAL APPROACH TO ENCODE PROTEIN FOLDING WITH CONVOLUTIONAL AUTO-ENCODERS

1. 64 filters, 3x3 window, 1x1 stride, RELU
2. 64 filters, 3x3 window, 1x1 stride, RELU
3. 64 filters, 3x3 window, 2x2 stride, RELU
4. 64 filters, 3x3 window, 1x1 stride, Sigmoid

Related work:
Hernandez 17 arXiv,
Doerr 17 arXiv
Deep clustering of protein folding simulations
D Bhowmik, S Gao, MT Young, A Ramanathan - BMC bioinformatics, 2018
Source code: http://ramanathanlab.org
RUN ENV

Radical EnTK

ADIOS 2 (Adaptable Input Output System)

- High performance I/O Framework
- Suitable for streaming workflow design
- Scalability
USER CASE 1: EXPLOITER

Protein Folding Pathway – BBA

- OpenMM software package
- Amber99sb-ildn force field
- Amber99_obc implicit water model
- 300K Langevin integrator with 2 fs time-step
- 1.0 nm cut-off
- The same starting conf. as Anton runs
- 10-dimension latent space
- dbscan outlier search with RMSD rankings
- 120 simulation runs on CUDA
- 12-hr runs on Summit/Lassen
RESULTS - UC1

• Lowest RMSD: 1.56 Å
RESULTS - UC1

<table>
<thead>
<tr>
<th></th>
<th>RMSD only</th>
<th>RMSD + ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>2.37 ± 0.30</td>
<td>1.81 ± 0.21</td>
</tr>
<tr>
<td>min</td>
<td>1.85</td>
<td>1.55</td>
</tr>
<tr>
<td>max</td>
<td>2.93</td>
<td>2.20</td>
</tr>
</tbody>
</table>
RESULTS - UC1

Sampling efficient 500 conformational states

- DDMD (ML + RMSD) samples significantly more states
  - Faster sampling overall
- ML-only approach samples the near-by space, as there is no guide physical property
- MD is slow and steady with inferencing, and dependent on the simulation condition.
RESULTS - UC1

Embeddings:
• The first 3 of 10 latent dimensions
• Clustering of conformers correlating to their RMSD
• The embedding space retains the configurational information of 3D protein models
• Similar conformations are packed as close neighbors
• Outlier detection picks out rare sampled conformations
RESULTS - UC1

Why is it working?

- Restarting points
  - ML-only: black/green stars
  - greedy RMSD: red stars
  - ML+RMSD: green star

- Playout
  - ML-only: keep sampling low-sampled regions
  - greedy RMSD: might trapped in metastable/misfold
  - ML+RMSD: More prone to make good sampling decisions
UC2

Explorer: 120 PLPro proteins with different ligands

- OpenMM software package
- Amberff14sb force field with tip3p explicit water model
- 300K Langevin integrator with 2 fs time-step
- 1.0 nm cut-off
- 10-dimension latent space
- DBScan and LOF outlier search
- 120 simulation runs on CUDA platform
- 12-hr runs on Summit/Lassen
RESULTS - UC2

Sampling behavior
**RESULTS - UC2**

**Sampling behavior**

- DDMD automatically picks the "interesting" conformer/ligand to simulation
- It reverts to previous ligands when the current runs are sufficiently modelled
UC3

Large multimer system

- Covid-19 nsp RNA system (7egq), 6760 residues, 1.09M atoms

A. Trifan, …, A. Ramanatha, SC21, GB covid-19 finalist
UC3
New env

- **Balsam**
  - Job launching on multiple HPC platforms
  - From your own laptop

- **Globus**
  - Data transfer platform

- **NAMD3**
  - Parallel multiple-GPU MD engine

- **(Fluctuating Finite Element Analysis) FFEA**
  - Blob simulation

- **(Graph Neural Operator) GNO**
  - ML PDE solver learning MD/FFEA trajectories
Multi-scale models

Making the tools do what they’re good at

• FFEA
  • Directly from cryo-EM blobs
  • Global fluctuation info for ML
• All-atom MD
  • Local interactions info for ML
• Hierarchical AI models
  • Record the input info from simulations
  • Generate/identify novel states for MD simulations
  • Update FFEA interfacial interactions or blob rotation
RESULTS - UC3 – MD

A

nsp10A  nsp14A
nsp9A    nsp13-1A
nsp12A  nsp8-1A
nsp14B  nsp13-2A
nsp10B  nsp8-2A

B

RNA unwinding

Initial Structure  Final Structure
The MD and FFEA results correspond to the CryoEM results.

Fluctuations at the nsp13

FFEA mesh interaction needs improvement, at the nsp10-nsp14 blob

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Non-equilibrium Sampling Methods</th>
<th>GFLOPs/step</th>
<th>Total sampling (μs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equilibration (pre-unwinding)</td>
<td>None</td>
<td>24.9</td>
<td>2.107</td>
</tr>
<tr>
<td>RNA unwinding</td>
<td>Extrabonds, Distance colvar, MDFD grid</td>
<td>25.9</td>
<td>0.004</td>
</tr>
<tr>
<td>RNA post-unwinding</td>
<td>Atom restraint, MDFD</td>
<td>11.04</td>
<td>1.904</td>
</tr>
<tr>
<td>RNA post-unwinding</td>
<td>None</td>
<td>11.04</td>
<td>3.916</td>
</tr>
</tbody>
</table>
UC3
MD surrogate models

- (A) FFEA latent embeddings separates various states
- (B) GNO perturbation in latent space from low-RMSD to high-RMSD
- (C) Overlay of high- and low- RMSD states from GNO

Z. Li, ..., A. Anandkumar, arXiv:2003.03485v1
CONCLUSIONS AND FUTURE WORK

- DeepDriveMD utilizes AI/ML to pivot MD simulation runs to speed up BBA folding
- Modes: Exploiter, Explorer, …
- It is flexible to incorporate different frameworks and implement multiple sampling strategies.
  - Better ML models?

- It needs more validations to understand the underlying mechanism
- Possible MD surrogate models from AI/ML, GNO or other sequence models
ACKNOWLEDGEMENT

- The teams:
  - Arvind Ramanathan, Rick Stevens, Alex Brace (CELS, UChicago)
  - Ian Foster, Igor Yakushin (DSL)
  - Anda Trifan, Defne Ozgulbas, John Stone (UIUC)
  - Venkatram Vishwanath (ALCF)
  - Shantenu Jha (Rutgers/BNL)
  - Sarah Harris (Leeds), Geoffrey Wells (UCL)
  - Lillian Chong, Anthony Bogetti (UPitt)
  - Anima Anandkumar, Zongyi Li (Caltech)
  - And many more

- Computing support:
  - ALCF, OLCF, LLNL computing
  - XSEDE: TACC, NERSC, SDSC,

- Funding:
  - DOE NVBL, Exascale Computing Project
  - ANL LDRD
THANK YOU

heng.ma@anl.gov