AdaptiveMD & PySFD: Large-Scale Computations and Analysis of Molecular Dynamics Simulations

John Ossyra*[,1,2], Sebastian Stolzenberg*[,3], Eugen Hruska[4], Vivek Balasubramanian[5], Manuel O. Maldonado[5], Matteo Turill[5], Shantenu Jha[5,6], Cecilia Clementi[4], Frank Noe[3], Jeremy Smith[1,2]

* contributed equally

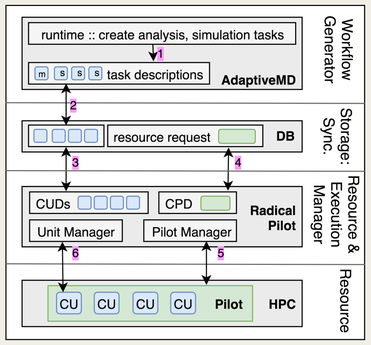
[1]UT-ORNL Graduate School of Genome Science and Technology
[2]UT-ORNL Center for Molecular Biophysics
[3]Freie Universität Berlin Computational Molecular Biology
[4]Rice University Department of Chemical and Biomolecular Engineering
[5]Rutgers University Department of Electrical and Computer Engineering
[6]Brookhaven National Laboratory

BIP149
INCITE 2018

Biomolecular Simulation Sampling Problem

- Regular Molecular Dynamics (MD) simulations fail to reach biologically relevant length and time scales due to large energy barriers separating relevant conformational states. To reach these scales, we have developed:
- **AdaptiveMD,** a novel, supercomputer-independent Python package which guides swarms of trajectories with adaptive sampling algorithms and Markov State Models.
- **PySFD,** a novel, multi-processing Python package to obtain mechanistic information from the resulting MD data sets by efficiently selecting significant differences in user-defined molecular features among many simulated states.

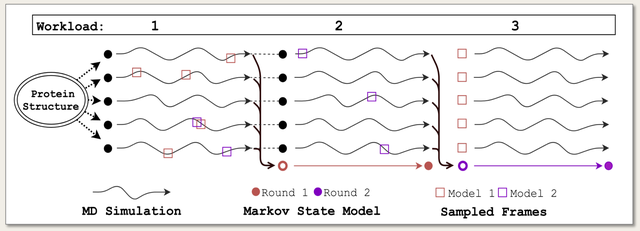
AdaptiveMD Workflow Platform



Architecture & Sequence

1) New task descriptions created from a workflow function at runtime
2) Tasks synchronize to storage, along with a request to specify the target resource and job parameters
3) Radical Pilot translates task to compute unit descriptions (CUD)
4) and requests to compute pilot descriptions (CPD)
5) Pilot enqueued in the HPC scheduler becomes active
6) and executes tasks in compute units

- 3 distinct environments distributed over network layers
  - *Application* environment, *storage* layer, and *resource* environment
- Storage layer database synchronizes all workflows to all layers

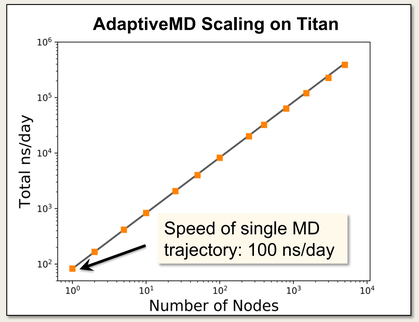


Asynchronous Workflow of Uncoupled Tasks

- MSM data from the analysis (filled circles) synchronized to application
  - Labels trajectory frames with properties:
    - (micro/macro) state
    - estimated equilibrium population.
- Properties used to sample the data and define new workloads

- Workflows consist of **simulation** and **analysis** tasks
  - Application has: *view* of trajectory data, *actual* MSM data from analysis
  - Task Kernels: Simulation- OpenMM, Analysis- PyEMMA

AdaptiveMD Workflow Performance



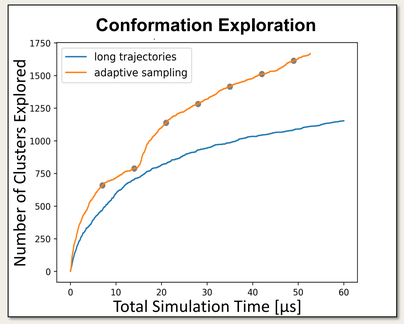
Weak Scaling of Workflow

- 90% efficiency up to 5000 tasks
- Limited by heavy database use
- Some redesign to limit number of connections, transfers
- 1 task per node
- Simulation tasks fully utilize Titan's GPU capability.

- We currently utilize Titan to execute workloads of up to 4000 tasks
  - Application on login node, storage on a VM in ORNL OpenStack Cloud
- Distributed platform design allows multiple concurrent workflows
  - Target resources best suited for workloads

Effect of Adaptive Sampling

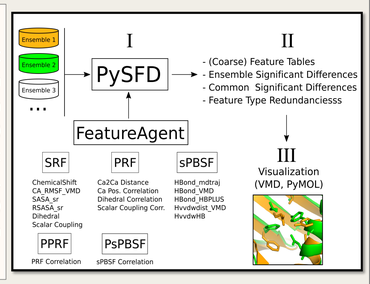
- NTL9 protein folding system
- Adaptive sampling (40 replicates, 150ns trajectory length)
  - 7 restarts (grey dots)
- Long trajectory (40 replicates)
- RMSD clustering of Cα positions from both datasets
- Conformations explored against aggregate simulation time



- Adaptive sampling increases rate of conformation exploration
- Generates large dataset with robust sampling in many states

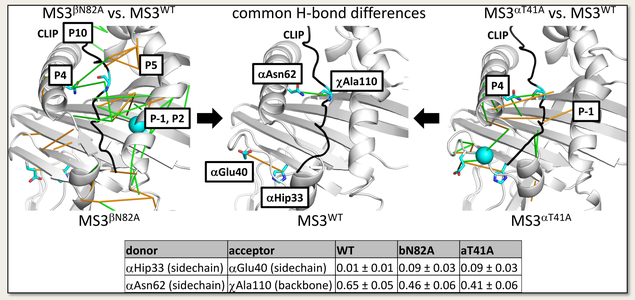
PySFD Data Analysis Workflow

Significant Feature Differences Analyzer:
(I) Given input trajectories realizing different molecular states (ensembles) and a FeatureAgent-derived class, the PySFD main class occupies multiple CPUs to compute (II) and visualize with VMD and PyMOL (III): (coarse-grained) feature tables per state, feature type redundancies, and (common) significant feature differences between (groups of) states.



PySFD Applied on 3 Variants of MHCII

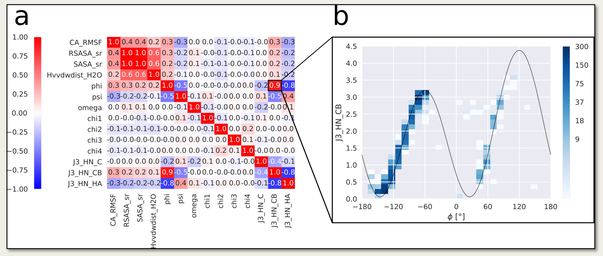
- In the adaptive immune system, Major Histocompatibility Complex class II (MHCII) proteins recognize antigenic (e.g. viral) peptides via peptide exchange under possible catalysis of the HLA-DM protein.
- βN82A (αT41A) mutant promotes uncatalyzed (catalyzed) exchange



donor	acceptor	WT	bN82A	aT41A
αHip33 (sidechain)	αGlu40 (sidechain)	0.01 ± 0.01	0.09 ± 0.03	0.09 ± 0.03
αAsn62 (sidechain)	χAla110 (backbone)	0.65 ± 0.05	0.46 ± 0.06	0.41 ± 0.06

(Common) Significant differences in hydrogen bond frequencies between MHCII's ground states of the βN82A and αT41A mutants with respect to wildtype (WT)

- Ground states (MS3): bound CLIP peptide (black), pts. of mutation (cyan spheres)
- Significantly more (orange)/ less (green) frequent in βN82A/αT41A vs. WT
- Two significantly different H-bonds are **common** to both state comparisons, as tabulated below the molecular graphics (χ indicates CLIP).



Potential feature type redundancies detected with PySFD

a. Heat map of Spearman correlation coefficients between SRF feature types along individual residues and ensemble means of nine meta-stable states.
b. Redundant feature $^3J_{HN\text{-}C\beta}$ vs. ϕ: heat map showing sampled feature values overlaid with a black plot of the Karplus relation.

Acknowledgements

This work has been supported by the Deutsche Forschungsgemeinschaft (DFG) "Eigene Stelle" grant no. STO 1177/1-1. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory (project IDs: BIP103 and BIP149), which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

References

[1] J. Ossyra. *AdaptiveMD* (in preparation), https://github.com/markovmodel/adaptivemd
[2] S. Stolzenberg. *PySFD: Molecular Mechanistic Clues from Significant Feature Differences selected among many Simulated Ensembles* (submitted), https://github.com/markovmodel/PySFD

THE UNIVERSITY OF TENNESSEE KNOXVILLE
Freie Universität Berlin
Center for Molecular Biophysics UT / ORNL
RICE