Cyberinfrastructures for In Situ Data Analytics for Next Generation Molecular Dynamics Workflows

Michela Taufer[1], Michel Cuendet[2], Ewa Deelman[3], Trilce Estrada[4], Rafael Ferreira Da Silva[2], Harrel Weinstein[2]

[1] University of Tennessee Knoxville
[2] Weill Cornell Medical College of CORNELL University
[3] University of South California
[4] University of New Mexico

Molecular dynamics simulations studying the classical time evolution of a molecular system at atomic resolution are widely recognized in the fields of chemistry, material sciences, molecular biology and drug design; these simulations are one of the most common simulations on supercomputers. Next-generation supercomputers will have dramatically higher performance than do current systems, generating more data that needs to be analyzed (i.e., in terms of number and length of molecular dynamics trajectories). The coordination of data generation and analysis cannot rely on manual, centralized approaches as it does now. This interdisciplinary project integrates research from various areas across programs such as computer science, structural molecular biosciences, and high performance computing to transform the centralized nature of the molecular dynamics analysis into a distributed approach that is predominantly performed in situ. Specifically, this effort combines machine learning and data analytics approaches, workflow management methods, and high performance computing techniques to analyze molecular dynamics data as it is generated, save to disk only what is really needed for future analysis, and annotate molecular dynamics trajectories to drive the next steps in increasingly complex simulations' workflows.

This project tackle the data challenge of data analysis of molecular dynamics simulations on the next-generation supercomputers by (1) creating new in situ methods to trace molecular events such as conformational changes, phase transitions, or binding events in molecular dynamics simulations at runtime by locally reducing knowledge on high-dimensional molecular organization into a set of relevant structural molecular properties; (2) designing new data representations and extend unsupervised machine learning techniques to accurately and efficiently build an explicit global organization of structural and temporal molecular properties; (3) integrating simulation and analytics into complex workflows for runtime detection of changes in structural and temporal molecular properties; and (4) developing new curriculum material, online courses, and online training material targeting data analytics. The project's harnessed knowledge of molecular structures' transformations at runtime can be used to steer simulations to more promising areas of the simulation space, identify the data that should be written to congested parallel file systems, and index generated data for retrieval and post-simulation analysis. Supported by this knowledge, molecular dynamics workflows such as replica exchange simulations, Markov state models, and the string method with swarms of trajectories can be executed ?from the outside? (i.e., without reengineering the molecular dynamics code).