# IMPROVING NWCHEM SCALABILITY USING THE DATASPACES FRAMEWORK

TN

**Allocation:** Innovation and Exploration/525 Knh
**PI:** Gregory Bauer[1]
**Collaborators:** Victor Anisimov[1], Manish Parashar[2], Melissa Romanus[2]

[1]National Center for Supercomputing Applications
[2]Rutgers University

## EXECUTIVE SUMMARY

Molecular Dynamics (MD) is the major computational technique to bridge the gap between experiment and theory in materials science, engineering, and biomedical research. However, the predictive ability of MD simulations strongly depends on the quality of underlying parameters. The purpose of the present phase of this project is to develop a parameter optimization tool for the AMBER classical force field for DNA, with the potential for extending the underlining methodology to optimization of other popular force fields. The project employs the previously generated data set of experimental quality base–base interaction energies prepared by conducting high-level quantum-mechanics CCSD(T) computations in NWChem on molecular clusters extracted from experimental crystallographic data for DNA bases. The parameter optimization procedure performs a grid-based scan on a set of parameters by trying all their possible combinations, computing the interaction energy for each grid point using the Amber force field, and comparing the result to the reference interaction energy. The computation runs in parallel and returns a set of parameters that best reproduces the target data.

## RESEARCH CHALLENGE

The present computational challenge is handling a combinatorial explosion in problem size that accompanies the task of identifying the global minimum in parameter space. The natural solution to this challenge is to employ parallelization. However, the use of a large number of parallel processes in the grid search exacerbates the I/O load on the filesystem since each process frequently performs read and write operations. That places a limit on the number of parallel tasks that can be practically used in the computation without overloading the filesystem. When the grid search is done, each process carries a multidimensional array of results that associates the point in the parameter space with the optimization function. The aggregate distributed array holds billions of records for an average parameter optimization problem. Sorting an array of such size to determine the promising parameter sets for further analysis represents a practically unsolvable problem that requires a creative solution.

BI

## METHODS & CODES

Generation of target data for parameter optimization uses a set of in-house scripts to extract molecular clusters from the experimental crystallographic data of DNA bases. Quantum mechanical computations, which follow, determine the base–base interaction energy in the molecular clusters. The computation employs the previously optimized CCSD(T) method [1] in the NWChem package [2]. A scalable tool to optimize Lennard–Jones parameters in the AMBER (Assisted Model Building with Energy Refinement) force field to fit the parameters to intermolecular interaction energies for experimental geometry of monomers has been developed. It has been tested to run on 16,384 XE nodes using 32 cores per node resulting in the use of 524,288 processing units on Blue Waters. The optimization tool generates an adjustable number of alternative parameter sets of comparable quality for further testing in molecular dynamics simulations.

## RESULTS & IMPACT

This project introduces a procedure for systematic improvement of classical force field by determining the global minimum in the parameter space for an expandable set of the training data. The beneficiary of the optimized parameter set is the entire molecular dynamics community. As the number and quality of the training data increase with time, rerunning the parameter optimization tool will deliver the improved parameter set. The developed fractional parallel sorting procedure drastically reduces time spent in sorting as well as the required RAM per node. The use of RAM disk for read / write operations on compute nodes eliminates the filesystem overhead and makes the code applicable to compute systems beyond Blue Waters' size.

## WHY BLUE WATERS

Blue Waters, with its fast interconnect and large memory per core, is unique in its ability to conduct CCSD(T) computations of molecular systems encountering a thousand basis functions, which is vital for the success of the developed parameter optimization procedure. Since the parameter optimization procedure is extremely resource demanding, the availability of large numbers of nodes is essential for the exhaustive exploration of parameter space.





*Continued from page 213*

Observed by Molecular Dynamics. *J. Chem. Theory Comp.*, 12 (2016), pp. 3382–3389.

Galindo-Murillo, R., et al., Assessing the current state of Amber force field modifications for DNA. *J. Chem. Theory Comp.*, 12 (2016), pp. 4114–4127.

Heidari, Z., et al., Using Wavelet Analysis to Assist in Identification of Significant Events in Molecular Dynamics Simulations. *J. Chem. Inf. Model.*, 56 (2016), pp. 1282–1291.

Hao, Y., et al., Molecular basis of broad-substrate selectivity of a peptide prenyltransferase. *PNAS*, 113 (2016), pp. 14037–14042.

Hayatshahi, H.S., et al., Computational Assessment of Potassium and Magnesium Ion Binding to a Buried Pocket in the GTPase-Associating Center RNA. *J. Phys. Chem. B*, 121 (2017), pp. 451–462.

Zgarbova, M., et al., Influence of BII Backbone Substates on DNA Twist: A Unified View and Comparison of Simulation and Experiment for all 136 Distinct Tetranucleotide Sequences. *J. Chem. Info. Model.*, 57 (2017), pp. 275–287.

Wang, Y., et al., Application of thiol-yne/thiol-ene reactions for peptide and protein macrocyclizations. *Chemistry*, 23 (2017), pp. 7087–7092.

Hayatshahi, H.S., C. Bergonzo, and T.E. Cheatham III, Investigating the ion dependence of the first unfolding step of GTPase-associating center ribosomal RNA. *J. Biomol. Struct. Dyn.*, 1:11 (2017), pp. 243–253.

Bergonzo, C., and T.E. Cheatham III, $Mg^{2+}$ binding promotes SLV as a scaffold in Varkud Satellite Ribozyme SLI-SLV kissing loop junction. *Biophys. J.*, 113 (2017), pp. 313–320.

Galindo-Murillo, R., and T.E. Cheatham III, Computational DNA binding studies of (-)-epigallocatechin-3-gallate. *J. Biomol. Struct. Dyn.*, 3 (2017), pp. 1–13 (2017).

Hayatshahi, H.S., N.M. Henriksen, and T.E. Cheatham III, Consensus conformations of dinucleotide monophosphates described with well-converged molecular dynamics simulations. *J. Chem. Theory Comp.*, 14 (2018), pp. 1456–1470.

Cornillie, S.P., et al., Computational modeling of stapled peptides toward a treatment strategy for CML and broader implications in the design of lengthy peptide therapeutics. *J. Phys. Chem. B*, 122 (2018), pp. 3864–3875.

Galindo-Murillo, R., T.E. Cheatham III, and R.C. Hopkins, Exploring potentially alternative non-canonical DNA duplex structures through simulation. *J. Biomol. Struct. Dyn.*, (2018) DOI:10.1080/07391102.2018.1483839.