HIGH-THROUGHPUT MATERIALS MODELING OPTIMIZATION

Allocation: Illinois/796.196 Knh
PI: Dallas R. Trinkle[1]
Co-PIs: Dan Katz[2], Joshua Vita[1]

[1]University of Illinois at Urbana–Champaign
[2]National Center for Supercomputing Applications

EXECUTIVE SUMMARY

Empirical potentials are "course-grained" models of atomic interactions and are fundamental to materials modeling. They allow molecular dynamics simulations of processes involving 10^6–10^9 atoms and timescales of nano- to microseconds or longer, and are necessary for both length- and time-bridging methods that span orders of magnitude in scale. Their optimization to reproduce computationally demanding quantum mechanics-based simulation methods is a significantly challenging problem.

Recently, the researchers developed a new approach that relies on a combination of Bayesian sampling of potential parameters [1] with the optimization of the fitting database [2]. The team's algorithm optimizes the target structures and properties, as well as their "weight," to guide the optimization of a potential to make accurate predictions [3]. This automated approach can work both for predictions where experimental or theoretical guidance is missing by including related structures and also to determine when an empirical potential form may be too limited to capture the predictions of interest.

RESEARCH CHALLENGE

The algorithm to optimize a coarse-grained empirical potential (Fig. 1) has recently been demonstrated in the team's publications. Recently, they achieved the next step to reach increased complexity and, hence, significantly greater impact across materials science, physics, chemistry, and biology: improved parallelization of the algorithm to reach large scale. The algorithm relies on Bayesian sampling of parameter space to determine optimal parameters along with error estimates for predictions from the model. The parameters are optimized against a "fitting database": a selection of structures with density functional theory (DFT) energies and force, and with relative weights capturing the importance of each entry. The database is optimized by using a genetic algorithm over the weights. At the center of this algorithm is their massively parallel evaluation engine that takes a list of structures (atomic positions and chemistries) and a large vector of parameters θ (the spline values) for the empirical potential, and evaluates the energies and forces for each parameter. As the researchers include more structures—through prediction of new structures using both DFT and the empirical parameter optimization—and more parameters, the need for a massively parallel approach becomes apparent. For example, if the team were to consider 10^6 empirical potential parameters applied to 10^3 structures with 10^3 evaluations (energy, forces, and derivatives of predictions with respect to parameters), although each individual calculation requires less than a second to complete, there are approximately 10^9 calculations providing about 8 terabytes of data to be used simultaneously; the complete calculation requires ~10^3 cores for memory. Increasing the number of structures by one order of magnitude and the number of parameters by two orders of magnitude increases the scale to approximately 10^6 cores with about 8 petabytes of data. These calculations are embarrassingly parallel and are utilized within a master–worker approach, but the large amount of simultaneous data is spread across workers. By moving to these larger scales, the structural and energy landscape of an empirical potential can be fully analyzed for accurate and predictive empirical potentials. This transforms the problem of coarse-grained parameter optimization into a "Big Data" problem that is of the scale appropriate for a machine like Blue Waters, and provides for a big impact.

METHODS & CODES

This project uses the research team's newly developed parallel evaluation engine, implemented in Python. The code continues to be in active development and is available through Github (https://github.com/TrinkleGroup/s-meam). The underlying parallel algorithm is worker–manager, where individual workers are tasked with evaluating forces or energies for a specific structure; sets of parameters can be passed to a given worker and the forces or energies sent back to the manager. At the beginning of a run, each worker analyzes its structure to convert the spline calculations into vector-matrix operations for efficient evaluation. This helps to keep each worker's evaluation for one parameter set efficient but also permits even faster evaluation of sets of parameters through vectorized operations. The code uses NumPy and SciPy along with Message-Passing Interface for communication. Current runs of up to 512 cores have shown that the calculations—using a genetic algorithm for optimization—have been compute-bound rather than I/O-bound. All development of the algorithm has been on Blue Waters.

RESULTS & IMPACT

Generating a highly efficient and massively parallel materials modeling optimization engine will enable new approaches to the development of empirical potentials that leverage machines at Blue Waters' scale. The research team is designing a general framework that is maximally efficient for large sets of parameters to be evaluated against a fixed set of candidate structures. This computational problem is related, but distinct, from the evaluation typically needed for molecular dynamics calculations where a single parameter set is evaluated against a very large number of structures. The computational engine will be used to take advantage of genetic algorithms for parameter optimization, Monte Carlo evaluation of Bayesian estimates of uncertainty, and cyclic improvement of databases, but other optimization schemes (such as Pareto optimization) could be considered as well.

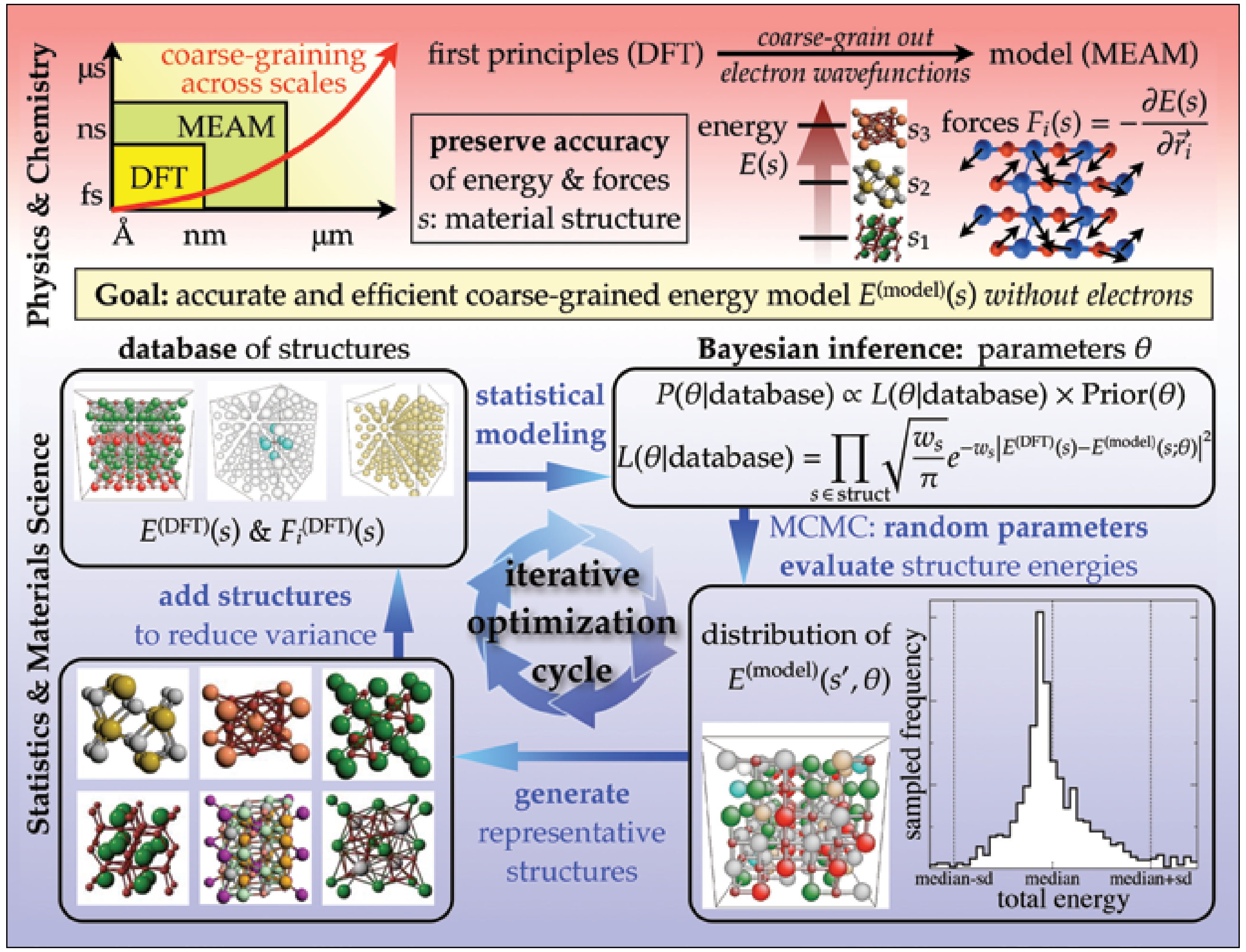
Physics & Chemistry
μs
ns
fs
Å
nm
μm
coarse-graining across scales
MEAM
DFT
first principles (DFT)
coarse-grain out electron wavefunctions
model (MEAM)
preserve accuracy of energy & forces
s: material structure
energy E(s)
s3
s2
s1
forces $F_i(s) = -\frac{\partial E(s)}{\partial \vec{r}_i}$
Goal: accurate and efficient coarse-grained energy model $E^{(\text{model})}(s)$ without electrons
Statistics & Materials Science
database of structures
$E^{(\text{DFT})}(s)$ & $F_i^{(\text{DFT})}(s)$
statistical modeling
Bayesian inference: parameters θ
$P(\theta|\text{database}) \propto L(\theta|\text{database}) \times \text{Prior}(\theta)$
$L(\theta|\text{database}) = \prod_{s \in \text{struct}} \sqrt{\frac{w_s}{\pi}} e^{-w_s|E^{(\text{DFT})}(s)-E^{(\text{model})}(s;\theta)|^2}$
MCMC: random parameters
evaluate structure energies
iterative optimization cycle
add structures to reduce variance
distribution of $E^{(\text{model})}(s',\theta)$
sampled frequency
median-sd
median
median+sd
total energy
generate representative structures

Figure 1: The parallel Bayesian optimization for coarse-graining complex interactions and directed knowledge discovery cuts across materials science, statistics, physics, and chemistry. While the research team's first application of this method is for the modeling of materials at atomistic length scales—a general problem—the approach for predictive coarse-graining will be useful for other applications. The iterative optimization cycle produces a dual representation of the coarse-grained information and an accurate, predictive model of material properties.

WHY BLUE WATERS

The computational engine is designed specifically to leverage massively parallel architecture with a worker–manager structure; access to Blue Waters has been instrumental for the implementation and testing of the code as well as preliminary runs. This work would have been impossible otherwise.