A Database Lookup to Replace Quantum Mechanical Calculations in QM/MM Monte Carlo Simulations

COLLEGE OF ARTS AND SCIENCES CHEMISTRY
COLLEGE OF ARTS AND SCIENCES COMPUTER SCIENCE
MIAMI

Sloan Atkins, Jaime Mizrahi, Dr. Orlando Acevedo

Department of Chemistry University of Miami, FL

Introduction

BOSS (Biochemical and Organic Simulation System) performs mixed quantum and molecular mechanics (QM/MM) simulations couped to Monte Carlo (MC) sampling and free energy perturbation (FEP) to model reaction mechanisms in solution. Throughout the entire simulation, BOSS calls Gaussian 16 (G16), an external QM code, to compute the energy and partial charges associated with the reacting solute configuration thousands to millions of times. This repeated QM calling is the primary computational cost of the method. Therefore, we sought to replace these per-configuration QM calculations with a pre-built database of geometries and their associated energies and charges, so that a simulation can query stored values instead of computing new ones at each step.

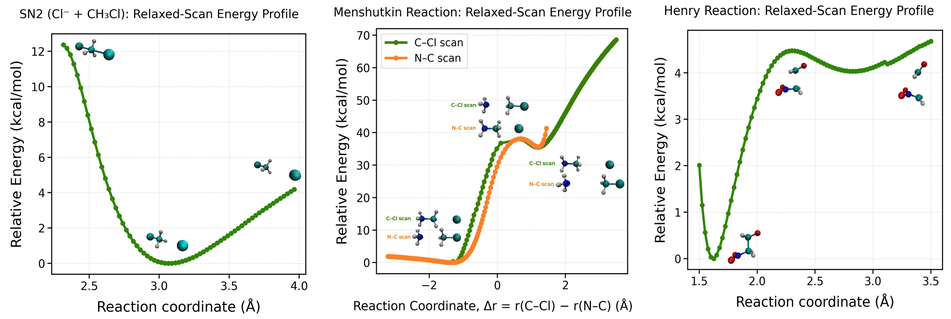
Objective

Our project builds a reaction-agnostic pipeline that generates a reaction-energy database for a given reaction coordinate, using geometries and energies and charges in place of the calculations BOSS would otherwise perform itself. In this workflow, the commercial software G16 has been replaced with the open-source ORCA QM package. This pipeline is paired with a Python implementation of BOSS, built separately, designed to query that database rather than call a QM code during the simulation. The database pipeline has been built for three unique reactions (S_N2, Menshutkin, and Henry) and is currently being extended to a fourth (Diels-Alder).

Molecule Generator

The relaxed scan performs a series of constrained geometry optimizations across the reaction coordinate, holding the coordinate fixed at each point while ORCA optimizes the remaining degrees of freedom. Each point is seeded from the previous point's converged geometry, and the scan proceeds bidirectionally from the transition-state geometry, expanding outward towards reactants and products. Two different scanning methods are used, depending on how many reaction coordinates a given reaction requires.

1D scan: For a reaction with a single forming or breaking bond, the reaction coordinate is scanned bidirectionally across all of its bins from the transition-state geometry. This is the scan type used for S_N2, Menshutkin, and Henry.



Representative reaction-coordinate scans generated for the S_N2, Menshutkin, and Henry reactions.

2D scan: For a reaction with two simultaneous forming or breaking bonds, such as Diels-Alder, the two reaction coordinates are scanned in two passes: one coordinate is held fixed while the other is scanned across its full range, and that process is then repeated for every value of the first coordinate.

Once a point converges, its geometry is used to generate a population of conformers rather than a single structure. Each atom is displaced relative to its own bonded parent, identified using covalent-radius bond distances, rather than a single fixed anchor atom, so that distant substituents are perturbed by a chemically appropriate amount. Bond lengths are perturbed by up to ±0.05 Å, angles by up to ±2.0°, and dihedrals by up to ±15.0°. Conformers with any atom pair closer than 0.8 Å are discarded and regenerated, up to fifty times, to keep the final population chemically reasonable.

Orca Pipeline

The ORCA QM pipeline computes a final energy and a set of atomic charges for every geometry produced by the molecule generator. ORCA input files are generated automatically and run in parallel across all available CPU cores, with each job capped at a one-hour timeout and any failures logged separately rather than discarded. Output files are parsed for the final energy and Hirshfeld atomic charges, and those charges are converted to the CM5 model, which corrects Hirshfeld charges using each atom's element and its distance to neighboring atoms. Final energies and both sets of charges are exported to a single CSV file, one row per geometry, for use in the balancing stage.

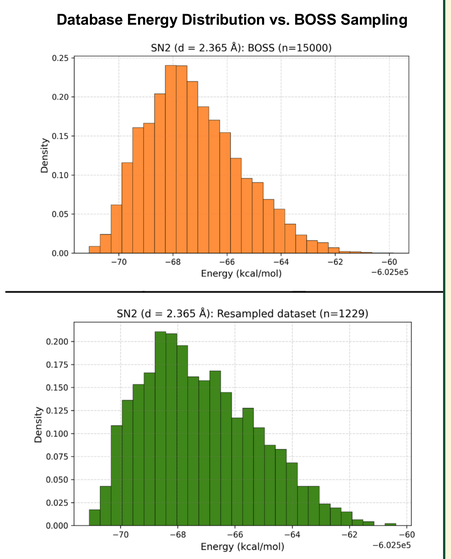
Atom	Hirshfeld	CM5
C	+0.0257	−0.1262
Cl (Incoming)	−0.5821	−0.5894
Cl (Leaving)	−0.5681	−0.5755
H	+0.0470	+0.1029
H	+0.0369	+0.0914
H	+0.0406	+0.0967
Sum	-1.000	-1.000

Hirshfeld and CM5 atomic charges for the SN2 transition state

Data Balancer

Before the exported CSV becomes part of the database, it passes through a dedicated balancing stage with the following steps:

1. **Energy distribution analysis:** The energy distribution is analyzed and binned, and statistical outliers are flagged using both a z-score cutoff and an interquartile-range test.
2. **Chemistry-aware quality scoring:** Every molecule is scored on atomic clashes, net atomic charge, fragment consistency, and energy outlier status, penalizing clashes closer than 0.6 and net charges greater than 0.5 e.
3. **Quality-filtered bin balancing:** Overrepresented bins are trimmed and capped on a per-bin basis using the quality scores from the previous step, so low-quality molecules are removed first.
4. **RMSD filtering:** For reaction-coordinate bins left underpopulated after trimming, each geometry is matched atom-by-atom against reference structures using the Hungarian algorithm, then aligned with a Kabsch rotation to compute a true minimum RMSD. Only candidates within 0.50 Å of a reference are kept.
5. **SMOTE-style oversampling:** As an alternative to RMSD filtering, sparse bins are identified directly, and new molecules are synthesized by interpolating between existing ones within a bin, weighted toward higher-quality neighbors, then passed through the same clash and geometry checks from step 2 before being accepted.



New Database Link & BOSS Changes

1. **Replace the Gaussian 16 (G16) Link: The original BOSS workflow communicated directly with G16 by writing a reaction input file** containing the molecular geometry (XYZ coordinates). G16 then performed a QM calculation on this XYZ structure and returned the computed energy and atomic partial charges (CM5 charge model) to BOSS. To substitute a database of precomputed QM data instead of running a new QM calculation each time, the original G16 interface was replaced with a new database link.
2. **Redesign the Data Exchange: Since the QM properties are stored in the database, the full G16 input is no longer required.** Instead, BOSS was modified to output only the information needed to identify the correct database entry, specifically the reaction coordinate (bin) corresponding to its current position on the potential energy surface. A Python interface then uses this information to locate within the database a random representative molecule and retrieve its pre-computed QM energy, atomic charges, and optimized geometry.
3. **Create a New Data Flow into BOSS: The original G16 interface returned only the QM energy and atomic charges.** Since the database also provides optimized atomic coordinates, BOSS was modified to accept and use these XYZ coordinates. This required creating a new pathway through the code so that all three quantities—energy, charges, and molecular coordinates—were correctly read, stored, and propagated throughout the existing QM/MM workflow.
4. **Validate the Modified Workflow: To ensure the new interface produced accurate behavior within BOSS.** Validation routines were implemented throughout the code to track the QM energy, atomic charges, and XYZ coordinates before and after each major processing step. These checks verified that the imported data remained unchanged as it passed through multiple subroutines and that the modified workflow correctly replaced the original G16 interface without affecting the simulation logic.

Future Prospects

Future work includes validating database-generated results against previously published BOSS studies of these reactions to confirm the database reproduces established experimental energetics rather than simply being internally consistent. Another direction is moving to ORCA more fully across the pipeline, reducing dependence on commercial QM software such as G16. A machine learning (ML) link is also planned, connecting the database to a trained model the same way it currently connects to BOSS, so that the same database used for BOSS queries can also be used to train the ML model. Finally, the appropriate database size for accurate results, whether queried directly or used for training, still needs to be tested rather than assumed.

Acknowledgments

We want to thank our mentor and P.I., Dr. Orlando Acevedo, Dr. Burton Rosenberg, director of the REU Computer Science program, and Eryn Anderson and Leslie Knecht, coordinators of the summer research program. This material is based upon work supported by the National Science Foundation under Grant No. CNS-1949972 and CHE-2102038. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.