Enhancing Quantum Data Analysis through Machine Learning and Numerical Techniques

Aaron Dai[1], Vivienne Pelletier[2], Dr. Christopher Muhich[2]

1 College of Arts and Sciences, University of Virginia, Charlottesville, VA, USA
2 School for Engineering of Matter, Transport, and Energy, Arizona State University, Tempe, AZ, USA

Introduction

Computational quantum chemistry generates valuable insights into chemical processes that cannot be easily obtained through experiment alone, but its computational demands quickly become prohibitive.

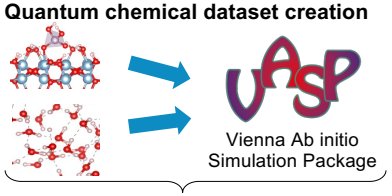
This research aims to accelerate quantum chemical calculations by improving popular machine learning approaches via a new methodology named Aggregated Gaussian Processes (AGP).

This AGP approach improves the efficiency of Gaussian Process Regression (GPR), a widely used machine learning method in quantum chemistry, by allowing the use of parallel data generation with a new methodology for aggregating and optimizing the training data, minimizing the impact of GPR's $O(N^3)$ scaling.

This new approach will be applied to a large dataset of quantum calculations of phosphate interacting with a metal oxide surface in an aqueous solution, which is too large for the application of classical GPR due to the memory requirements. The AGP approach aims to make the construction and application of these large datasets feasible while maintaining the performance of GPR.

Methods

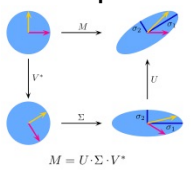
Quantum chemical dataset creation



Vienna Ab initio Simulation Package

Embedding of chemical structures

Basis set optimization



$M = U \cdot \Sigma \cdot V^*$

Singular Value Decomposition/ PCA

AGP Method

Parallel generation of quantum chemical data

↓

Extraction of chemical information from these data

↓

Optimization of dataset via SVD transformation & down selecting over-represented regions of data-space

Results

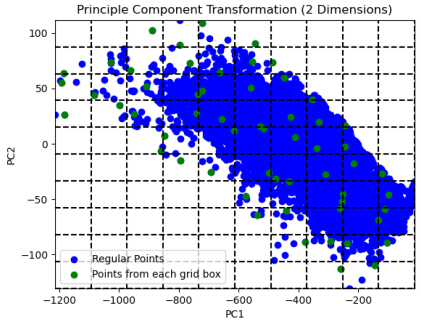


Figure 1. *Principle component transformation of oxygen in two dimensions*

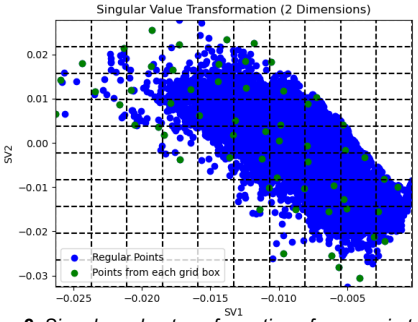


Figure 2. *Singular value transformation of oxygen in two dimensions*

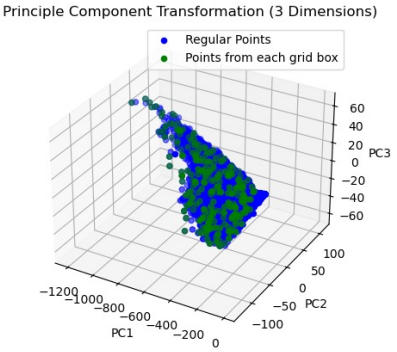


Figure 3. *Principle component transformation of oxygen in three dimensions*

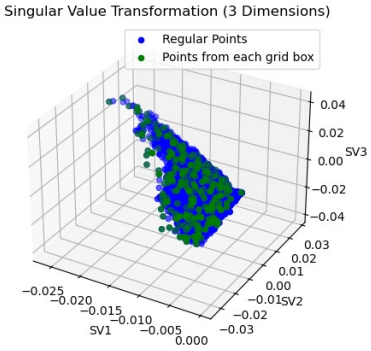
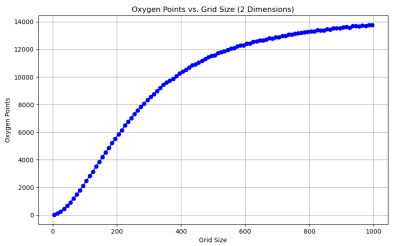
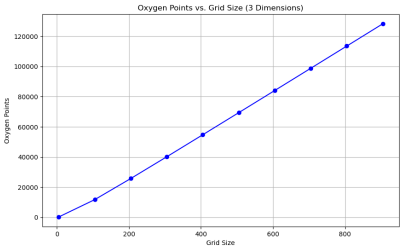


Figure 4. *Singular value transformation of oxygen in three dimensions*





Figures 5 & 6. *Number of oxygen data points compared with the grid size in each singular value decomposition process in two and three dimensions.*

The difference in scaling behavior of the down-selection process in 2 and 3 dimensions is shown in these figures.
In both cases, the curve should follow a logistic shape,
However, with the same range of selection grid sizes, the full sigmoid is seen in the 2D case, but the 3D case remains in the linear region.

Discussion

- ***The full AGP method has not yet been completed,*** this will continue in future work.
- **This project has created a framework** on which the AGP approach can be implemented.
- Understanding the distribution of the principal components of the chemical structures is necessary to the ability to optimize the datasets
- This project has obtained this through the characteristic selection curves shown in Figures 5 & 6

Acknowledgements

The Muhich Lab Group

ASU Research Computing

STEPS REU Program

This material is based upon work supported by the National Science Foundation CBET-2019435.