JOHN T. BARBER, Department of Natural Sciences and Mathematics, West Liberty University, West Liberty, WV, 26074, and RAYMOND W. HOLSAPPLE, Department of Natural Sciences and Mathematics, West Liberty University, West Liberty, WV, 26074. A machine learning approach for predicting molecular energy.

The PubChem library houses data for more than 100 million molecules. The features for each molecule in this dataset include the molecular energy, shape multipole, and the constituent atoms' 3-D coordinates. Simulations of molecular properties can be computationally expensive. In this project, we demonstrate how supervised learning may be used to predict a molecule's molecular energy using its molecular structure alone. Training a regressor to make such predictions requires feature engineering. We use a molecule's Coulomb matrix as our training feature. This matrix uses atomic numbers and distances between atoms to describe a molecule's structure. Initial model evaluation considered six regression models: gradient descent, linear regression with ridge regularization, decision trees, bagging regression (BG), adaptive boosting, and random forests (RF). Each was evaluated using root mean squared error (RMSE) and the coefficient of determination (R2). The BG and RF regressors were separated by less than 0.6% in both metrics. Both BG and RF performed approximately 23% better in RMSE and 12% better in R2 compared to the next best performing regressor. The top three models were then put through an extensive hyperparameter tuning pipeline. RF and BG remained the top performing regressors, with RF execution time approximately 13% faster than BG. Finally, a sensitivity analysis was performed to ensure that model performance was not dependent upon random partitioning of data training sets. Our results showed that the random forest regressor outperforms the other models considered for predicting molecular energy.