Predicting Molecular Properties: A Novel Data Driven Approach

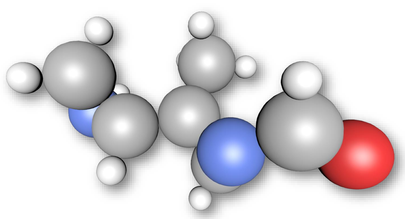
Kishore Rajesh[1], Amy Schneider[1], Riley Tallman[1], Nathanial White[1] and Yassin Youssfi[1]
Mentors: David Schneider[1], Abhishek Kothari[1], Chris Yoo PhD[1]
[1]Systems Imagination
[2]Chemistry and Mathematics in Phase Space (CHAMPS)

Abstract

The goal of the Kaggle competition "Predicting Molecular Properties" was to predict the scalar coupling constant (the magnetic interaction) between atom pairs in molecules, given the two atom types (e.g., C and H) and the molecular structure in three-dimensional space. Before the competition, the calculation of the constant, while possible, was time consuming and impractical. In order to improve the efficiency of calculations, we took a data driven approach with machine learning to estimate the scalar coupling. Using an ensemble of boosted decision trees, we accurately estimated the scalar coupling constant and placed in the 80th percentile on the competition leaderboard.

Introduction

Using state-of-the-art methods from quantum mechanics, it is possible to accurately calculate scalar coupling constants given only a 3D molecular structure as input. However, these quantum mechanics calculations are extremely expensive (days or weeks per molecule), and therefore have limited applicability in day-to-day workflows. A fast and reliable method to predict these interactions will allow medicinal chemists to gain structural insights faster and cheaper, enabling scientists to understand how the 3D chemical structure of a molecule affects its properties and behavior. Ultimately, such tools will enable researchers to make progress in a range of important problems, like designing molecules to carry out specific cellular tasks, or designing better drug molecules to fight disease. This project aims to develop an algorithm that can predict the magnetic interaction between two atoms in a molecule (i.e., the scalar coupling constant), and bypass the expensive quantum mechanics calculations altogether[1].

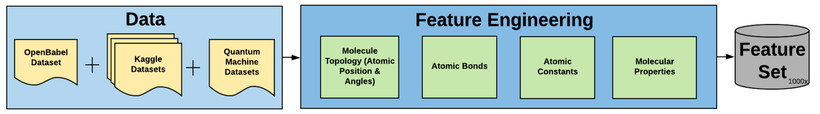


Materials

- NVIDIA DGX Workstation: 4X Tesla V100 GPUs, 128 GB GPU RAM, 2,560 Tensor Cores, 20,480 CUDA cores, Intel Xeon E5 2.2 GHz 20-core, 256 GB system RAM, running Ubuntu Linux OS 16.04.4 LTS
- Kaggle Molecular Properties Dataset: molecular structure (3D coordinates), atom types, and bond type for 85k molecules [1]
- Quantum-Machine Datasets: computed geometric, energetic, electronic, and thermodynamic properties for 134k stable small organic molecules made up of Carbon, Hydrogen, Oxygen, Nitrogen, and Flourine [3]
- Open-Babel: python library used to extract 27 features like orbital energies, molecular mass, and orthogonal angles between atoms for the Kaggle molecules [4]

Feature Engineering

Original Data: The original data from Kaggle consists of atom types (carbon, hydrogen, etc), relative Cartesian coordinates of atoms within each molecule (x, y, and z), and the number of intermediate bonds between two atoms (the bond type). The target is called the scalar coupling constant. This amounts to a total of five features and a single continuous (regression) prediction target [1].



Feature Engineering: After aggregating the provided Kaggle dataset with other open-source datasets from OpenBabel and Quantum Machine, we created new features with statistics based on categorical features. For instance, grouping by molecule and calculating the mean distance between atoms creates a new feature called `molecule_dist_mean`. This process is repeated for all categorical features and all numerical features, and for other statistics like standard deviation, minimum, and maximum. Next, additional features were created by adding, subtracting, and dividing numerical features by other numerical features. In turn, this approach generated over **1000 features** for our models.

Results

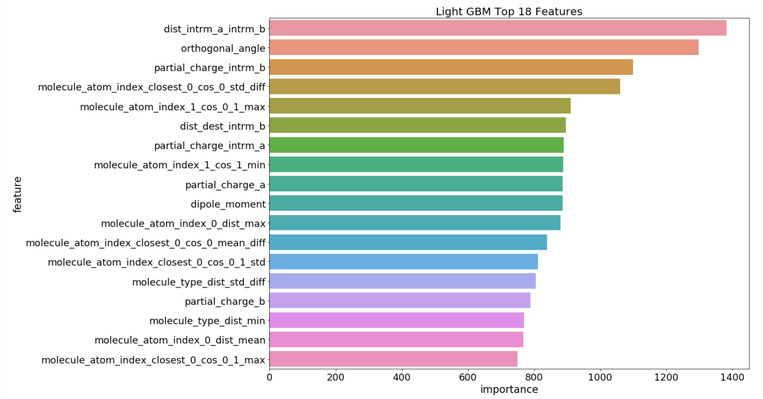
The team tested the feature set against multiple supervised learning algorithms including polynomial regression, gradient boosted random forests, and neural networks. We built numerous neural networks varying in depth, width, regularization, dropout, and optimization algorithms using Keras. Furthermore, the team took various unorthodox approaches to improve performance, like converting the regression problem into a classification problem (by rounding the scalar coupling constant to one decimal place and making each unique number a category) and creating eight different models based on a categorical feature with eight categories.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Scoring Metric: Mean Absolute Error (MAE)

Our highest performing algorithm used eight models, each employing an ensemble of gradient-boosted random forests. Using the LightGBM library developed by Microsoft to construct the models, we achieved a mean absolute error of 0.27 [5]. Overall the team placed in the **80th percentile** of the competition, out of over 2,000 teams.

Feature Analysis



Out of all 1000 features, we tested each one's individual importance to the scalar coupling constant using machine learning. As shown above, many of the best features were engineered features. Our experimental results support the hypothesis that the data-driven approach can significantly improve a predictive model. For a complex scientific competition such as this, the data-driven approach was very effective and revealed new insights surrounding the scalar coupling constant.

Future Work

The winner(s) of the competition will be submitting their source code as open source software under the MIT license (no limitations to use, copy, modify, distribute, and/or sell copies of the software). This way, researchers will have unrestricted access to a powerful tool that may assist in creating new medicines and developing drugs.

Acknoledgements

We would like to acknowledge David Schneider for his mentorship and guidance throughout the project, and to thank the entire Systems Imagination team for fostering a conducive learning environment and instilling an innovative spirit in every team member.

References

[1] "Predicting Molecular Properties." *Kaggle*, www.kaggle.com/c/champs-scalar-coupling/overview.
[2] "CHAMPS." *CHAMPS*, champsproject.com/
[3] L. C. Blum, J.-L. Reymond, 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13, J. Am. Chem. Soc., 131:8732, 2009.
[4] J. Cheminf. 2011, 3:33
[5] Ke, Guolin, et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." *Microsoft Research*, 9 Nov. 2017, www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/.