

Optimal scheduling of quantum chemical calculations under a time budget

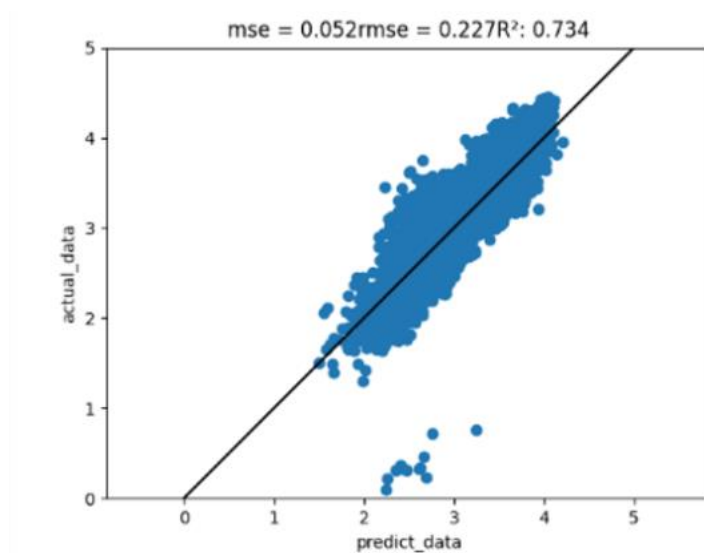
Laboratory of Large-scale Knowledge Discovery

Liu Yubin

Introduction

Quantum chemical calculations such as density functional theory (DFT) calculations are used to predict the behavior of electrons in molecular systems, thereby estimating molecular properties such as fluorescence [1]. However, DFT calculation is much more time consuming than classical calculation and its computational time is highly inconsistent even for the molecules of equal size. In the tasks of drug design and materials design using generative models [1,2], it is required to process as many molecules as possible by DFT calculations to increase the size of training data under a time budget.

In this work, we employ machine learning to predict the computational time of DFT calculation, and then use it for scheduling the given molecules for DFT calculation. How we should schedule the molecules depends on our task at hand. For example, one can consider several different evaluation criteria such as maximum molecular diversity and maximum value of target properties. Here we focus on scientific value function (SVF) [3] which is a recently proposed criterion to measure the scientific value of a molecule set based on the inter-sample distances in the molecular and property spaces.



Results

Figure 1: Prediction of DFT calculation time. X and Y axes correspond to predicted and true values, respectively.

Our dataset consists of 64,712 molecules from our previous study [2]. Each molecule has its DFT calculation time. It is divided into 80% training and 20% test sets. Using

random forest over 11 dimensional chemical descriptors including molecular weight, cLogP, TPSA, the number of acceptors and donors, we achieved relatively accurate prediction of DFT time ($R^2=0.734$, Figure 1).

Next, we scheduled all the molecules using two different algorithms to maximize SVF. One is to diversify the molecules by applying a clustering algorithm and pick up one from each cluster. The other is a genetic algorithm where the total DFT time is constrained within the predetermined threshold. The cost of each molecule is defined as the predicted DFT time. In comparison to random choice, both methods achieved better performance in achieving higher SVF within a limited time budget (Figure 2).

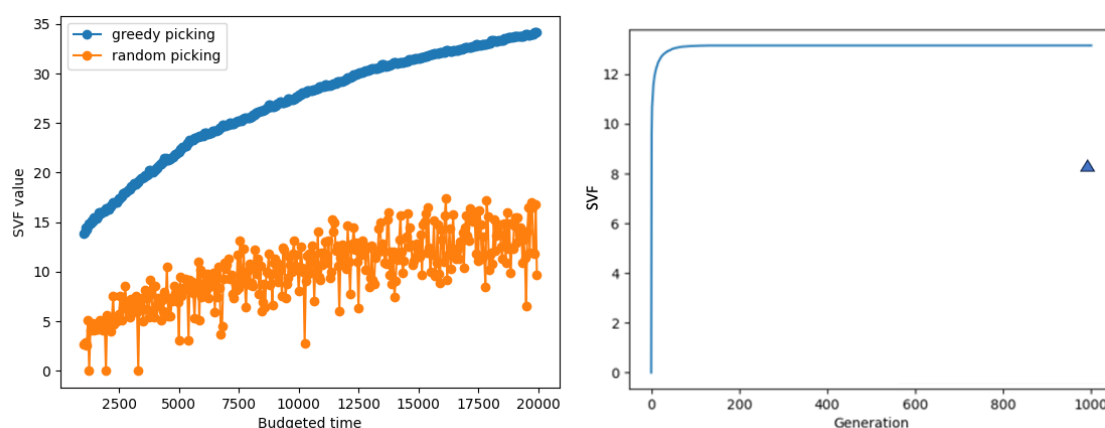


Figure 2: (Left) Selection of molecules by clustering. (Right) Results of the genetic algorithm. The time budget is set to 20,000. The triangle shows the result of random mutations.

Future Work

We only tried simple algorithms for scheduling, but we will apply more sophisticated scheduling algorithms to our problem. Also, in addition to SVF, other quality measures need to be applied. In addition, it is important to quantify how the time prediction error affects the scheduling results.

Reference

- [1] B. Huang et al., The central role of density functional theory in the AI age, *Science*, 381(6654), 170-175, 2023.
- [2] M. Sumita et al., De novo creation of a naked eye–detectable fluorescent molecule based on quantum chemical computation and machine learning, *Science Advances*, 8, eabj3906, 2022.
- [2] M.R. Carbone et al., Flexible formulation of value for experiment interpretation and design, *Matter*, 7, 1-12, 2023.