AP275 Project Update

Cooper Lorsung

April 17, 2020

1 Introduction

This project aims to implement a machine learning algorithm that uses limited Degrees of Freedom (DoF) to increase precision of the model. Using limited DoF means using a crude estimate of the property we are interested in. This can help give insights into the physics that matters for the quantity of interest. For example, the paper that developed this method uses Debye frequency as part of feature space, which is very directly relevant to calculating thermal conductivity[5]. The aim here is to both determine a good model and good features. I have reformulated my initial proposal to calculate formation energies of copper-tin alloys, with plans to expand this set. I could not find good reference data for vacancy formation energies of copper alloys, so I had to change direction. I chose this specific alloy due to the numerous applications of bronze in architecture and engineering. The compounds of interest are: Cu_3Sn, $Cu_{10}Sn_3$, $CuSn_3$, Cu_5Sn_4, $Cu_{81}Sn_{22}$, $Cu_{10}Sn_3$, Cu_6Sn_5, CuSn, $Cu_{41}Sn_{11}$, and $Cu_{19}Sn_{13}$. All lattice structures available on materialsproject[1] were used for these calculations. For example, Cu_3Sn has two structures available, one of Pmmm spacegroup and one of $P6_3/mmc$. Both structures were included in the dataset.

I find this topic especially interesting because it potentially allows for 'fast and dirty' calculations to be used in place of very expensive calculations. In this case, the expensive calculations used as reference come from MaterialsProject.org. The api is used to gather data on the compounds, and I am specifically comparing against formation energy per atom. I originally was going to only use stable compounds, but decided to use all Cu-Sn compounds after checking the phase-hull diagram[3]. I define a 'fast and dirty' calculation as one that does not make use of the most accurate and computationally expensive methods. A coarse k-point grid, inaccurate functional, or unexpressive pseudopotential may be used (in this case, all three). The model is then used to correct for these innaccuracies. The promise comes from points we have no reference for. Say, for example, we want to know if a newly posited compound is stable. This may be a very complicated compound that will take a lot of resources to accurately calculate. With a pre-trained model of similar compounds, we may be able to accurately estimate the formation energy of this new compound, with much cheaper calculations. From there, it can be determined if it is worthwhile to continue onto more expensive calculations. I am answering the question of

whether or not the DoF approach to modeling can achieve this goal in this context. Previous work has shown promise in calculating band gaps in binary semiconductors and thermal conductivity, and elastic modulus of zeolites[5]. Additionally, techniques have been used to calculate formation energy of interstitial atoms in HCP crystals[4].

2 Methods

Thus far, I have run quick, cheap calculations for Cu-Sn alloys. I have used the scalar relativistic LDA ultra-soft pseudopotential with non linear core correction for all calculations. Additionally, I have used a relatively coarse k-point grid. For primitive cells containing less than 10 atoms, I used 5x5x5 k-points, between 10 and 30 atoms I used 3x3x3, between 30 and 40 atoms I used 2x2x2 k-points, and above 40 atoms I used one k-point (the gamma point). I used quantum espresso for these calculations, and used ASE to convert cif files to quantum espresso input. No magnetization was needed for these calculations. Additionally, I will assume high-quality calculations are available already for the base elements. Copper and Tin calculations have been done and aren't going anywhere. As such, I used the PBESOL functional as well as converged the energy per atom with respect to k-points and supercell.

Once the calculation were run, I got the following results:

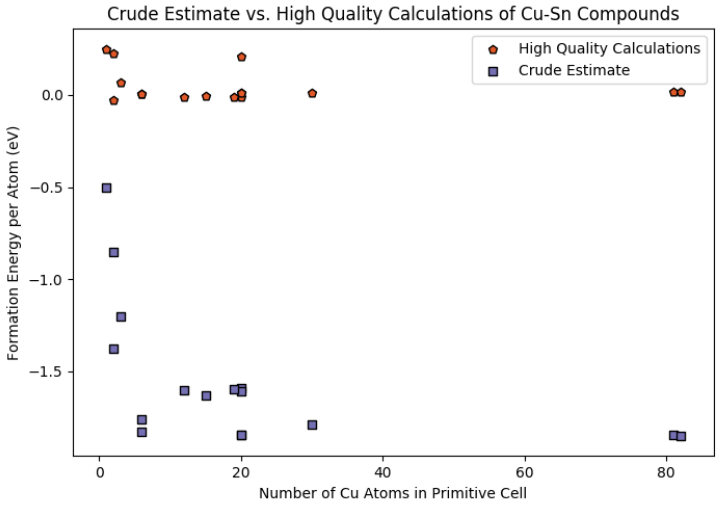

Figure 1: Comparison between my crude calculations and high quality calculations from MaterialsProject.org

we clearly see systematic underestimation of formation energy for every compound. While a consistent pattern such as this is promising and relatively easy

to correct for, the aim is to be more accurate than an average correction.

To accomplish this end, I applied machine learning techniques[2]. The models used were Lasso Regression and Kernel Ridge Regression with an RBF kernel as outlined by Zhang and Ling. Feature space was constructed from readily available quantities from the materialsproject API. Assuming now knowledge of which quantities are useful for calculating energy, I iteratively checked every combination of features. Those features are: unit cell density, unit cell volume, total magnetization, number of copper atoms, number of tin atoms, lattice parameters and lattice angles. Total magnetization was included in the potential features as a test to make sure the model would not prefer useless features. A comparison was done between models that included and excluded the crude estimate.

3 Results and Challenges

So far results have not been somewhat disappointing. We see below:

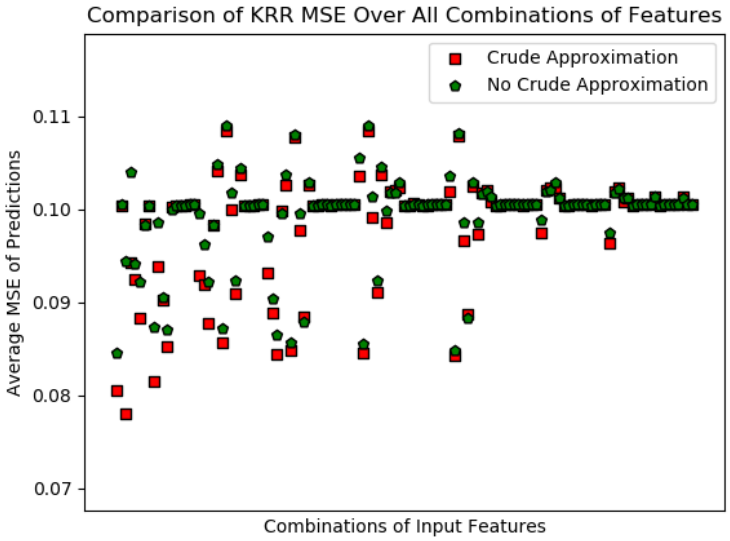

Figure 2: Comparing average mean squared error (MSE) for each data point using leave-one-out validation. The x-axis represents a different combination of input features. Labeling each point of the x-axis would be illegible, so this plot is used as a tool to direct further exploration, rather than come to hard conclusions.

On average, learning with the crude approximation does better, although only slightly, than learning without the crude approximation. The disappoint-

ment comes from the fact that the model is not robust to useless input. The best combinations of features from this experiment is crude approximation and total magnetization. Total magnetization was not used in the DFT calculations because it is so small, a maximum value of 7.197×10^{-5} Bohr magnetons, rounded down to 0 on the web page.

The best model made the following predictions

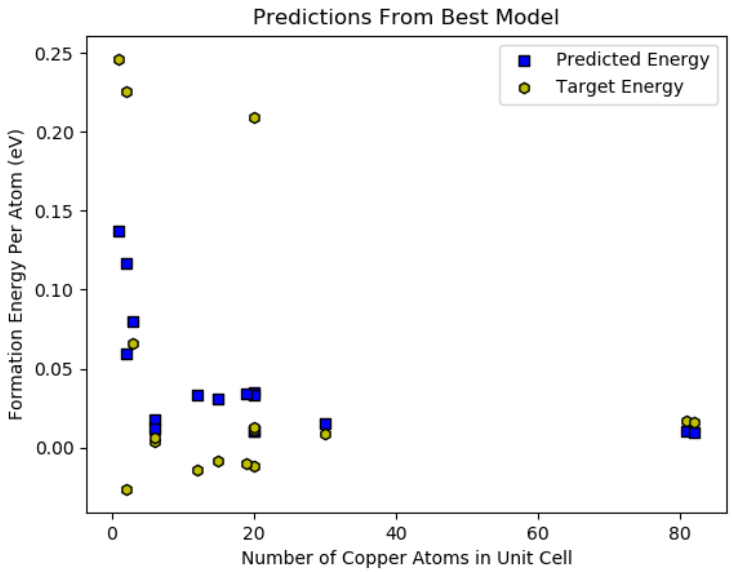

Figure 3: Predictions for each value from the best model. This was KRR with total magnetization and crude approximation as input features.

We see that the model is close for some points, but very wrong for others, notably the predictions are quite accurate for the largest compounds, and are very wrong for the smallest.

It does not appear that there is a linear relationship between any of the parameters studied so far and the formation energy per atom. The feature space likely needs to be expanded. Elasticity data is available for each compound, which has thus far not been included due to additional postprocessing needed.

Some of the next steps are to try new models and and expand current models. I have had trouble implementing hyperparameter tuning using cross-validation, which was used in the original paper. This is a very obvious next step that needs to be included for fair comparison. Different kernels in KRR could provide better results, such as cosine similarity or linear kernels. I am unsure on how applicable these kernels are to the task at hand, but the option will be explored. Additionally, qualitative checks will be done to see if the model's corrections are reasonable. I will recreate the hull diagram to check if it predicts the correct

compound as stable, and approximately gets the correct energy above the hull. Time permitting, I will expand this to another set of materials to see if the idea is generalizable. I was originally looking at steel alloys. These present the additional challenge of magnetization. However, expanding to this set of materials is looking unlikely.

References

[1] Anubhav Jain et al. "The Materials Project: A materials genome approach to accelerating materials innovation". In: *APL Materials* 1.1 (2013), p. 011002. ISSN: 2166532X. DOI: 10.1063/1.4812323. URL: http://link.aip.org/link/AMPADS/v1/i1/p011002/s1%5C&Agg=doi.

[2] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 0262018020.

[3] Shyue Ping Ong et al. "Li-Fe-P-O2 Phase Diagram from First Principles Calculations". In: *Chemistry of Materials* 20.5 (Mar. 2008), pp. 1798–1807. ISSN: 0021-9606. DOI: 10.1021/cm702327g.

[4] Gaegun You et al. "Machine learning-based prediction models for formation energies of interstitial atoms in HCP crystals". In: *Scripta Materialia* 183 (July 2020), pp. 1–5.

[5] Y. Zhang and C. Ling. "A strategy to apply machine learning to small datasets in materials science". In: *npj Compututational Materials* 4 (1 May 2018).