Machine Learning for Correcting Inaccurate DFT Calculations

By: Cooper Lorsung

AP 275 Final Project Spring 2020

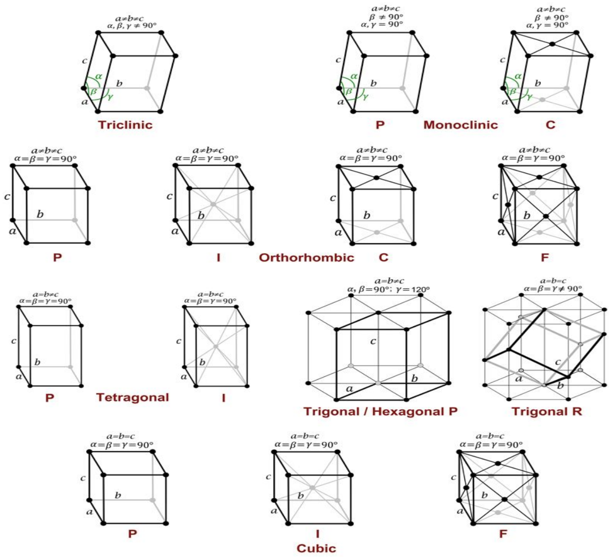
Introduction

- Formation energy of Cu-Sn alloys
  - CuSn, CuSn3, Cu3Sn, Cu5Sn4, Cu81Sn22, Cu10Sn3, Cu6Sn5, Cu41Sn11, Cu19Sn13
  - LDA functional, ultra soft pseudopotential
- Cheap and inaccurate calculations
- Very small data sets (15 points)
- Use machine learning to correct for errors (Y. Zhang and Z. Lin)
- Target energies come from MaterialsProject.org
  - GGA functional, PAW potential

Source: http://conexsprings.com/subproduct/bronze-helical-springs.html

Crude Estimate of Property

$$E_{\text{Formation}} = \frac{E - N_{Cu}E_{Cu} - N_{Sn}E_{Sn}}{N_{Cu} + N_{Sn}}$$



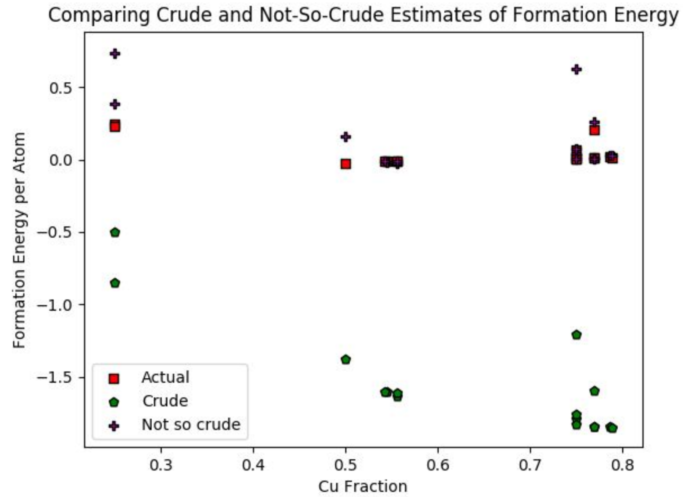
Source: https://www.xtal.iqfr.csic.es/Cristalografia/parte_03_4-en.html

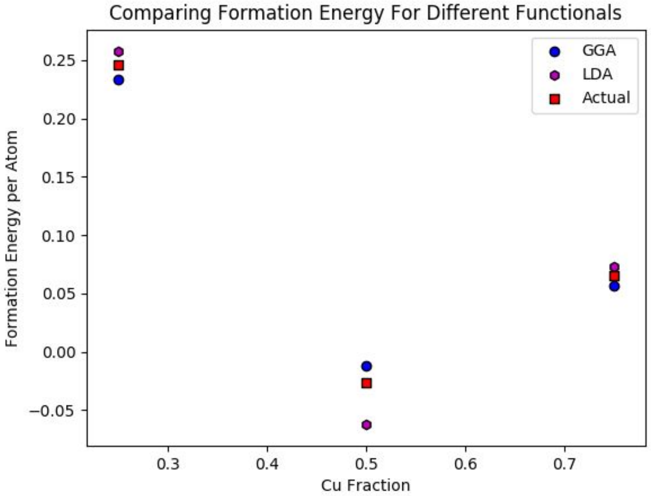
- Crude Estimate of property
  - Use crude estimate of target property (i.e. fast, inaccurate calculations)
  - Reduces degrees of freedom
- Feature Selection
  - Carefully select features to give insights into relevant physics
  - Use readily available features such as density, lattice parameters, etc.
  - I found no clear trends in with any single feature and formation energy

Sources of Error

- K-point error
  - Primary source of error
  - Mostly from non-converged reference material energies
  - Can compare crude estimate and ‘not-so-crude’ estimates

- Functional error
  - With k-point error converged, GGA performed slightly better than LDA
  - I was unable to perfectly replicate MaterialsProject data





Models

Kernel Ridge Regression

$$\hat{f}_{\mathrm{KRR}}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$$

$$C(\alpha_1, \ldots, \alpha_n) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \hat{f}_{\mathrm{KRR}}(\mathbf{x}_i)\right)^2 + \eta \sum_{ij} \alpha_i K(\mathbf{x}, \mathbf{x}_i)\alpha_j$$

Source: Y. Zhang and Z. Lin

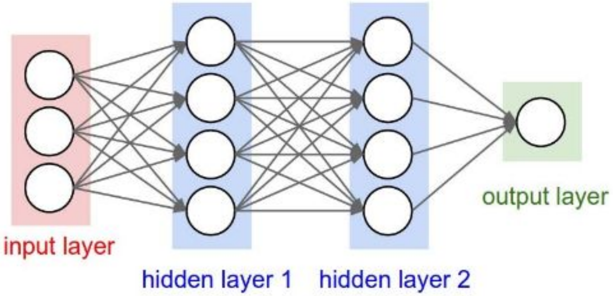
$$K_{RBF}(\mathbf{x}, \mathbf{x}') = \exp\left(\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

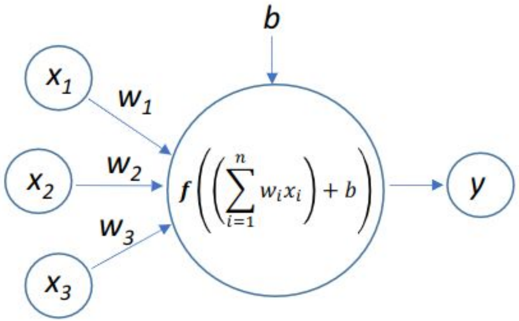
$$K_{CS}(\mathbf{x}, \mathbf{x}') = \frac{\mathbf{x}\mathbf{x}'^T}{\|\mathbf{x}\| \, \|\mathbf{x}'\|}$$

Source: https://en.wikipedia.org/wiki/Radial_basis_function_kernel

Source: https://scikit-learn.org/stable/modules/metrics.html

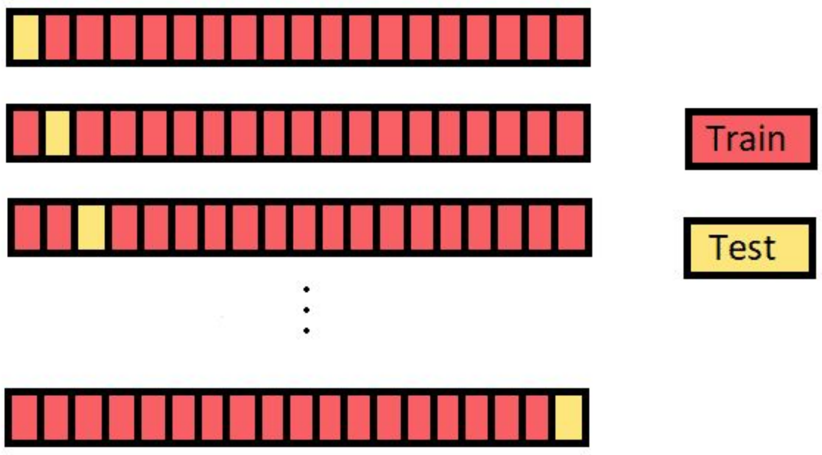
Multilayer Perceptron





Source: AP275 SP20 Lecture 16 Slide 12

Model Selection

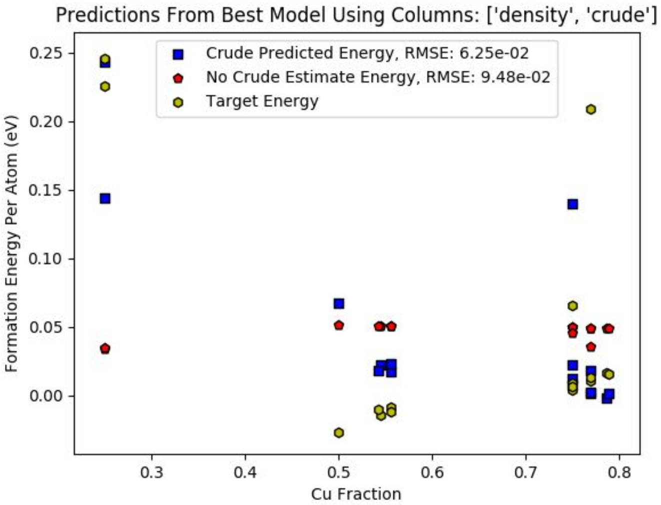


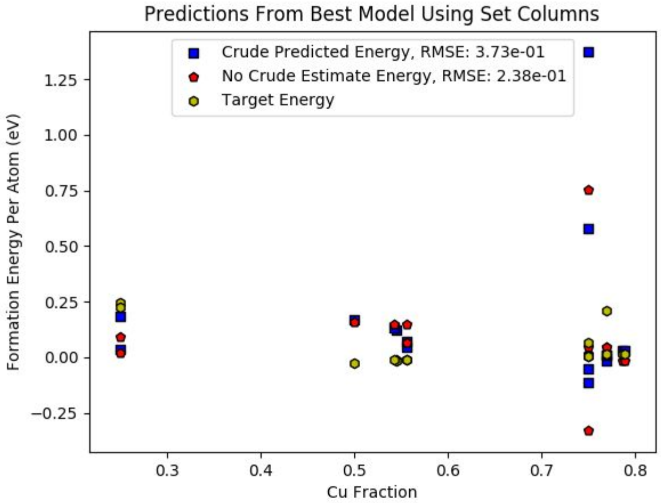
Source: https://aiaspirant.com/cross-validation/

- Leave one out Cross Validation
  - Excludes one datapoint, trains on the rest, predicts that datapoint
  - Average error after all data points have been used as test point
- All possible combinations of features were used
- Selected model with lowest average RMSE

Results

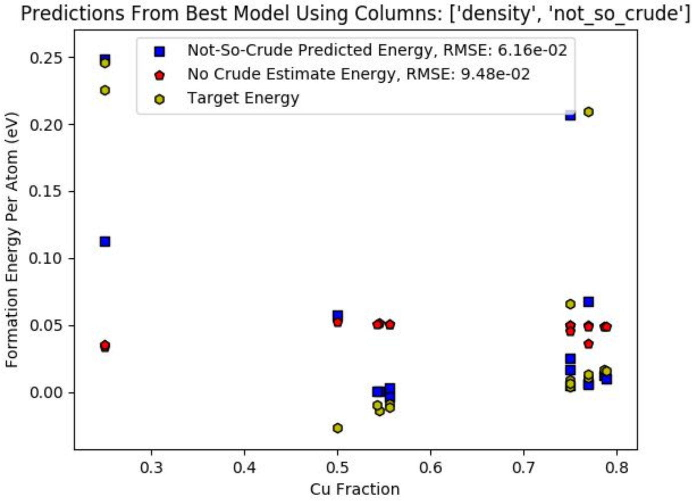
- Using the crude estimate gives much better results than not using it.
- Kernel Ridge Regression with cosine similarity is the best model in both cases
- Using additional physical features seems to qualitatively improve the no crude estimate predictions



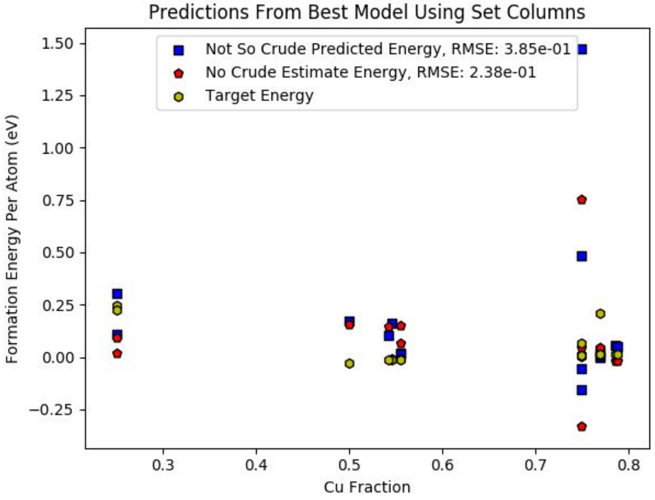


Results Cont.

- Best columns agree with cruder estimate
- Best model with set columns is Kernel Ridge Regression
- Best model with adaptive columns is the Neural Network
- With set columns an outlier seems to be causing issues



Predictions From Best Model Using Columns: ['density', 'not_so_crude']

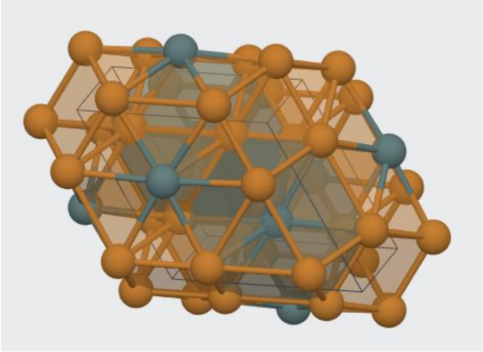


Predictions From Best Model Using Set Columns

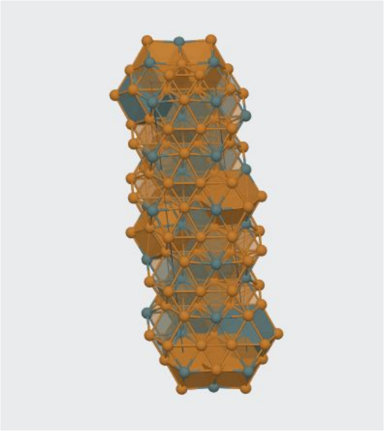
Conclusion

- Using crude estimates improves predictive quality
- Using more accurate crude estimates increases predictive quality further
- CuSn3 with spacegroup 63 was an outlier for most calculations.
  - Could be due to unusual shape
  - Is very oblong compared to other materials
- This idea shows promise but is not ready to make high-stakes predictions

P63/mmc Cu3Sn

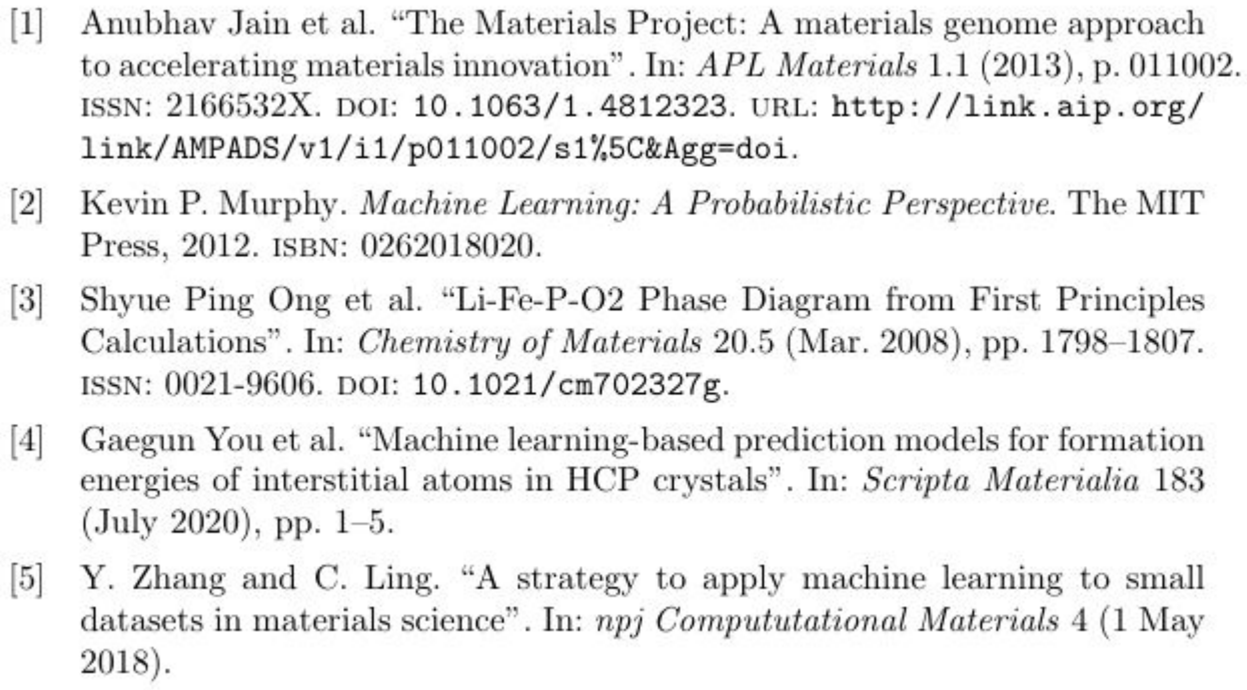


Cmcm Cu3Sn



Source: MaterialsProject.org

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).