Rapid Prediction of Formation Energy Using Machine Learning

Phase equilibria play a central role in the development of new materials with advanced properties. However, assessing a single binary or ternary phase diagram is a laborious process that often takes years. Rapid prediction of phase equilibria remains a grand challenge for accelerating the materials development cycle. To estimate phase equilibria at 0 K, the 0 K formation energy of each phase must be known. This work employs machine learning to develop a model that predicts the 0 K formation energy of a phase using only its chemical formula as input. The model could be used to elucidate 0 K phase equilibria several orders of magnitude more rapidly than conventional methods.

The data used to train the model was scraped from the Materials Project, a database containing the calculated formation energies of 69,640 inorganic compounds. The formation energy and chemical formula of each compound was gathered and stored in two matrices. A matrix **X** contains the features of the compounds: the mole fractions of a particular element, $x_{element}$, in the compound (e.g. $x_{Ni} = 0.5$ and $x_{Al} = 0.5$ for NiAl), and a matrix **y** contains the properties: the calculated formation energy of the compounds.

The heat map in Figure 1 visualizes the chemical makeup of compounds in the dataset. Each entry represents the number of times a pair of elements is encountered together in the compounds. For example, the compound Li_2O would contribute to entries O-O and Li-Li along the trace and Li-O and O-Li on the off-diagonal. Importantly, the values have been normalized against the total number of compounds in the dataset. Darker entries signify that a particular combination of elements is encountered more frequently in the compounds. Based on the figure, it is apparent that the vast majority of the compounds are formed by elements with lower atomic number and particularly by the elements Li, O, F, P, S, Mn, and Fe. Furthermore, there are many compounds with Li and O, P and O, and Li and P (reflecting the large number of battery materials in the Materials Project).

The data was divided into a training set (60%), a cross-validation set (20%), and a test set (20%) and a least-squares linear regression model was fit. Regularization was not employed because the number of features is much smaller than the number of examples. Figure 2 plots the formation energy predicted by the model against the calculated formation energy. The RMS error on the CV set was calculated to be ~500 meV/atom. To determine if a non-linear hypothesis could improve the error, a neural network was also fit to the data. The neural network had a single hidden layer with five neurons. Figure 3 plots the formation energy predicted by the neural network model against the calculated formation energy. The RMS error on the CV set was calculated to be ~240 meV/atom. Thus the neural network outperforms the linear regression model and reduces the error by a factor of two. Several other neural networks with different numbers of hidden layers and neurons per layer were tested, but exhibited higher error.

The RMS error of the neural network on the test set is quite large, ~225 meV/atom. As comparison, formation energies used in computational assessment of phase diagrams are usually accurate to about 10-20 meV/atom. Thus our algorithm does not reach the desired performance by a factor of about 10-20. The learning curve in Figure 4 suggests that our

model is underfit and that implementing more features could improve the quality of the model. One feature that was not implemented but that could substantially improve the error is the crystal structure of the compound. The crystal structure should improve the error for several reasons. First, the current algorithm cannot distinguish different polymorphs of the same compound. For example, there are 17 instances of ZrO_2 in the dataset that correspond to 9 different polymorphs. The formation energies for these polymorphs span a range of 100 meV/atom. Secondly, the crystal structure contains information about the character of the bonds in the compound, e.g. difference in electronegativity, which influence the strength of the bonds and hence the formation energy.

In conclusion, an algorithm that predicts formation energies of phases using only their chemical formula was developed. However, the RMS error of the model, ~200-250 meV/atom, is too large for phase diagram prediction. Adding additional features, particularly information relating to crystal structure, is expected to improve the model, but it is unclear whether or not the desired performance can be achieved.

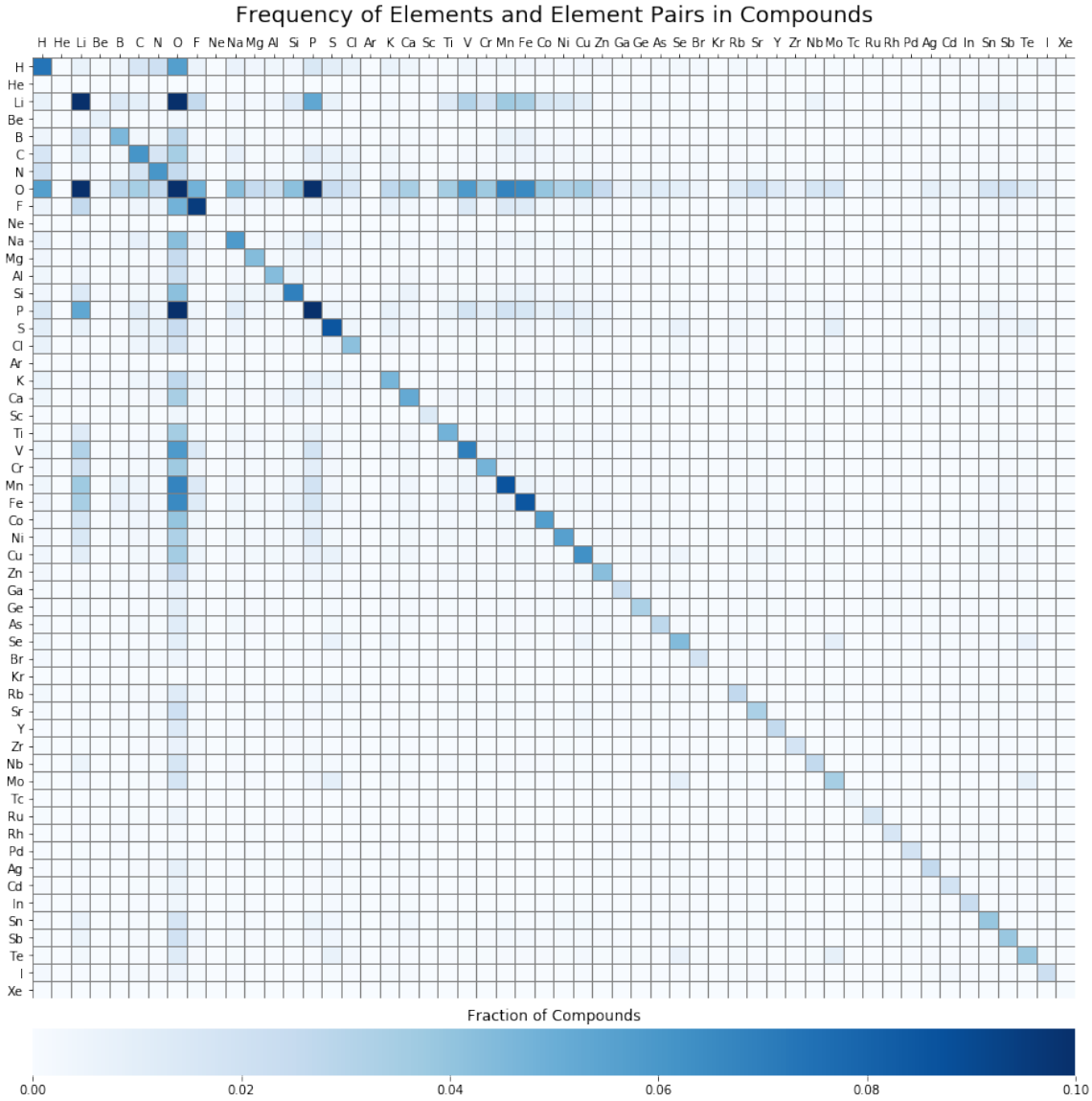

Figure 1. Heat map showing the frequency of elements and element pairs in compounds in the dataset. Only the first five rows of the periodic table are shown due to the lower frequency of higher atomic number elements. The darkest blue color signifies that 10% or more of the compounds include a particular combination of elements.

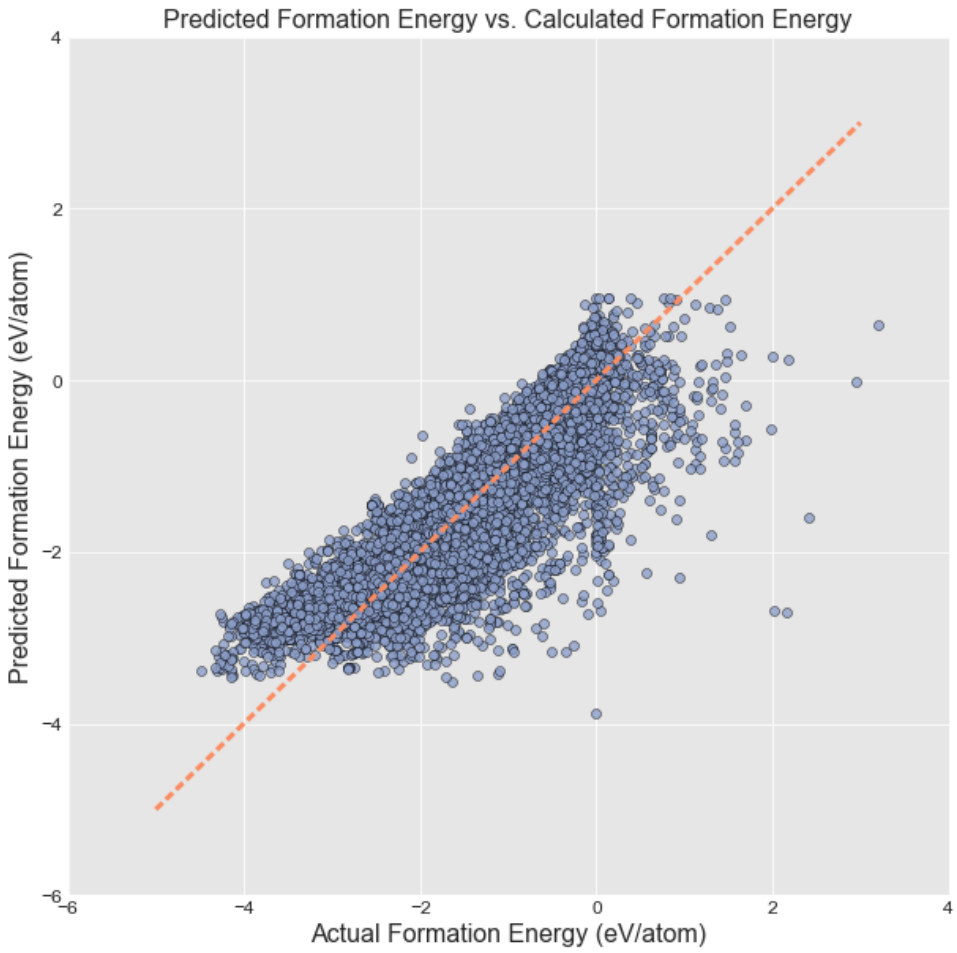

Figure 2. Plot showing the predicted formation energy from the least-squares linear regression model against the calculated formation energy. The dotted orange line shows a perfect model.

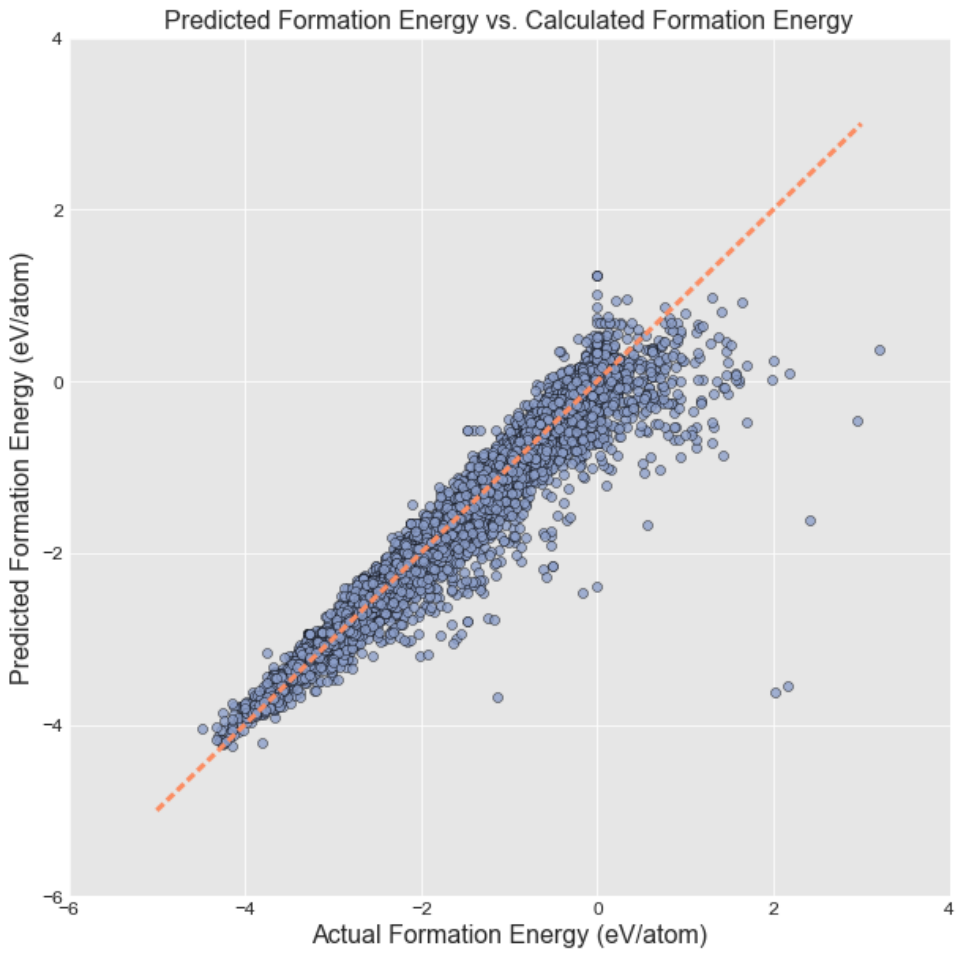

Figure 3. Plot showing the predicted formation energy from the neural network against the calculated formation energy. The neural network had a single hidden layer with five neurons and outperforms the linear regression model. The dotted orange line shows a perfect model.

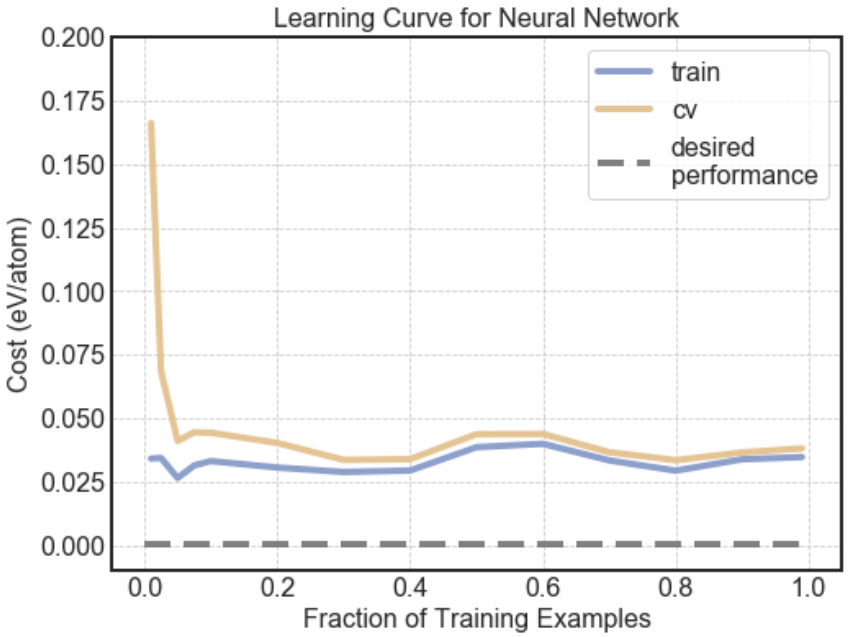

Figure 4: Learning curve for the neural network. Note that the desired performance is not set at 0 but at (Desired RMS)^2 / 2 = 0.0002 (desired RMS of 20 meV/atom).