Machine Learning Methods for Interatomic Potentials: Application to Boron Carbide

Qin Gao

Department of Physics and Machine Learning Department, Carnegie Mellon University

qingao@andrew.cmu.edu

Committee: Jeff Schneider, Michael Widom, and Geoff Gordon

May 11, 2015

Abstract

Total energies of crystal structures can be calculated to high precision using quantum-based density functional theory (DFT) methods, but the calculations can be time consuming and scale badly with system size. Boron carbide exhibits disorder in the distribution of boron and carbon atoms among the crystallographic sites. A cluster expansion of the DFT energy in a series of pairs, triplets, etc. is prohibitive owing to the structural complexity. We fit the energies using machine learning methods like neural network, Gaussian process and support vector regression based on pair correlations only in order to capture nonlinear effects associated with many-body interactions. We use our interaction model in Monte Carlo simulations to evaluate the phase diagram.

1 Introduction

Boron carbide is an extremely hard and very light material with wide range of applications in industry and in the military. It is also used as efficient neutron absorbent in nuclear power plants. Despite its importance, the phase diagram of boron carbide is not precisely known [1, 2] due to its structural complexity, difficulty of equilibration and the small difference between the atomic numbers of boron, which is 5, and carbon, which is 6. The phase diagram describes the stable structures of boron carbide at certain conditions like temperature and composition which is crucially important both for industry and in fundamental physics.

Two major problems exist in the widely accepted experimental boron carbide phase diagrams [3, 4]. The solubility range of carbon is $9\% - 19.2\%$ and does not shrink as temperature decreases. Extrapolation of this range to low temperature will violate the third law of thermodynamics at $T = 0\ K$. The second issue is that the upper limit of carbon solubility range is 19.2% which is less than 20%, the composition of a theoretically

predicted low temperature stable structure. Since experimental measurement is not reliable at low temperature (< 1000 K), theoretical research is warranted to resolve these remaining problems.

The primitive cell of boron carbide is shown in Fig. 1. The primitive cell (smallest repeating unit) of $B_{13}C_2$ has 15 atoms, a C-B-C chain in the center and a 12-atom boron icosahedron at every vertices. Boron atoms on top and bottom of the icosahedron can be substituted by carbon atoms with very low energy cost. This degree of freedom leads to infinitely many possible structures in the thermodynamic limit. To study thermodynamics, especially phase transitions of boron carbide, we need to well sample the configuration space of large supercells (e.g. $8\times8\times8$ cell = 7680 atoms) and accurately determine the energies of the structures. The task of this project is to accurately determine the energies of boron carbide structures by combining quantum mechanical calculations and machine learning models.

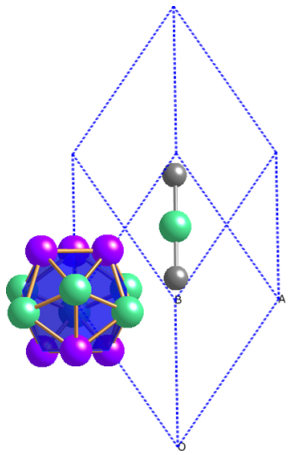

Figure 1: The primitive cell of boron carbide ($B_{13}C_2$). Dashed blue lines depict the cell. Two grey atoms are carbon atoms. Both green and purple atoms are boron atoms. The top and bottom of the icosahedron (purple) can be substituted by carbon atoms at low energy cost.

2 Background

2.1 Density functional theory

In quantum mechanics, the energy of a system is determined by solving its Schrodinger Equation:

$$H(\vec{R}_1, \vec{R}_2, ..., \vec{r}_1, \vec{r}_2, ...)\Phi_n(\vec{R}_1, \vec{R}_2, ..., \vec{r}_1, \vec{r}_2, ...) = E_n\Phi_n(\vec{R}_1, \vec{R}_2, ..., \vec{r}_1, \vec{r}_2, ...), \quad (1)$$

where H is the Halmitonian of the system, Φ_n is the wavefunction of the nth excited stated (n=0 for ground state), E_n is the energy of nth excited state, $\{\vec{R}_i\}$ are the position of

ions and $\{\vec{r}_i\}$ are the position of electrons in the system. Eq. 1 is a linear second order differential equation. However it may have hundreds to thousands variables and thus is usually intractable to solve directly.

Hohenberg and Kohn [5] theoretically proposed and proved that the ground state energy of the system can be uniquely determined by the electron density $n(\vec{r})$, namely how many electrons exist at position $\vec{r}$. Kohn and Sham [6] showed how to replace Eq. 1 with a set of coupled nonlinear equations with a single 3-D variable $\vec{r}$. These theories, which are referred to as density functional theory (DFT), won Kohn the Nobel Prize in Chemistry in 1998. In DFT, instead of directly calculating $\Phi_n(\vec{R}_1, \vec{R}_2, ..., \vec{r}_1, \vec{r}_2, ...)$ which has hundreds to thousands variables, the electron density $n(\vec{r})$ is calculated, which has only one 3-D variables. DFT thus dramatically decreases the computational complexity. Standard DFT software packages, like the Vienna Ab initio Simulation Package (VASP) [7, 8], are well developed and widely used in physics, chemistry and material science communities. The DFT calculated energies and various other physical properties of crystals are very accurate and can be directly compared with experiments.

2.2 Interatomic Potentials

Despite the great success of DFT, it is still expensive to calculate energies of structures with primitive cells of hundreds of atoms or more. For example, it takes around a week for a 4-core machine to calculate the energy of one boron carbide 3×3×3 structure (405 atoms). Since the time complexity scales as $O(N^3)$ in DFT, where N is the number of atoms, it is almost impossible to calculate the energy of even larger cells.

On the other hand, to study the phase transitions of boron carbide, the energies of millions large-cell structures (at least 8×8×8 = 7680 atoms) have to be accurately determined. We thus need to construct an accurate energy model based on DFT calculated energy dataset of relatively small cells (2×2×2 and 3×3×3). We call this energy model an interatomic potential, which ultimately only depends on the positions of the atoms in the cell. The underlying assumption is that the interaction between atoms are the same in cells with different sizes which is generally true for most crystals.

2.3 Physical Model

One way of modeling the energies of structures is using cluster expansion [9, 10]. In principle, if we know the numbers of pairs (2-atom clusters) of all types, numbers of triplets (3-atom clusters) of all types, numbers of quadruplets (4-atom clusters) of all types, and so on, we can uniquely identify the structure. Ignoring noise, if we assign (through a fit) an energy for every such cluster, we can in principle perfectly recover the energies of the structures by the cluster expansion,

$$E(\vec{N}^{\text{pairs}}, \vec{N}^{\text{triplets}}, ...) = \sum E_i^{\text{pair}} N_i^{\text{pair}} + \sum E_j^{\text{triplet}} N_j^{\text{triplet}} + ... \quad (2)$$

where $\vec{N}^{\text{pairs}}$ denotes the collection of numbers of pairs at certain distances, $\vec{N}^{\text{triplet}}$ denotes the collection of numbers of triplets of certain types.

However, due to the complexity of the boron carbide structure, there are too many triplets or higher order clusters to be included. One feasible approximation to the exact cluster expansion is to truncate the expansion at the pairwise level. The numbers of carbon-boron pairs and boron-boron pairs are determined if we know the numbers of carbon-carbon pairs. Moreover, since boron carbide is a hard material, the atoms always stay close to the fixed lattice sites. Thus the relaxed energy is a deterministic functions of the initial positions of carbon atoms. Our linear model based on this physical approximation, which serves as a baseline model, is

$$E(\vec{N}) = E(N_0, ..., N_{23}) = E_0 + \sum_{i=0}^{23} \beta_i N_i, \tag{3}$$

where N_i's are the features we use. Most of the features are number of carbon-carbon pairs in the structure. A detailed description of the features is in Section 4. We use 24 different pairs, which include carbon-carbon bonds up to 6.5 Å.

We notice that information in the higher order terms like triplets might be expressed as nonlinear function of pairs. We thus should try nonlinear models that capture the local nonlinear properties to improve the fit. This motivates our study with machine learning methods.

3 Machine Learning Method

We fit the DFT calculated energies with four machine learning models. We first exploit two parametric models the L_1-penalized polynomial regression (PR) and neural network (NN) and then two nonparametric models, Gaussian process (GP) and support vector regression (SVR). We discuss the essence of them in this section. Since we eventually decide to use GP and SVR, we describe them in more detail here.

3.1 Polynomial Regression

One direct nonlinear generalization of the linear model in Eq. 3 is the polynomial models. Due to the limited size of data (~ 600), we choose second order polynomial model,

$$E(\vec{N}) = E(N_0, ..., N_{23}) = E_0 + \sum_{i=0}^{23} \beta_i N_i + \sum_{j=0}^{23} \sum_{k=j}^{23} \gamma_{jk} N_j N_k, \tag{4}$$

which fully characterize the second order interactions between number of pairs with 325 parameters. To avoid overfitting and perform feature selection we add a L_1 penalty term. The resulting optimization problem is,

$$\min_{\vec{\theta}} \sum_{m=1}^{M} \frac{1}{2} (E_m^{DFT} - E(\vec{N}_m; \vec{\theta}))^2 + \lambda \|\vec{\theta}\|_1, \tag{5}$$

where $\vec{\theta}$ is the collection of all parameters, E_0, $\{\beta_i\}$ and $\{\gamma_{ik}\}$, λ is a tuning parameter, E_m^{DFT} is the DFT calculated energy of the mth structure, $\vec{N}_m$ is the 24 dimensional feature vector of mth structure and M is the size of the training set. To clarify, in this paper, we use i, j and k as the index for the features and l, m and n for the index of samples.

3.2 Neural Network

In our neural networks, the input layer contains 24 nodes corresponding to the components of the 24 dimensional input vector. We choose one to two hidden layers and one to ten nodes in every layer. We use nonlinear activation functions like "tanh" and "sinh" in the hidden layers and linear function in the single node output layer. However, since we only have around 600 data, complex neural network can easily overfit. We use the Bayesian regularization for the parameters in the model to reduce/avoid overfitting. A detailed description of neural network can be found in [11].

3.3 Gaussian Process

In GP, we assume the energies of structures are gaussian distributed,

$$\begin{pmatrix} E_{\text{train}} \\ E_{\text{pred}} \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma), \text{ and } \Sigma = \begin{pmatrix} \Sigma_{tt} & \Sigma_{tp} \\ \Sigma_{tp}^T & \Sigma_{pp} \end{pmatrix},$$

where the E_{train} and E_{pred} vectors denote the energies of training structures and structures whose energies to be predicted (predicting structures) respectively, μ is the mean of the energy distributions (set to zero in later derivation for simplexity), Σ is the covariance matrix. Σ_{tt} is the covariance matrix of training structures, Σ_{tp} is the covariance matrix between training structures and predicting structures, and Σ_{pp} is the covariance matrix of predicting structures.

The mth row and nth column of the covariance matrix Σ is,

$$\Sigma_{mn} = k(\vec{N}_m, \vec{N}_n), \tag{6}$$

where $\vec{N}_m$ and $\vec{N}_n$ are the feature vectors of the mth and nth structure respectively, and $k(\vec{N}_m, \vec{N}_n)$ is called the kernel function.

The kernel function characterizes the similarity between two feature vectors, and therefore structures. In our study, we use and compare the polynomial kernel $(1 + \beta \vec{N}_m \cdot \vec{N}_n)^d$, the Gaussian kernel $\exp(-\|\vec{N}_m - \vec{N}_n\|^2/\gamma^2)$, and the Laplacian Kernel $\exp(-\|\vec{N}_m - \vec{N}_n\|_1/\gamma)$. A constant variance term δ^2 is added to the kernel when $m = n$ to model the noise. The parameters β, d, γ and δ in the kernels are called hyperparameters. We can optimize these hyperparameters by maximizing the likelihood of the training data, which is a convex optimization problem that can be efficiently solved [12].

The covariant matrix Σ can be calculated using the feature vectors only. Under the assumptions of GP, the conditional distribution of energies of predicting structures are,

$$E_{\text{pred}}|E_{\text{train}} \sim \mathcal{N}(\Sigma_{tp}^T \Sigma_{tt}^{-1} E_{\text{train}}, \Sigma_{pp} - \Sigma_{tp}^T \Sigma_{tt}^{-1} \Sigma_{tp}). \tag{7}$$

We take the mean values of the distribution as the predicted energies. More explicitly, for a new structure with feature vector $\vec{N}_l$, the predicted energy is,

$$E(\vec{N}_l) = \sum_{m=1}^{M}\sum_{n=1}^{M} k(\vec{N}_l, \vec{N}_m)(\Sigma_{tt}^{-1})_{mn} E_{\text{train,n}}. \tag{8}$$

Moreover, GP also provides the variance of the predicted energy which implies the accuracy or confidence of the prediction. For a new structure with feature vector $\vec{N}_l$, the variance of the predicted energy is,

$$\sigma^2(\vec{N}_l) = k(\vec{N}_l, \vec{N}_l) - \sum_{m=1}^{M}\sum_{n=1}^{M} k(\vec{N}_l, \vec{N}_m)(\Sigma_{tt}^{-1})_{mn} k(\vec{N}_n, \vec{N}_l). \tag{9}$$

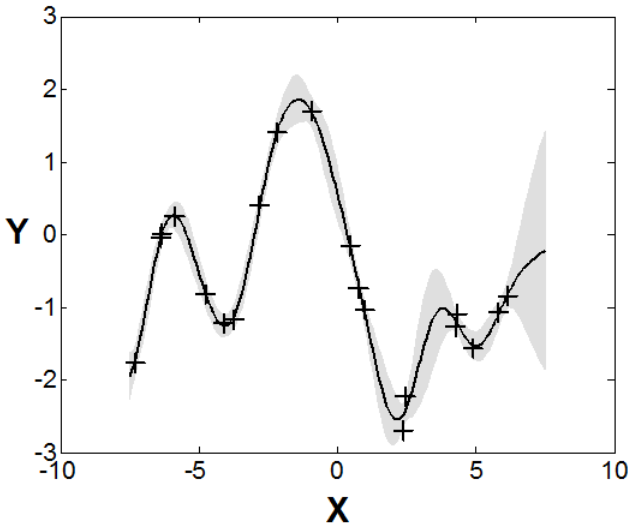

Figure 2: A GP fit of a toy one-dimensional function. A "+" denotes a data point. The solid curve is the GP fitted function. The shaded region is the region within one standard error of the GP prediction.

To illustrate the properties of GP, Fig. 2 shows a toy example of fitting a one dimensional function. The fitted nonlinear function well captures the local properties of the data. Moreover, the fit provides the standard error of the prediction, where standard error indicates small data density in nearby region. The standard error of prediction thus can be used to check whether the data well sample the feature space.

3.4 Support Vector Regression

In SVR [11] the fitted function $E(\vec{N}) = \vec{\omega} \cdot \vec{\Phi}(\vec{N}) + b$ minimizes the target function,

$$C\sum_{m=1}^{M} \max(0, |E(\vec{N}_m) - E_m^{\text{DFT}}| - \epsilon) + \frac{1}{2}\|\vec{\omega}\|_2^2, \tag{10}$$

where C and ϵ are positive real numbers, and $\vec{\Phi}(\vec{N})$ is a collection of chosen functions of $\vec{N}$. We can introduce the slack variables, construct the Lagrangian and obtain the dual form of this optimization problem as,

$$\max_{\vec{a},\vec{e}} -\frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{M}(a_m - e_m)(a_n - e_n)k(\vec{N}_m, \vec{N}_n) - \epsilon\sum_{m=1}^{M}(a_m + e_m) + \sum_{m=1}^{M}(a_m - e_m)E_m^{\mathrm{DFT}}$$
$$\text{subject to } 0 \le a_m \le C \text{ and } 0 \le e_m \le C,\ \forall m = 1, 2...M, \tag{11}$$

where $k(\vec{N}_m, \vec{N}_n) = \vec{\Phi}(\vec{N}_m)^T \cdot \vec{\Phi}(\vec{N}_n)$ is called the kernel function, similar as in GP. The kernel function rather than $\vec{\Phi}(\vec{N})$ is used to perform feature transformation and make prediction with,

$$E(\vec{N}) = \sum_{m=1}^{M}(a_m - e_m)k(\vec{N}, \vec{N}_m) + b, \tag{12}$$

where only data points on or outside the ϵ tube has nonzero a and e values, and are called support vectors. b can be calculated by support vectors,

$$b = E(\vec{N}_l) - \epsilon - \sum_{m=1}^{M}(a_m - e_m)k(\vec{N}_l, \vec{N}_m), \tag{13}$$

where $\vec{N}_l$ is any one of the support vectors. The usage of SVR in prediction is similar as GP but just need to sum over support vectors in Eq. 12 rather than all the training samples in Eq. 8.

4 Data Description

4.1 Data generation

We have a dataset of around 600 structures. Every structure has a 24 descriptors and the DFT calculated energy which is treated as the ground truth energy. We use the package VASP [7, 8] to perform DFT calculations with the projector augmented wave (PAW) [13,14] method utilizing the PBE generalized gradient approximation [15] as the exchange-correlation functional. We subtract the reference energies of B_4C and $B_{13}C_2$. The final energy in the dataset is within 0 to 600 meV/cluster, where a cluster contains 15 atoms.

The first feature is the carbon concentration, the remaining 23 features are the numbers of C-C pairs at certain discrete distances. Since the structures are of different size (mainly of sizes 2×2×2 and 3×3×3), we normalize all the features and the energies to be values per cluster independently. After that, we normalize all the features to be between 0 and 1 independently.

4.2 Exploratory data analysis

The histograms and pairwise scatter plots of three selected features (C concentration, number of nearest C-C bonds, number of second nearest C-C bonds) and energy are shown in Fig. 3. All the variables are screw distributed due to the physical distribution of the boron carbide structures. Transformations like logarithmic or Box-Cox types could be performed. However, to maintain the interpretability we stick to the untransformed features. The distribution of energy is more normal-like than the features but still has some positive screwness. We also do not take logarithmic transform of the energy, to avoid large prediction errors in certain energy ranges. Moreover, the scatter plot shows the features are correlated but not collinear. The energy correlates with these three features. The variances of energies at different feature values are not the same. The screwed distributions and non-constant variances might bring difficulty to linear model and other models with similar assumptions.

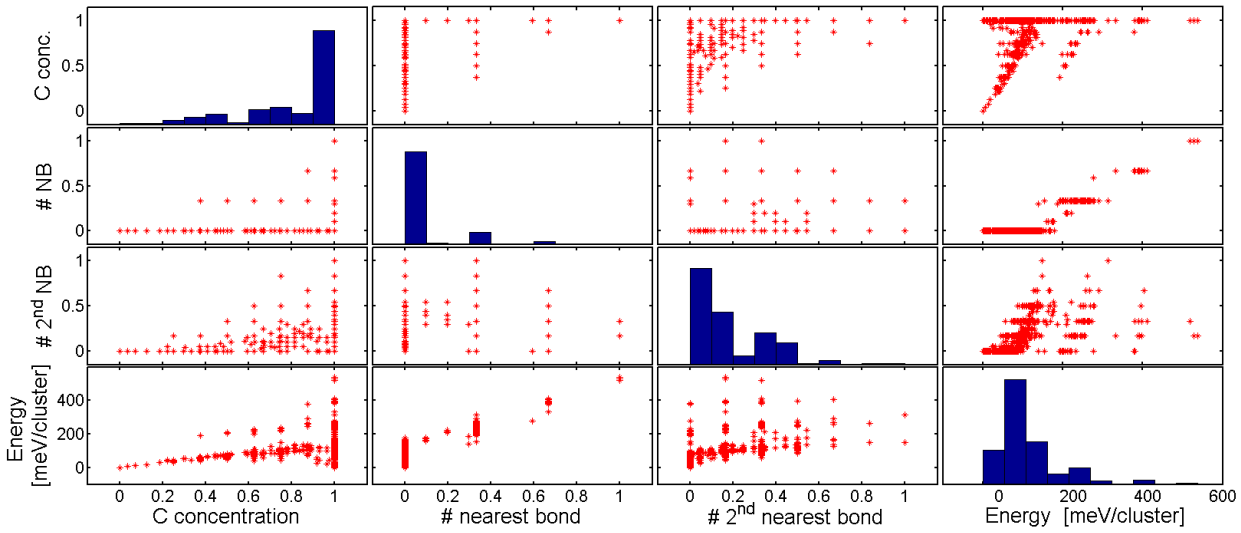

Figure 3: The histograms (diagonal plots) and pairwise scatter plots (off-diagonal plots) of data points. The variables from left to right and from top to bottom are the C concentration, number of nearest C-C bond, number of second nearest C-C bond and the ground truth energy. All the three features are normalized. Every red star represents the value of one structure.

5 Results

5.1 Performances and Analysis

We first perform 5-fold cross validation (CV) to evaluate the models. The root mean square errors (RMSEs) of CV, the improvement, and the total variance explained (TVE) of these

models are shown in Table 1. The improvement is defined as the percentage decrease of RMSE compared to the baseline linear model. TVE of a model M is defined as,

$$\mathrm{TVE(M)} = 1 - \frac{\mathrm{RMSE(M)}^2}{\mathrm{Var(E)}}, \tag{14}$$

where the RMSE is from the validation sets in CV, and Var(E) is the variance of the DFT energies of the whole dataset, which is 78 meV/cluster.

MODEL	RMSE (meV/cluster)	improvement	TVE
Linear	7.2 ± 0.06	0%	99.16±0.02%
PR2	5.8 ± 0.09	20±1%	99.46±0.02%
NN	5.6 ± 0.10	24±1%	99.49±0.02%
GP	**4.8 ± 0.10**	**33±1%**	**99.63±0.02%**
SVR	4.9 ± 0.08	32±1%	99.61±0.02%

Table 1: RMSE of 5-fold cross validation, improvement and TVE of different models. The standard error of these quantities are obtained from the statistics of ten times of cross validation.

The second order polynomial regression with L_1 penalty (PR2) outperforms the baseline linear model by a decrease of 20% in RMSE error. The RMSE minimizes at 197 nonzero parameters.

The neural network (NN) performs similarly to PR2 in CV. The best performing NN has 24 input nodes (features), one hidden layer of 3 or 4 nodes with "tanh" activation function and a single-node output layer with linear activation function. We used the Bayesian regularization of the parameters.

The nonparametric GP and SVR models decreases the CV error by around 33%. Since GP has a probabilistic interpretation, we choose the hyperparameters by maximizing the likelihood of the training data. However, SVR does not have such interpretation, thus we perform an extensive search over the grid of hyperparameters to find the set of hyperparameters that minimizes the 5-fold CV error.

The goodness of fit is shown in Fig. 4 by comparing the predicted energies of the validation sets in the 5-fold CV with the corresponding DFT energies. The points generally lie near the $y = x$ line. Both GP and SVR perform better than the baseline linear model. To illustrate the fine details, we only show structures with energies upto 200 meV/cluster rather than the maximum energy of around 600 meV/cluster. GP and SVR predictions are slightly different and can be used as a cross check to identify suspicious predictions.

To further compare the performance of linear model and GP, the residuals of the validation sets in the 5-fold CV are shown in Fig. 5. The residuals of linear models are generally larger than residuals of GP. More obvious patterns exist in the residuals of the linear model which indicates underfitting. Moreover, the variance of the linear model residuals is clearly not constant for different feature values.

Since we need to predict energies of large cell structures in our Monte Carlo simulation, we also study the performance of our models when generalize to large cells. We use 2x2x2 and 3x3x3 cell structures as the training set and the remaining 12 larger cell structures (3x3x4 and 4x4x4) as the generalization set. The generalization error are 6.5, 6.9 and 12.9 meV/cluster for SVR, GP and the linear model respectively. All these models have larger generalization errors than the CV errors of the whole dataset. This difference is less pronounced for SVR and GP. Moreover, GP and SVR have generalization error 47% and 50% smaller than the linear model.

We perform a greedy stepwise feature selection, as shown in Fig. 6 where the 24 features yield the smallest CV errors. Note that the CV error is still decreasing near 24 features, which suggests that our description of the structures with these 24 features is insufficient and further improvement could be made by adding more effective features. Moreover, Fig. 6 shows that both GP and SVR CV errors are insensitive to the choice of kernel.

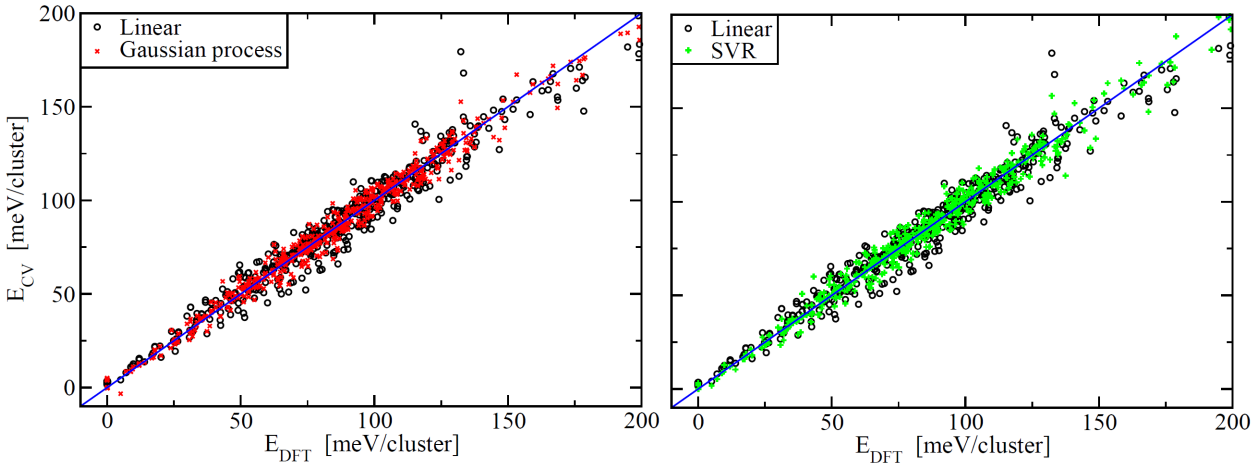

Figure 4: The predicted energies in 5-fold CV vs the ground truth DFT energies. The blue lines are $y = x$ line. Left panel shows linear model (black) and GP (red) predictions, and the right panel shows linear model (black) and SVR (green) predictions.

5.2 Model Selection and Acceleration

Since the CV error is insensitive to the choice of kernel, we use GP with polynomial kernel of degree two to predict the energies fast enough for the Monte Carlo themodynamics simulation. In the Monte Carlo simulation, millions to billions structures are generated one-by-one sequentially and energies has to be predicted one by one. Directly using Eq. 8 or Eq. 13 to predict is very slow since to predict one energy we have to sum over the whole training set (~ 600) or all support vectors(~ 400). It thus takes several hundreds inner products of 24-dimensional vectors to predict one energy. However, we realize with polynomial kernel of degree two, we can use a change of summation order trick to accelerate the prediction,

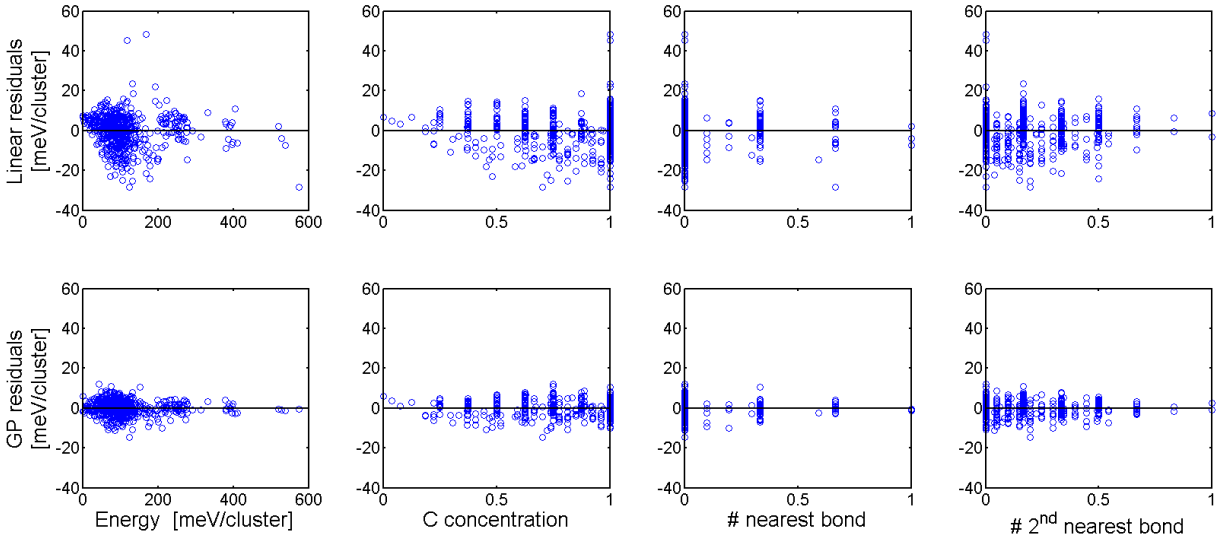

Figure 5: The residuals of predicted energies in 5-fold CV of linear model (top) and GP (bottom). From left to right the x-axis represents the DFT energy, carbon concentration, number of nearest bonds and number of second nearest bonds, respectively.

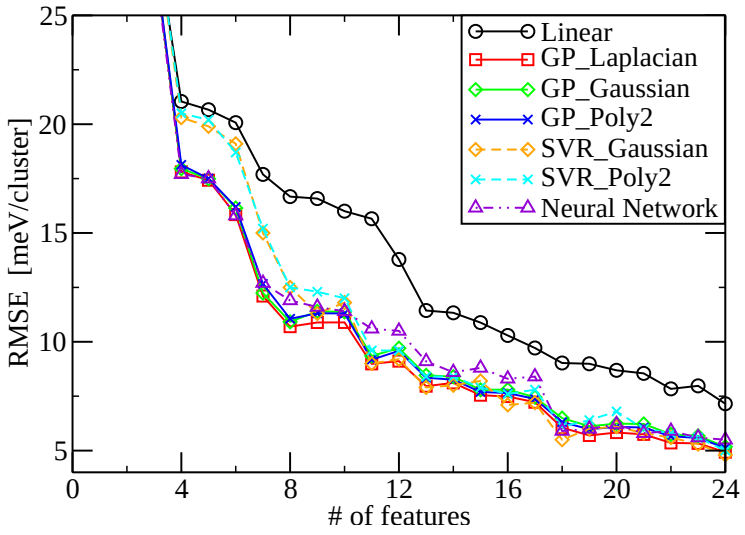

Figure 6: RMSE of linear model, GP and SVR vs number of features. The black curve is linear model, solid curves are GP with different kernels. Dashed curves are SVR with different kernels.

which is essentially rewriting the GP prediction in a parametric form.

Define $\vec{\alpha} = \Sigma_{tt}^{-1} E_{\text{train}}$ which can be easily calculated offline before the Monte Carlo

simulations, the prediction can be rewritten as,

$$E(\vec{N}_l) = \sum_{m=1}^{M} \alpha_m k(\vec{N}_l, \vec{N}_m) = \sum_{m=1}^{M} \alpha_m (1 + \beta \vec{N}_l \cdot \vec{N}_m)^2 = c + \vec{v} \cdot \vec{N}_l + \vec{N}_l^T A \vec{N}_l, \qquad (15)$$

where $c = \sum_{m=1}^{M} \alpha_m$, $\vec{v} = \sum_{m=1}^{M} 2\beta\alpha_m \vec{N}_m$ and the matrix $A = \sum_{m=1}^{M} \beta^2 \alpha_m \vec{N}_m \vec{N}_m^T$. Since c, $\vec{v}$ and A can be easily calculated offline using the training set, the online calculation to predict one structure only needs 25 vector multiplications, which is 30 times fewer than directly using Eq. 8. In practice, the prediction is fast enough for Monte Carlo simulation.

5.3 CV error vs Data Size

To examine whether our data set is large enough, we calculated the CV errors using parts of the whole dataset. In Fig. 7, the CV error of linear model saturates at around 450 data points. In contrast, the CV errors of NN, GP and SVR are still decreasing with more data which suggests that the dataset is not large enough for these three models. The CV error and presumably the generalization error can be further reduced by effectively adding new data points.

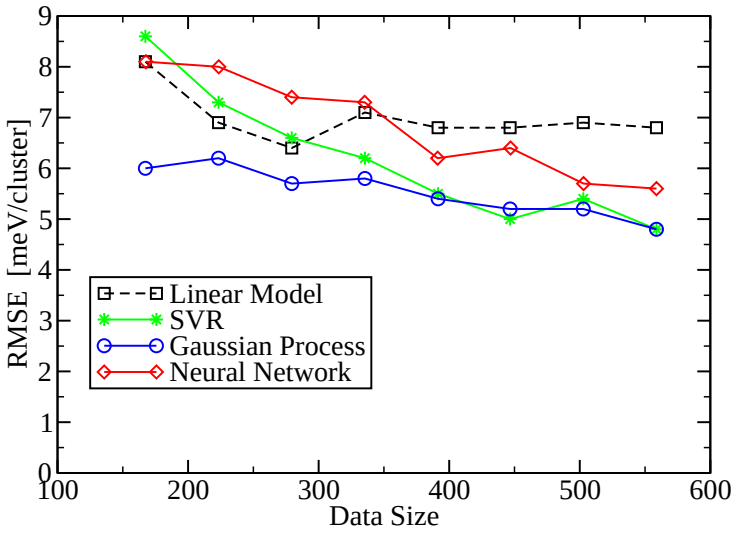

Figure 7: 5-fold CV error vs the data size. The dashed black curve is linear model, the green curve is SVR, blue is GP and red curve is NN.

6 Improvement

6.1 More Features

As shown in Fig. 6, the CV errors are not saturated with the number of features. As a direct generalization, we count the numbers of C-C pairs with distance between 6.5 Å and 9 Å and add them as new features. The total number of features is 51 in this new dataset. The CV errors with different number of features are shown in Fig. 8. The CV error of GP decreases from 4.8 to 3.6 meV/cluster (23%) and CV error of linear model decreases from 7.2 to 5.8 meV/cluster (20%). Fig. 8 shows we almost reach the limit of using this type of features. However, since the DFT uncertainty is still less than our CV error, our model might be further improved by adding other effective features. As a future work we plan to add in many-body interaction-related features. The interaction between nuclei are coulomb interactions which is a two-body interaction. However, the surrounding electrons effectively induce many-body interactions between the nuclei. Although the product of two pairs includes some information relating to many-body interaction at the whole supercell level, we expect adding selected descriptors that directly relate to many-body interactions at local structural level could possibly improve the models.

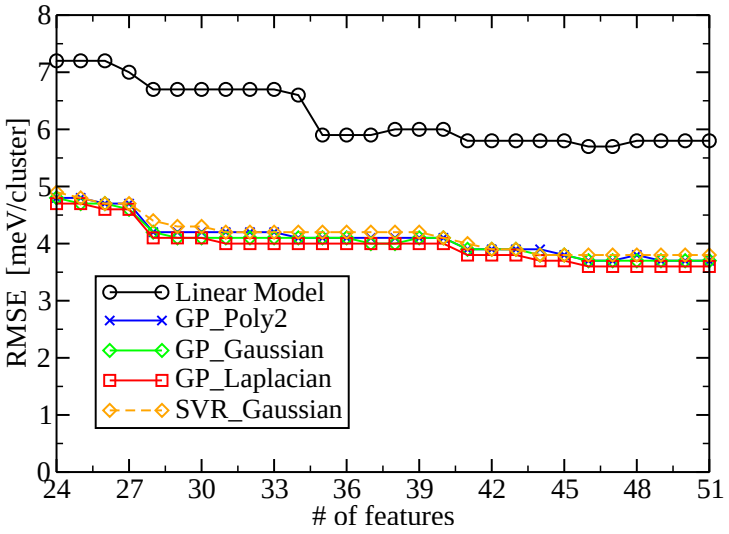

Figure 8: CV errors of linear model (black), SVR with Gaussian kernel (orange) and GP with three different kernels with 24 to 51 features.

6.2 Improve dataset

Since as shown in Fig. 7 the CV error has not saturated yet with the whole dataset. Moreover, since we need accurately predicted energies in the whole structural space for Monte

Carlo simulation, we need more data to sample the blank regions of the structural space. One effective way of adding data is guided by GP using a large set of structures with unknown energies. GP not only predicts the energies of structures but also provides the standard error σ of every prediction. A structure with large σ implies a low density of data in that region. Adding that structure to the dataset by calculating its energy with DFT will improve the fit.

Our recipe to add data to the dataset is to first generate a large number of structures ($\sim$1 million) through Monte Carlo simulation, and then use the GP model to predict the energies and σ's of the energies. We then pick out the structures with large predicted σ's and calculate their energies with DFT and finally add them into the dataset. We can perform this update of the dataset iteratively. An example of the histogram of predicted σ's of unlabeled data is shown in Fig. 9. The majority of structures have σ smaller than 5 meV/cluster, the CV error of the model. We thus obtain 207 structures with σ larger than the RMSE of the model as candidates to perform DFT calculations. We pick the 39 largest σ structures from several simulations, calculate them and add them to the dataset. By always using these 39 points as training data, the CV RMSE decreases slightly with GP and increases slightly with linear model. This discrepancy might be because the linear model does not have the flexibility to learn those hard points well. We can still add in more data until the CV error saturates and the standard errors of GP predictions are all small.

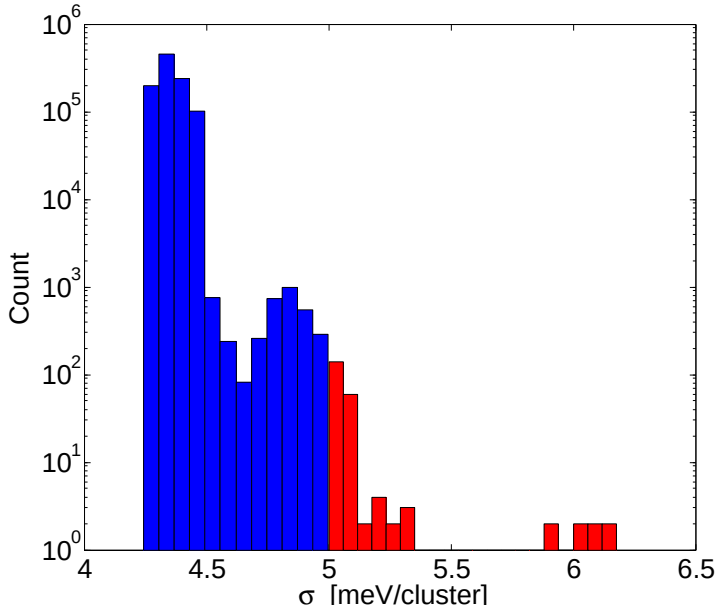

Figure 9: The histogram of the predicted standard error of the one million unlabeled structures. The Y-axis is in log scale. The red bars are for structures with predicted standard errors larger than 5 meV/cluster.

7 Monte Carlo Simulation

In the MC simulation, we find our current GP model predicts nonphysically low energies to some structures. This is probably because our dataset size is limited and GP is too flexible to be well extrapolated. We thus need to modify our GP model. One observation is that the GP model with polynomial kernel of degree two is similar to a parametric polynomial regression of degree two (PR2). We can try less flexible PR2 models with physics-guided feature selections to achieve similar accuracy with good extrapolation results. For example, physically we know that the interaction between C-C nearest bonds and C-C second nearest bonds is much more important than the interaction between C-C nearest bonds and C-C 10th nearest bonds. We thus start with the 24 features N_i and add the products N_iN_j successively as new features for $i = 0, 1, 2$ and $j = i, i + 1, ..., 10$. The 41-feature PR2 model has smallest CV error, which is 5.2±0.1 meV/cluster, only 0.5 meV/cluster higher than the GP model. The generalization error to the 12 large cell structures is 7.6 meV/cluster, 1 meV/cluster larger than GP but 6.3 meV/cluster smaller than linear model. Moreover, the 41-feature PR2 model does not predict nonphysical energies in our simulations.

We use the 41-feature PR2 model to perform Monte Carlo (MC) simulations to calculate the phase diagram of boron carbide and compare with the linear model results. In the MC simulation, the conventional Metropolis method is used. We accept a new trial structure with probability,

$$\min\{1, \exp(\frac{E_1 - E_2 + \mu(Nc_2 - Nc_1)}{k_BT})\}, \tag{16}$$

where E_1 and Nc_1 are the energy and the number of carbon of current structure respectively, E_2 and Nc_2 are the energy and the number of carbon of the new trial structure, μ is the chemical potential difference of carbon and boron, k_B is the Boltzmann constant, and T is temperature.

The MC simulation and analysis is similar as in [16]. We perform MC simulation at given T and μ to obtain how many structures occur at certain energies, which we call the histogram at T and μ. We simulate at various T's and μ's and make sure nearby histograms are well overlapped. Histograms of the linear model and the 41-feature PR2 model with $T = 600K$ and different μ's are shown in Fig. 10. The μ's we use is in the unit of k_BT. Rapidly changing histograms, or multiply-peaked histograms indicate a possible phase transition. The histograms of the linear model indicate a phase between μ =0.8 and 1.0. The 41-feature PR2 model indicate two phase transtitions, one between μ =0.8 and 1.0 and the other between μ =-0.3 and 0.0. To evaluate which model is physically better, we need to study the order parameter and the symmetry breaking path which are still ongoing.

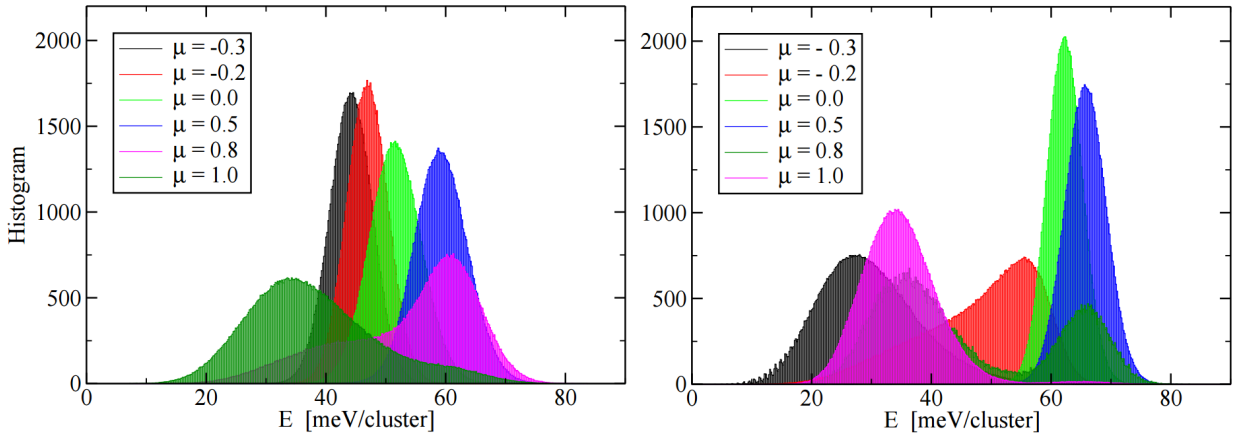

Figure 10: Energy histograms of the linear model (left) and the 41-feature PR2 model (right).

8 Conclusion

In this project we performed data analysis on a dataset of boron carbide structures with quantum mechanical (DFT) calculated energies. Constructing an accurate energy model is crucial for the Monte Carlo simulation of thermodynamics and phase diagram of boron carbide. We started with a physics motivated linear model as the baseline. We then exploit the nonlinear interaction between features using parametric models like an L_1-penalized polynomial model, neural network and nonparametric models like Gaussian process and support vector regression.

Comparing with the baseline linear model, our result shows that the L_1-penalized polynomial model and neural network decrease the cross validation error by 20% and 24% respectively but not as much as GP and SVR of around 33%. The accuracy of GP and SVR is insensitive to the choice of kernel in our problem. We chose the GP with polynomial kernels with degree two to adapt to the requirement of fast prediction in our Monte Carlo simulation. We also observed that the cross validation error are not saturated with the number of data and number of features. We made further improvements by adding more features and by effectively adding training data guided by the Gaussian process predicted standard error. We performed Monte Carlo simulations and found a possibly new phase transition unseen from the linear model. As future work, we plan to add many-body interaction-related features and more DFT calculated structures in the dataset. We also plan to generalize our model to the extended physics problem of allowing chain substitutions.

9 Acknowledgement

I would like to thank my committee members Jeff Schneider, Michael Widom, and Geoff Gordon, for their advice, support, and help. I would like to thank Sanxi Yao and Michael Widom for providing the dataset. Financial support from the McWilliams Fellowship and

the ONR-MURI under Grant No. N00014-11-1-0678 is gratefully acknowledged.

10 Reference

[1] V. Domnich, S. Reynaud, R.A. Haber, M. Chhowalla, Boron carbide: structure, properties, and stability under stress, J. Am. Ceram. Soc. 94 (11) (2011) 3605-3628.

[2] P.F. Rogl, J. Vrestal, T. Tanaka, S. Takenouchi, The B-rich side of the B-C phase diagram, Calphad 44 (2014) 3-9.

[3] K.A. Schwetz, P. Karduck, Investigations of the boron-carbon system with the aid of electron probe microanalysis, J. Less Common Met. 175 (1991) 1-100.

[4] H. Okamoto, B-C (boron-carbon), J. Phase Equilib. 13 (1992) 436.

[5] Hohenberg P, Kohn W (1964) Inhomogeneous electron gas. Phys Rev 136(3B):864871.

[6] Kohn W, Sham L, (1965) Self-Consistent Equations Including Exchange and Correlation Effects. Physical Review 140 (4A): A1133A1138.

[7] G. Kresse, J. Hafner, Ab initio molecular dynamics for liquid metals, Phys. Rev. B 47 (1993) 558.

[8] G. Kresse, J. Furthmuller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B 54 (1996) 11169.

[9] de Fontaine D (1994) Cluster approach to order-disorder transformation in alloys, Solid State Physics 47:33.

[10] Ducastelle F (1991) Order and phase stability in alloys, Elsevier.

[11] C. Bishop, Pattern Recognition and Machine Learning, 2007, Springer .

[12] Carl Edward Rasmussen, Gaussian processes for machine learning, 2006, MIT Press.

[13] P.E. Blochl, Projector augmented-wave method, Phys. Rev. B 50 (1994) 17953.

[14] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B 59 (1999) 1758.

[15] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (1996) 3865.

[16] S. Yao, W.P. Huhn, M. Widom, Phase transitions of boron carbide: Pair interaction model of high carbon limit, Solid State Sciences, 1293-2558 (2014).