Bayesian Active Machine Learning for Cluster Expansion Construction

Hantong Chen[a], Sayan Samanta[a], Siya Zhu[a], Hagen Eckert[b], Jan Schroers[c], Stefano Curtarolo[b], Axel van de Walle[a]

[a]*School of Engineering, Brown University, Providence, RI 02912, USA*
[b]*Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27708, USA*
[c]*Department of Mechanical Engineering and Materials Science, Yale University, New Haven, CT 06511, USA.*

Abstract

The Cluster expansion (CE) is a powerful method for representing the energetics of alloys from a fit to first principles energies. However, many common fitting methods are computationally demanding and do not provide the guarantee that the system's ground states are preserved. This paper demonstrates the use of an efficient implementation of a Bayesian algorithm for cluster expansion construction that ensures all the input structural energies are fitted perfectly while reducing computational cost. The method incorporates an active learning scheme that searches for new optimal structures to include in the fit. As performance tests, we calculate the phase diagram of the Fe-Ir system and study the short range order in an equimolar MoNbTaVW system. The new method has been integrated into the Alloy Theoretic Automated Toolkit (ATAT).

Keywords: cluster expansion, Bayesian algorithm, active learning

1. Introduction

Cluster expansion (CE) is a powerful method enabling thermodynamic modeling of alloys from first principles. To achieve a good CE fit, researchers have consistently employing new algorithms in clusters expansion. These algorithms include compressive sensing [1, 2], group lasso [3], ridge regression [4], quadratic programming [5] and linear programming [6]. As the idea of applying Bayesian methods in computational material science becomes increasingly popular, multiple groups have proposed using such methods [7–9] for CE construction. The main difference is that Bayesian methods explicitly formalize how a priori information (known before the input data is observed) is used in the fitting procedure. Here, we build upon the method proposed in [9], which offers major advantages over other fitting methods. First, our method ensures that the energies of all ground state structures in the training set are fitted perfectly, as ground state structures are important in applications such as phase diagram calculations. While methods based on quadratic or linear programming [5, 6] share this property, we seek improvements along other dimensions. For instance, a second advantage of our method is that it enjoys favorable convergence properties as it allows users to incorporate physics-based priors on the magnitude of the effective cluster interaction (ECI). Third, the output ECI is guaranteed to depend smoothly on the input data, which is important for uncertainty quantification purposes. Fourth, the method is computationally efficient, because enumerative searches for the "best" model (or, more generally, non-smooth optimization problems) are replaced by a smooth optimization problem. In this paper, we propose a number of improvements over this approach. First, we propose a procedure based on the multivariate optimization of the hyperparameters of the prior to improve the predictive power of the CE, as measured by the cross-validation (CV) score. Second, we devise an efficient algorithm to calculate the CV score in linear time. Third, we use an active machine learning scheme that autonomously searches for new data points to incorporate in the CE training set. Finally, all mechanisms described above are integrated into the latest version of ATAT [10–18], which is an atomistic simulation toolbox containing multiple functions performing thermodynamic calculations including cluster expansion. The new algorithm is built as an extension of the MIT Multicomponent Ab initio Phase Stability code (`mmaps`) through the plug-in functionality [19]. This, by construction, ensures that the input and output files are maintained in their original format, thus allowing users to seamlessly adopt the new method.

2. Method

2.1. Cluster expansion method in general

In the cluster expansion method, the energy $E(\sigma)$ (per atom) of an atomic configuration σ is expressed as

$$E(\sigma) = \sum_{\alpha} m_\alpha J_\alpha \langle \sigma_\alpha \rangle \tag{1}$$

In equation (1), the summation is over clusters α and, for each of them, J_α is the ECIs while m_α is cluster multiplicity, which indicates the number of clusters that are equivalent by symmetry to α (divided by the number of lattice sites). The expectation $\langle \sigma_\alpha \rangle$ depends solely on the

atomic configuration σ and encodes the correlations between atomic occupations of the sites on all clusters symmetrically equivalent to α. Its specific expression for arbitrary multi-component multisublattice alloys can be found in [18]. Currently, `mmaps` can perform cluster expansion using that specific expression with a least square fit method, which is also the method we use in the paper as a comparison.

Equation (1) is a convergent expansion of the energy $E(\sigma)$ of any structure σ in the limit where all clusters are included. However, in practice, the cluster expansion function is typically truncated to a finite number of terms. Our new method avoids an excessive dependence on the chosen truncation point by including many more terms than in a traditional least square method and instead forcing ECIs to converge to zero gradually as the size of the cluster increases through a Bayesian prior. This form of regularization is the key mechanism allowing a perfect fit to the data without sacrificing predictive power.

2.2. Bayesian method

2.2.1. Basic formulas

Our method relies on Bayes's theorem:

$$P[J|E] = \frac{P[E|J]P[J]}{P[E]} \tag{2}$$

where the prior probability density of the ECIs is denoted as $P[J]$, which embodies our prior knowledge regarding the possible magnitude of the ECIs. Here, J is an $n_J \times 1$ matrix containing the ECI J_α for all n_J clusters. Also, E is the $n_E \times 1$ energy vector containing n_E energies obtained from ab initio calculations. In this equation, P[E] is a normalization constant and $P[E|J]$ is a delta function requiring the CE to predict energies exactly. Thus, to find best ECI values, we maximize the posterior probability $P[J|E]$ given the constraint of $P[E|J]$. Note that this expression assumes that the number of ECIs n_J is larger than the number of reference structures n_E, thus enabling the known structural energies to be fitted exactly. This approach can be seen as Bayesian inference in the limit where the likelihood function $P[E|J]$ is degenerate.

For simplicity, we consider a Gaussian prior

$$\begin{aligned} P[J] &= \prod_\alpha (2\pi)^{-1/2} w_\alpha^{-1} \exp(-w_\alpha^{-2} J_\alpha^2) \\ &\propto \prod_\alpha \exp(-w_\alpha^{-2} J_\alpha^2) \end{aligned} \tag{3}$$

where the widths w_α are proportional to how large we expect each ECI J_α to be. A more specific physics-based choice of widths parameter w_α will be motivated later. Note that the ECI can take both positive and negative values, and the fact that the prior centered at $J_\alpha = 0$ helps ensure that the fit won't include an unnecessarily large number of ECIs. Using a normal distribution prior will also simplify the structure selection algorithm, since with this specific choice of prior, our method can be seen as a form of Gaussian Process Regression [20]. We need to maximize the posterior probability under the constraints of the known energies, and the question becomes finding J_α that maximize

$$\prod_\alpha \exp(-w_\alpha^{-2} J_\alpha^2) \tag{4}$$

subject to the constraint

$$\sum \rho_{i\alpha} m_{i\alpha} J_\alpha = E_i \tag{5}$$

for $i = 1, ..., n_E$ and where $\rho_{i\alpha}$ are the correlations $\langle \sigma_\alpha \rangle$ associated with cluster α for structure i, and $m_{i\alpha}$ are the cluster multiplicities. Observing that maximizing the logarithm of the probability is equivalent, dropping irrelevant constants and introducing Lagrange multipliers, the Lagrangian for this constrained maximization problem is

$$L = -\sum_\alpha w_\alpha^{-2} J_\alpha^2 - 2\sum_i \lambda_i \left(\sum_\alpha \rho_{i\alpha} m_\alpha J_\alpha - E_i \right) \tag{6}$$

And the solution for J_α and λ_i is found by solving a system of linear equations which can be cast in matrix form as:

$$W^{-1} J + R^T \lambda = 0 \tag{7}$$

$$RJ = E \tag{8}$$

where the matrix W contains w_α^2 on the diagonal, λ is the $n_E \times 1$ vector of Lagrange's multipliers and R is the $n_E \times n_J$ correlation matrix containing all $\prod \sigma_\alpha$. Note that this system has as many unknowns as there are equations and the solution can thus be expressed in closed form:

$$J = WR^T \left(RWR^T \right)^{-1} E \tag{9}$$

Defining $\tilde{J} = W^{-1/2} J$ and $\tilde{R} = RW^{1/2}$, this can be rewritten as

$$\tilde{J} = \tilde{R}^T \left(\tilde{R}\tilde{R}^T \right)^{-1} E \tag{10}$$

In equations (9) and (10), the inversion exists only when the rows in the correlation matrix R are linearly independent, and in our code we have constructed a whole procedure to ensure that this inversion is valid all the time.

2.2.2. ECI Optimization

Based on equation (9), we see that with each choice of weighting matrix W, there is a different ECI vector that maximize the probability $P[J]$, which brings the question of how to choose the "best" W. To address this, we rely on (i) a physics-based parametrization of W and (ii) statistical measures of out-of-sample predictive power.

We expect interaction strength to decay with a cluster's spatial extent and number of sites and, accordingly, we propose the following prior widths:

$$w_\alpha = b^{|\alpha|} \prod_{r_{ij} \in \alpha} f(r_{ij}) \tag{11}$$

$$f(r_{ij}) = \min\{a^k (r_{ij} + c)^{-k}, 1\} \tag{12}$$

where $|\alpha|$ denotes the number of sites in cluster α and $r_{ij} \in \alpha$ denotes all pairwise distances between sites of cluster α. The adjustable hyperparameters are (i) $b \in [0, 1]$, controlling the decay as a cluster has more sites, (ii) a and c, controlling the slowly-varying short-range behavior and (iii) k, controlling the long-range decay. In addition, the function $f(r_{ij})$ is set to have a maximum of 1 to ensure that larger clusters will always have a smaller prior. For the empty and point clusters we set w_α to 1.

To optimize the hyperparameters, the leave one out cross validation (LOOCV) score [18] is used:

$$(CV)^2 = n^{-1} \sum (E_i - \hat{E}_{(i)})^2, \tag{13}$$

where $\hat{E}_{(i)}$ is the energy predicted using the fit that includes all the structures except structure i. While evaluating the CV score apparently requires re-fitting n times, it is possible to compute the CV score in order n operations. Such an algorithm is well-known for standard least-squares (see, e.g. [18]) but takes a rather different expression in the context of the specific Bayesian scheme we propose:

$$(CV)^2 = n^{-1} \sum_{i=1}^{n} \left(\frac{(BE)_i}{B_{ii}} \right)^2, \tag{14}$$

where $B = (\tilde{R}\tilde{R}^T)^{-1}$ only needs to be calculated once and $(BE)_i$ denotes the i element of the vector BE. In Appendix Appendix A.1, we derive this expression, which demands a completely different method of proof than for standard least-squares.

For any given values of the adjustable hyperparameters, the CV score can be computed for a given set of structures. These parameters can be optimized to improve the quality of the CV score. In our implementation, the 3 adjustable parameters are optimized using the Nelder-Mead [21] algorithm, as the gradient is difficult to compute. Boundary constraints ($b \in [0, 1]$, $c \geq 0$) are imposed by simply forcing the vertices of the simplex to move back into the allowed region if they ever violate the constraints.

Even though, in our approach, the number of ECIs could theoretically be infinite, this number must be truncated to a finite number for the purpose of numerical tractability. Note, however, that this truncation excludes ECIs that would already have had a very small value by construction due to the prior, thus having little effect on the statistical properties of the method. Here we provide heuristics to automatically determine the truncation point (which can change as more structures are added to the fit — the code generate new clusters on-the-fly if needed). First, the minimum number of ECIs to include increases as the number of atomic species types in the system increases. Second, all clusters with the same number of sites and same diameter will be jointly included or excluded in the optimization, and larger cluster will be included only after all smaller sub-clusters are included. Third, if an m-body cluster is included, any other m-body cluster with smaller diameter must be included as well. Finally, we found that keeping the number of ECIs at least twice the number of structures is typically sufficient to prevent the correlations of multiple structures from being linearly dependent.

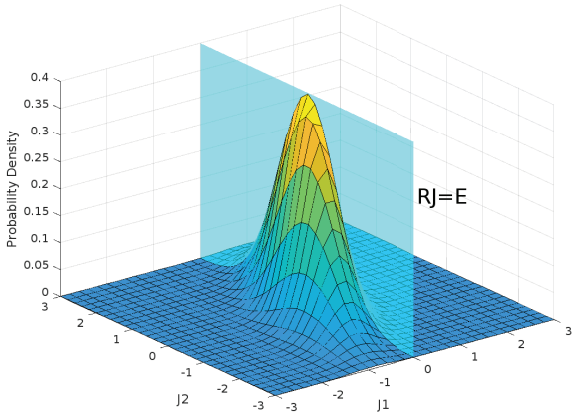


Figure 1: This figure exemplifies how the energy constraint $RJ = E$ works on the multivariate normal distribution $P[J]$. The energy constraint acts as a cross-section of the distribution and the remaining distribution is still a normal distribution

2.2.3. Structure selection based on variance reduction

Our new structure selection algorithm is in the same spirit as ATAT's original approach, but significant modifications were needed in the implementation due to the different statistical properties on the Bayesian approach.

As a starting point, the code includes the energies of all elemental end members in the system, as these are used in the calculation of formation energies. After obtaining the energies of all pure element structures, the code starts adding the smallest structures to the optimization until the number of structures reaches the minimum needed to fit a first-nearest-neighbor cluster expansion.

Then, the code performs active learning by selecting new structures based on a variance reduction criterion. The joint prior probability of the ECIs, $P[J]$, is a multivariate normal distribution. As the ECI are independent under the prior, their covariance matrix Σ is just a diagonal matrix containing all ${w_\alpha}^2$ on its diagonal. When the energy constraint is applied, the ECIs have to satisfy $RJ = E$. The new distribution will remain a multivariate normal distribution, but only defined a particular subspace of $\mathbb{R}^n$. This is exemplified in Figure 1. There thus exists a $n \times n$ projection matrix P, which projects the unconstrained ECI matrix to the constrained matrix. Therefore, we can express the posterior covariance matrix $\Sigma_{\text{post}} = P\Sigma P^T$, where

$$P \quad = \quad I - R^T(RR^T)^{-1}R. \tag{15}$$

Since we wish to minimize the variance of any structure picked at random in correlation space and since

$\int_{\|v\|=1} v'\Sigma_{\text{post}} v dv \propto \text{tr}\Sigma_{\text{post}}$, we use the trace of the covariance matrix as a criterion to optimize.

When considering the addition of a new structure to the fit, the code computes the reduction ΔV in $\text{tr}\Sigma_{\text{post}}$ and estimate the computational cost C for obtaining the energy of the structure (the code uses a simple order of magnitude estimate $C = N^3$, where N is the number of atoms in the unit cell). The structure that maximizes the ratio $\Delta V/C$ is chosen for addition. It can be shown that, if the structures are generated in increasing order of size N, it is possible to identify the best structure among the infinite set of structures while only iterating over a finite number of them. The maximum possible ΔV, denoted $\Delta V_{\max}$, is bounded by the largest eigenvalue of Σ_{post}. If one has already found a structure of size N_0 with certain ΔV_0, it follows that any structure of size

$$N > \left(\frac{\Delta V_{\max}}{\Delta V_0}\right)^{1/3} N_0 \tag{16}$$

cannot lead to a better $\Delta V/C$ ratio.

In addition to this variance reduction criterion, the code prioritizes the validation of new predicted ground states, since ground state structures are key determinant of phase diagram topology. When the code searches for possible structures to add, it first constructs a convex hull of the energies for all known structures at the current stage of the algorithm. Then, it generates a large number of new structures (up to a user-specified size), and test whether these new structures break the convex hull, thus indicating that they are candidate ground state structures. If new ground states are predicted, one with the largest $\Delta V/C$ ratio is added to the training set, otherwise, the search over all structures, described above, is performed.

2.3. Code usage

The new Bayesian algorithm is integrated into `mmaps` as a "plug-in" feature inside ATAT. The new algorithm is written in the format documented in `mrefine_skel.c++` file in the ATAT source file folder, which exemplifies how new fitting algorithms should be written and merged into `mmaps`. The usage of the new algorithm is identical with the original version of `mmaps` and more details on how to use the code can be found in the original `mmaps` paper [10]. To use the Bayesian cluster expansion method, the user just need to add `-fa=bayesian` to select the Bayesian method. The input files required are the `lat.in` file and `xxxx.wrap` file, and the format requirement remains exactly the same as with the original method. Further help is available by typing `mmaps -h`. A considerable advantage of this integration is that the new algorithm automatically inherits other `mmaps` functionalities, such as the availability of plug-ins for analytic long-range interaction that incorporate constituent strain [22] or electrostatic effects [23].

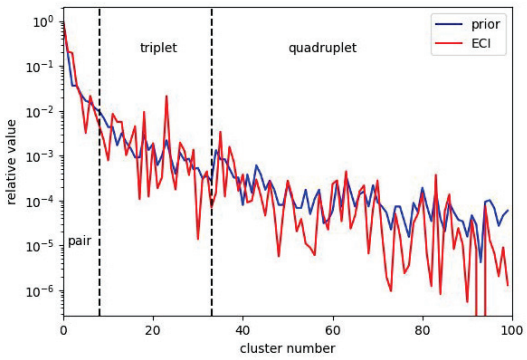


Figure 2: Normalized ECI value vs prior parameter for Mo-Ta bcc system

3. Test cases

To test the reliability of our method, we consider 8 alloy systems and compare our results with those obtained with least squares using cross-validation model selection. All ab initio calculations are performed using the Vienna Ab initio Simulation Package(VASP) [24–27], implementing the projector augmented wave (PAW) method [28, 29] with PBE functional [30] with an energy cutoff of 300eV. Suitable k-point meshes are automatically generated [18] to guarantee a density of at least 1000 k-points per reciprocal atoms. For our tests, we deliberately favor using real input energies over artificial energy generated from known ECIs, because, in the latter case, it would be too easy to select favorable generating models that closely agree with the prior by construction.

Nevertheless, it is instructive to show that the prior is indeed effective as a regularization scheme and that our specific form of prior does not clash with actual energy data. This can be done by plotting the absolute value of the fitted ECIs overlaid with the prior's width value w_α. Figure 2 shows the result of this exercise for the Mo-Ta bcc system and indicates that our method has no problem finding ECI values that are consistent with our form of postulated prior, resulting in a well-behaved decay of the fitted interactions.

3.1. Comparing Bayesian method with least squares

In this section, we test the Bayesian ECI fitting algorithm and compare it with `mmaps`' default least-squares (LS) algorithm with default parameters (see Table 1). To facilitate the comparison, the pool of possible input structural energies is the same for both methods. These input structures are generated with ATAT's default algorithm [18]. For the Bayesian method, we gradually add these structures in increasing order of cell size until we reach the same LOOCV score as the default least-squares method. (The Bayesian structure selection algorithm is not used in this section.)

The results of this exercise, reported in Table 1 for a number of alloy systems, show that the Bayesian method can achieve the same predictive power with far fewer input structures. As the target accuracy for binary and ternary systems are different, it's not very meaningful to directly compare the number of structures used for different systems. To further validate these findings, we also compute the least-squares prediction error for the structures included in the least-squares fit but not included in the Bayesian fit. These results support the fact that LOOCV provides an estimator of the out-of-sample predictive power and further confirm that the Bayesian considerably lowers the computational cost of obtaining a given accuracy. In addition, we have also compared the ECIs we have obtained using LS method and Bayesian method for the binary Mo-Ta system in Figure 3. The results show that the scale of the ECI values we have obtained using Bayesian method is close to the ECI values obtained from LS method. The main difference is that the smallest cluster of each size is replaced by a large number of clusters in the Bayesian method which enables the perfect fit.

3.2. Active learning test

To demonstrate the effectiveness of the active learning structure selection algorithm, we construct a cluster expansion from scratch for the Fe-Ir system. In this section, spin-polarization is turned on to ensure the accuracy of the calculation. Although the Fe-Ir system contains multiple lattices, we focus on the metastable phase diagram arising from fcc superstructures. This system is interesting because high-throughput ab initio calculations [31, 32] predict an fcc superstructure ground state at composition $FeIr_3$ that has not yet been observed experimentally.

At each step of the iteration towards convergence, the code performs a Bayesian fit and automatically determines the most informative structure to add to the fit for the next iteration. We have plotted the energy composition graph in Figure 3.2 to illustrate how our structure selection algorithm operates.

Our code will first add a few smallest structures, and then start ground state search based on variation reduction. If there is no possible new ground states, the code will simply look for the structure that reduce the variance the most. As shown in figure 3.2, when we add the smallest structure in 4(b), this structure is a new ground state so we construct the new convex hull with it and continue the loop. Finally, a CV score of 0.010eV is reached with only 40 structures and the associated cluster expansion comprises a total number of 165 clusters, including up to 4-body clusters.

Then, using the phb (PHase Boundary code) and memc2 (Multicomponent Eazy Monte Carlo Code) in ATAT [11, 17], we perform Monte Carlo (MC) simulations over a wide composition and temperature range and generate the Ir-rich portion of the metastable phase diagram (see Figure 4).

First, in the Ir-rich half of the phase diagram, we confirm the existence of two possible ordered ground states at 0K. The ground state at 50% is masked by an hcp solid solution [33] that is stable up to ∼900K. This leaves the $FeIr_3$ compound as a candidate new phase, which we find disorders at relatively low temperature (∼400K), thus explaining why it may not have been experimentally observed yet, given the slow kinetics as such low temperature.

This example demonstrates how, thanks to our new cluster expansion construction algorithm, a relatively small number of ab initio calculations can be used to effectively assess the relevance of high-throughput predictions of novel phases.

3.3. Short range order determination

The high-entropy alloys community is becoming increasingly aware that simple entropy estimates based on composition alone can be considerably misleading, given the likely presence of short range order (SRO) [34, 35]. The ECI obtained from our method can provide useful input to quantify SRO in these alloys. Here we have performed MC simulations on the equimolar BCC MoNbTaVW system using memc2 [16] with the ECI obtained from Section 3.1 to obtain the pair correlations functions. The temperature range we use is from 600K to 2500K as this system undergoes ordering below 600K. The correlations $\langle \sigma_\alpha \rangle$ entering the cluster expansion can be readily converted into occupation probability of different configurations on a cluster, by multiplication by the so-called V-matrix [36]. The V-matrix was generated using the cvmclus (Cluster Variation Method CLUSter generator code) in ATAT [37]. The resulting occupation probabilities can then be directly converged into the Warren-Cowley short range order parameter α_{ij} for all the nearest neighboring pairs i, j:

$$\alpha_{ij}^{(r)} = 1 - \frac{P_{ij}^{(r)}}{2c_i c_j} \tag{17}$$

where P_{ij} is the probability of pair containing atoms i and j, while c_i is the concentration of element i. The resulting SRO parameters for all 15 nearest neighbor pairs are shown below in Figures 5 and 6.

As shown in Figure 5 and 6, the resulting graph agrees well with earlier high-accuracy cluster expansion results [38] obtained with ATAT's least-squares algorithm. In particular, below 1000K the Mo-Ta pair has the most negative SRO value, followed by the V-W pair and Mo-Nb pair, which all have large negative values. Also, the Nb-V and Ta-V pair have the largest positive values close to each other above 600K, which all agrees with previous results. While this earlier study achieved a LOOCV error of about 8 meV using over 400 structures, the present study achieves a comparable accuracy (11 meV) using only 76 structures, which is a huge improvement in efficiency. Considering the fact that the structure selection algorithm in both methods tend to favor smaller systems, the computational cost of the first principles calculations can be greatly reduced.

Table 1: CV score(in eV) comparison between least-squares (LS) and Bayesian fitting

Alloy	number of structures (LS)	number of structures (Bayesian)	LOOCV score (LS)	LOOCV score (Bayesian)	out-of-sample error (Bayesian)
IrRu fcc	163	36	0.0030	0.0043	0.0023
MoNb bcc	93	34	0.0050	0.0045	0.0031
MoNbTa bcc	86	39	0.0093	0.0091	0.012
MoNbV bcc	451	36	0.0088	0.0093	0.0066
MoNbW bcc	148	33	0.0057	0.0070	0.0098
MoTa bcc	141	28	0.0096	0.0085	0.0047
MoTaW bcc	251	36	0.0082	0.012	0.0078
MoVWNbTa bcc	1154	121	0.011	0.0104	0.0110

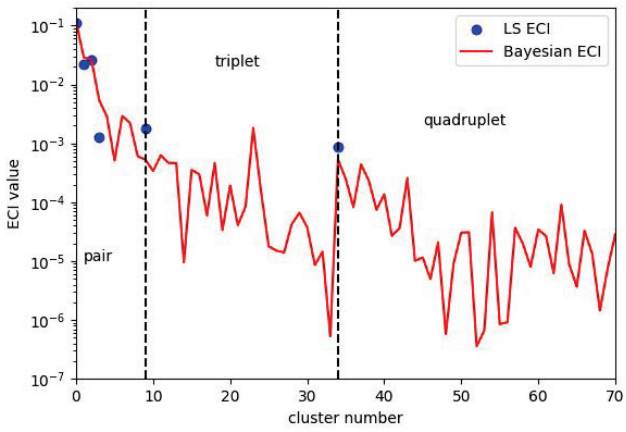
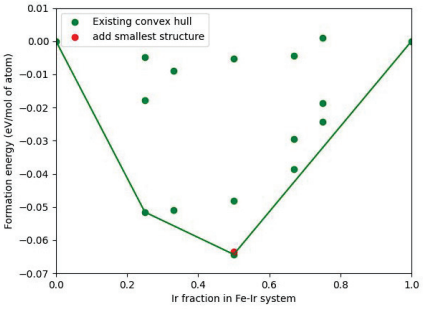
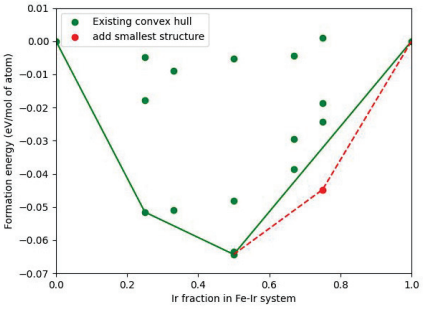


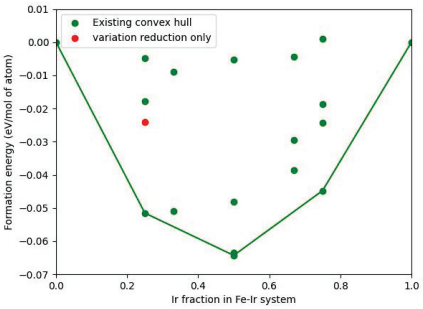
Figure 3: Comparison of ECI values obtained using LS method and Bayesian method for Mo-Ta system


Existing convex hull
add smallest structure
Formation energy (eV/mol of atom)
Ir fraction in Fe-Ir system

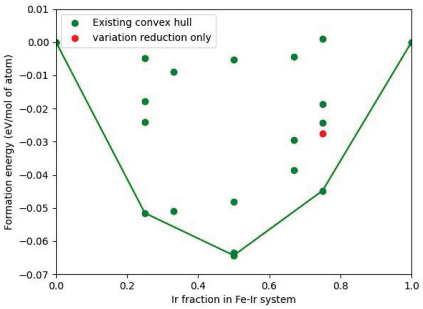
(a)


Existing convex hull
add smallest structure
Formation energy (eV/mol of atom)
Ir fraction in Fe-Ir system

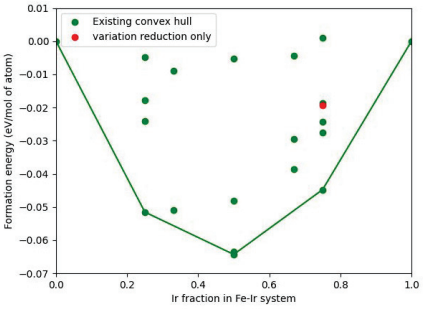
(b)


Existing convex hull
variation reduction only
Formation energy (eV/mol of atom)
Ir fraction in Fe-Ir system

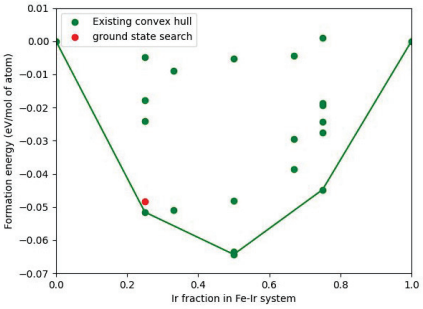
(c)


Existing convex hull
variation reduction only
Formation energy (eV/mol of atom)
Ir fraction in Fe-Ir system

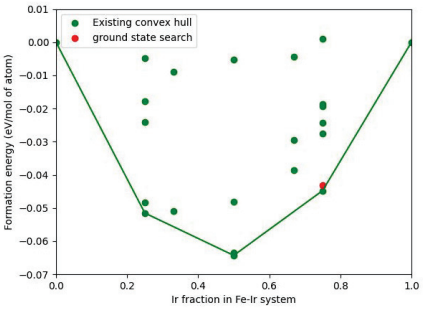
(d)


Existing convex hull
variation reduction only
Formation energy (eV/mol of atom)
Ir fraction in Fe-Ir system

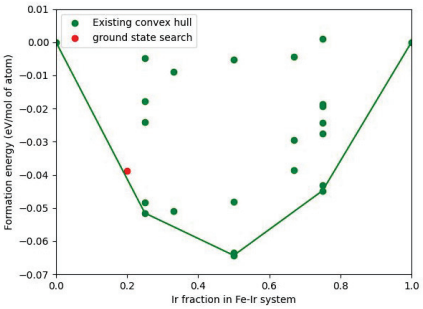
(e)


Existing convex hull
ground state search
Formation energy (eV/mol of atom)
Ir fraction in Fe-Ir system

(f)


Existing convex hull
ground state search
Formation energy (eV/mol of atom)
Ir fraction in Fe-Ir system

(g)


Existing convex hull
ground state search
Formation energy (eV/mol of atom)
Ir fraction in Fe-Ir system

(h)

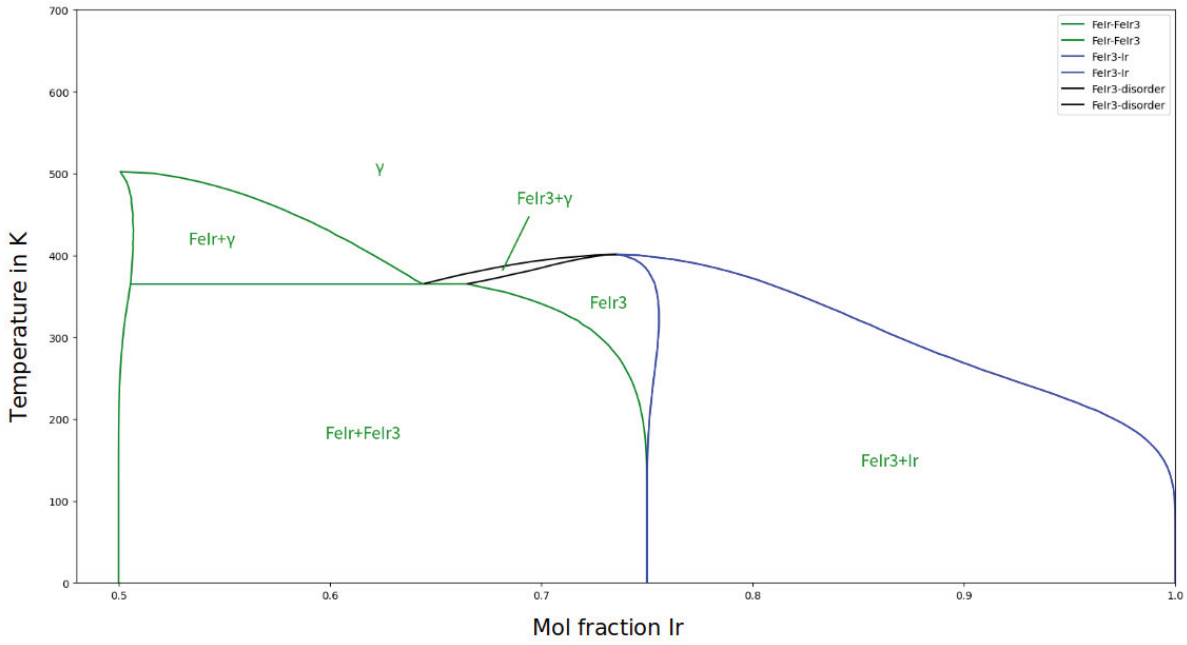


Figure 4: Ir-rich portion of Fe-Ir phase diagram

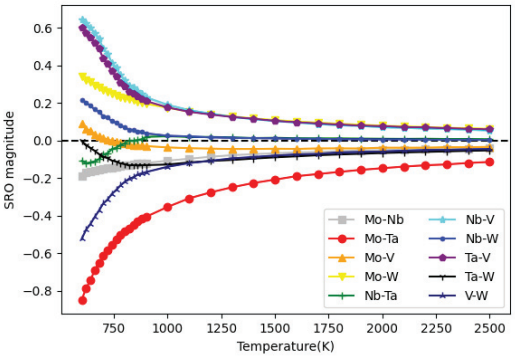


Figure 5: SRO parameter for the nearest-neighbor pairs of different elements in the MoNbTaVW system

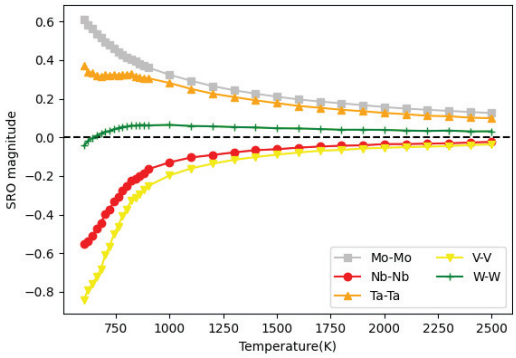


Figure 6: SRO parameter for the nearest-neighbor pairs of the same element in the MoNbTaVW system

4. Conclusion

In this paper we have described an improved implementation of a new Bayesian algorithm for cluster expansion construction. The new method exhibits major advantages over existing methods. First, the new method ensures, by construction, that the energies of predicted ground states are exactly reproduced, which considerably streamlines the thermodynamic model construction process. This is a feature that is absent from most cluster expansion construction algorithms, including the most advanced machine-learning-based [1] and Bayesian [7] schemes. As ground state structures are key determinants of a phase diagram's topology, this feature consistently improves the accuracy of calculated phase diagram. The new method also inherits the algorithms built into ATAT that prioritize ground states search, discovery and ab initio confirmation.

Second, we find that the new Bayesian method tends to require fewer structures compared to standard least-squares. In our test cases, we observe that we consistently achieve a LOOCV score comparable to least-squares while using only between 20% and 40% of the number of training structures. We also confirm by direct monitoring of the fitting ECIs, that they exhibit a physically highly plausible decay with distance and cluster size. These findings confirm that our proposed Bayesian prior acts as a very effective regularizing scheme that imposes physical plausibility requirements on the interactions. This Bayesian approach also ensures that the output model depends smoothly on the input data, since there is no discontinuous model selection step, which is important for uncertainty quantification purposes. The smoothness of the optimization problem also leads to improvement in the speed of the algorithm.

Finally, the implementation described herein fully integrates with the existing functionalities of ATAT [14, 15], thus flattening the user's learning curve. Therefore, we believe our new method provides a powerful new tool for atomistic thermodynamic modeling.

Acknowledgements

H.C., S. S., S. Z and A.V. acknowledge support from the US National Science Foundation through grant DMR-2001411, the US Army Research Office through grant W911NF-21-2-0161 and the US Office of Naval Research under grant N00014-20-1-2225. The authors also acknowledge support from the National Science Foundation under grant NRT-HDR DGE-2022040 (H.E. and S.C.) and the Office of Naval Research under grant N00014-20-1-2200 (H.E., S.C. and J.S.). Computational resources were provided by the Center for Computation and Visualization at Brown University. This work also used the Expanse and Bridges-2 clusters at UCSD and PSC, respectively, through allocation DMR010001 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants 2138259, 2138286, 2138307, 2137603, and 2138296.

Appendix A. Appendix

Appendix A.1. Efficient calculation of the CV score

Throughout this section, we let subscript $\bar{\imath}$ denote quantities obtained after removing observation i.

The energy of structure i predicted from all other structures can be expressed as:

$$\hat{E}_{\bar{\imath}} = \tilde{r}_i^T \tilde{J}_{\bar{\imath}} = \tilde{r}_i^T \tilde{R}_{\bar{\imath}}^T \left(\tilde{R}_{\bar{\imath}} \tilde{R}_{\bar{\imath}}^T\right)^{-1} E_{\bar{\imath}} \tag{A.1}$$

where $\tilde{r}_i^T$ is the i^{th} row of $\tilde{R}$. To find an efficient expression for $\hat{E}_{\bar{\imath}}$, we need to relate the inverse of a subblock, $\left(\tilde{R}_{\bar{\imath}} \tilde{R}_{\bar{\imath}}^T\right)^{-1}$, to the inverse of the full matrix $\left(\tilde{R}\tilde{R}^T\right)^{-1}$. To this effect, we re-order and partition the matrices $A = \tilde{R}\tilde{R}^T$ and $B = \left(\tilde{R}\tilde{R}^T\right)^{-1}$ as

$$\begin{aligned} A &= \begin{bmatrix} A_{\bar{\imath}\bar{\imath}} & a_{\bar{\imath}i} \\ a_{\bar{\imath}i}^T & a_{ii} \end{bmatrix} = \begin{bmatrix} \tilde{R}_{\bar{\imath}} \tilde{R}_{\bar{\imath}}^T & \tilde{R}_{\bar{\imath}} \tilde{r}_i \\ \tilde{r}_i^T R_{\bar{\imath}}^T & \tilde{r}_i^T \tilde{r}_i \end{bmatrix} \\ B &= \begin{bmatrix} B_{\bar{\imath}\bar{\imath}} & b_{\bar{\imath}i} \\ b_{\bar{\imath}i}^T & b_{ii} \end{bmatrix}. \end{aligned}$$

Using the Partitioned Inverse formula [39], we have:

$$\begin{aligned} A_{\bar{\imath}\bar{\imath}} &= \left(B_{\bar{\imath}\bar{\imath}} - b_{\bar{\imath}i} b_{ii}^{-1} b_{\bar{\imath}i}^T\right)^{-1} \\ a_{\bar{\imath}i} &= -\left(B_{\bar{\imath}\bar{\imath}} - b_{\bar{\imath}i} b_{ii}^{-1} b_{\bar{\imath}i}^T\right)^{-1} b_{\bar{\imath}i} b_{ii}^{-1} \end{aligned}$$

Substituting these expressions into Equation (A.1) using the facts that $\left(\tilde{R}_{\bar{\imath}} \tilde{R}_{\bar{\imath}}^T\right)^{-1} = \left(A_{\bar{\imath}\bar{\imath}}\right)^{-1}$ and $\tilde{r}_i^T \tilde{R}_{\bar{\imath}}^T = a_{\bar{\imath}i}^T$, we have:

$$\begin{aligned} \hat{E}_{\bar{\imath}} &= -b_{ii}^{-1} b_{\bar{\imath}i}^T \left(B_{\bar{\imath}\bar{\imath}} - b_{\bar{\imath}i} b_{ii}^{-1} b_{\bar{\imath}i}^T\right)^{-1} \left(B_{\bar{\imath}\bar{\imath}} - b_{\bar{\imath}i} b_{ii}^{-1} b_{\bar{\imath}i}^T\right) E_{\bar{\imath}} \\ &= -b_{ii}^{-1} b_{\bar{\imath}i}^T E_{\bar{\imath}} \\ &= -b_{ii}^{-1} \left((BE)_i - b_{ii} E_i\right) \\ &= -b_{ii}^{-1} (BE)_i + E_i. \end{aligned}$$

Rearranging and noting that $B_{ii} = b_{ii}$, we have

$$E_i - \hat{E}_{\bar{\imath}} = \frac{(BE)_i}{B_{ii}}. \tag{A.2}$$

Appendix A.2. Command used for phase diagram generation

To generate the phase diagram, we have calculated the phase boundary of multiple phases using different starting ground states. One example input command is
`phb -keV -gs1=0 -gs2=3 -dT=5 -er=20 -ltep=1.0e-3-dx=1.0e-3`.
Depending on the phases we are interested in the ground states -gs should be changed. Detailed documentation on the phb code can be found in reference [10].

Appendix A.3. Command used for SRO detection

Firstly, Monte Carlo simulation is performed using the ECI obtained from Bayesian method with command `memc2 -er=15 -n=100 -eq=500 -gs=-1 -keV`.
Secondly, use command `cvmclus -d` to generate the V-matrix. Detailed input file requirement can be found using command `cvmclus -h`.
Finally, perform a matrix multiplication of the V-matrix and the correlation to obtain the occupation of the clusters. Then do a dot product of the occupation of the clusters with the multiplicity of the clusters, which can be found in file clusmult.out generated by cvmclus. These operations can be done using a script.

References

[1] L. J. Nelson, G. L. W. Hart, F. Zhou, V. Ozolins, Compressive sensing as a paradigm for building physics models, Phys. Rev. B 87 (2013) 035125.

[2] J. H. Chang, D. Kleiven, M. Melander, J. Akola, J. M. Garcia-Lastra, T. Vegge, Clease: a versatile and user-friendly implementation of cluster expansion method, Journal of Physics: Condensed Matter 31 (32) (2019) 325901. doi:10.1088/1361-648X/ab1bbc.

[3] Z. Leong, T. L. Tan, Robust cluster expansion of multicomponent systems using structured sparsity, Phys. Rev. B 100 (2019) 134108. doi:10.1103/PhysRevB.100.134108.

[4] D. B. Laks, L. G. Ferreira, S. Froyen, A. Zunger, Efficient cluster expansion for substitutional systems, Phys. Rev. B 46 (1992) 12587.

[5] W. Huang, A. Urban, Z. Rong, Z. Ding, C. Luo, G. Ceder, Construction of ground-state preserving sparse lattice models for predictive materials simulations, npj Computational Materials 3 (1) (8 2017). doi:10.1038/s41524-017-0032-0.

[6] G. D. Garbulksy, G. Ceder, Linear-programming method for obtaining effective cluster interactions in alloys from total-energy calculations: application to the fcc Pd-V system, Phys. Rev. B 51 (1995) 67.

[7] T. Mueller, G. Ceder, Bayesian approach to cluster expansions, Phys. Rev. B 80 (2009) 024103. doi:10.1103/PhysRevB.80.024103.

[8] L. J. Nelson, V. Ozoliņš, C. S. Reese, F. Zhou, G. L. W. Hart, Cluster expansion made easy with bayesian compressive sensing, Phys. Rev. B 88 (2013) 155105. doi:10.1103/PhysRevB.88.155105.

[9] E. Cockayne, A. van de Walle, Building effective models from sparse but precise data: Application to an alloy cluster expansion model, Phys. Rev. B 81 (2010) 012104. doi:10.1103/PhysRevB.81.012104.

[10] A. van de Walle, M. Asta, G. Ceder, The Alloy Theoretic Automated Toolkit: A User Guide, Calphad 26 (4) (2002) 539–553. doi:10.1016/S0364-5916(02)80006-2.

[11] A. van de Walle, Multicomponent multisublattice alloys, nonconfigurational entropy and other additions to the alloy theoretic automated toolkit, Calphad 33 (2) (2009) 266–278, tools for Computational Thermodynamics. doi:10.1016/j.calphad.2008.12.005.

[12] A. van de Walle, R. Sun, Q.-J. Hong, S. Kadkhodaei, Software tools for high-throughput CALPHAD from first-principles data, Calphad 58 (2017) 70–81. doi:10.1016/j.calphad.2017.05.005.

[13] H. Liu, A. van de Walle, Rapid geometric screening of low-energy surfaces in crystals, Symmetry 14 (10) (2022) 2067. doi:10.3390/sym14102067.

[14] A. van de Walle, Q. Hong, S. Kadkhodaei, R. Sun, The free energy of mechanically unstable phases, Nature communications 6 (2015) 7559. doi:10.1038/ncomms8559.

[15] A. van de Walle, S. Kadkhodaei, R. Sun, Q.-J. Hong, Epicycle method for elasticity limit calculations, Phys. Rev. B 95 (2017) 144113. doi:10.1103/PhysRevB.95.144113.

[16] A. van de Walle, Multicomponent multisublattice alloys, nonconfigurational entropy and other additions to the alloy theoretic automated toolkit, Calphad 33 (2) (2009) 266–278, tools for Computational Thermodynamics. doi:10.1016/j.calphad.2008.12.005.

[17] A. van de Walle, M. D. Asta, Self-driven lattice-model monte carlo simulations of alloy thermodynamic properties and phase diagrams, Model. Simul. Mater. Sc. 10 (2002) 521. doi:10.1088/0965-0393/10/5/304.

[18] A. van de Walle, G. Ceder, Automating first-principles phase diagram calculations, Journal of Phase Equilibria 23 (2002) 012104. doi:10.1361/105497102770331596.

[19] H. Chen, A. van de Walle, Implementing user extensions to the alloy theoretical automated toolkit via the "plug-in" functionality (2023).

[20] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, MIT, Cambridge, MA, 2006.

[21] S. Singer, J. Nelder, Nelder-Mead algorithm, Scholarpedia 4 (7) (2009) 2928, revision #91557. doi:10.4249/scholarpedia.2928.

[22] V. Ozoliņš, C. Wolverton, A. Zunger, Effects of anharmonic strain on the phase stability of epitaxial films and superlattices: Applications to noble metals, Phys. Rev. B 57 (1998) 4816.

[23] A. van de Walle, D. Ellis, First-principles thermodynamics of coherent interfaces in samari um-doped ceria nanoscale superlattices, Phys. Rev. Lett. 98 (2007) 266101. doi:10.1103/PhysRevLett.98.266101.

[24] G. Kresse, J. Hafner, Ab initio molecular dynamics for liquid metals, Phys. Rev. B 47 (1993) 558–561. doi:10.1103/PhysRevB.47.558.

[25] G. Kresse, J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Computational Materials Science 6 (1) (1996) 15–50. doi:10.1016/0927-0256(96)00008-0.

[26] G. Kresse, J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B 54 (1996) 11169–11186. doi:10.1103/PhysRevB.54.11169.

[27] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B 59 (1999) 1758–1775. doi:10.1103/PhysRevB.59.1758.

[28] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B 59 (1999) 1758–1775. doi:10.1103/PhysRevB.59.1758.

[29] P. E. Blöchl, Projector augmented-wave method, Phys. Rev. B 50 (1994) 17953–17979. doi:10.1103/PhysRevB.50.17953.

[30] J. P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (1996) 3865–3868. doi:10.1103/PhysRevLett.77.3865.

[31] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, Dan andŠkinner, G. Ceder, K. A. Persson, The Materials Project: A materials genome approach to accelerating materials innovation, APL Materials 1 (2013) 011002.

[32] C. Oses, M. Esters, D. Hicks, S. Divilov, H. Eckert, R. Friedrich, M. J. Mehl, A. Smolyanyuk, X. Campilongo, A. van de Walle, J. Schroers, A. G. Kusne, I. Takeuchi, E. Zurek, M. B. Nardelli, M. Fornari, Y. Lederer, O. Levy, C. Toher, S. Curtarolo, aflow++: a c++ framework for autonomous materials design, Comput. Mat. Sci. 217 (2022) 111889.

[33] L. J. Swartzendruber, The fe-ir (iron-iridium) system, Bulletin of Alloy Phase Diagrams 5 (1) (1984) 266–278. doi:10.1007/BF02868724.

[34] C. Nataraj, E. J. L. Borda, A. van de Walle, A. Samanta, A systematic analysis of phase stability in refractory high entropy alloys utilizing linear and non-linear cluster expansion models, Acta Materialia 220 (2021) 117269. doi:10.1016/j.actamat.2021.117269.

[35] Y. Wu, F. Zhang, X. Yuan, H. Huang, X. Wen, Y. Wang, M. Zhang, H. Wu, X. Liu, H. Wang, S. Jiang, Z. Lu, Short-range ordering and its effects on mechanical properties of high-entropy alloys, Journal of Materials Science I& Technology 62 (2021) 214–220. doi:10.1016/j.jmst.2020.06.018.

[36] F. Ducastelle, Order and Phase Stability in Alloys, Elsevier Science, New York, 1991.

[37] S. Samanta, A. van de Walle, A cluster variation method based software toolkit to introduce short range order correction to calphad free energies (2023).

[38] A. Fernandez-Caballero, J. S. Wrobel, P. M. Mummery, D. Nguyen-Manh, Short-range order in high entropy alloys:theoretical formulation and application to mo-nb-ta-v-w system, Journal of Phase Equilibria and Diffusion (2017). doi:10.48550/ARXIV.1705.01844.

[39] C.-H. Hung, T. Markham, The moore-penrose inverse of a partitioned matrix m=(adbc), Linear Algebra and its Applications 11 (1) (1975) 73–86. doi:10.1016/0024-3795(75)90118-4.