Robust cluster expansion of multicomponent systems using structured sparsity

Zhidong Leong[*] and Teck Leong Tan[†]

Institute of High Performance Computing, Agency for Science,
Technology and Research, Singapore 138632, Singapore

(Dated: October 21, 2019)

Identifying a suitable set of descriptors for modeling physical systems often utilizes either deep physical insights or statistical methods such as compressed sensing. In statistical learning, a class of methods known as structured sparsity regularization seeks to combine both physics- and statistics-based approaches. Used in bioinformatics to identify genes for the diagnosis of diseases, *group lasso* is a well-known example. Here in physics, we present group lasso as an efficient method for obtaining robust cluster expansions (CE) of multicomponent systems, a popular computational technique for modeling such systems and studying their thermodynamic properties. Via convex optimization, group lasso selects the most predictive set of atomic clusters as descriptors in accordance with the physical insight that if a cluster is selected, so should its subclusters. These selection rules avoid spuriously large fitting parameters by redistributing them among lower order terms, resulting in more physical, accurate, and robust CEs. We showcase these features of group lasso using the CE of bcc ternary alloy Mo-V-Nb. These results are timely given the growing interests in applying CE to increasingly complex systems, which demand a more reliable machine learning methodology to handle the larger parameter space.

I. INTRODUCTION

Model building in physics requires both physical insights and statistics. In the cluster expansion (CE) of multicomponent systems[1], physical insights prescribe that the energies of atomic configurations obey a generalized Ising-like Hamiltonian. The energy $E(\sigma)$ of an atomic structure σ can be expanded in terms of atomic clusters α, where the cluster correlation functions $\Phi_\alpha(\sigma)$ serve as the basis set and the effective cluster interactions (ECIs) V_α as the coefficients:

$$E(\sigma) = \sum_\alpha \Phi_\alpha(\sigma) V_\alpha. \tag{1}$$

Statistically optimal values of the ECIs could be obtained via fitting to Eq. 1 the energies of a training set of structures, usually calculated from first principles. When appropriately truncated, the CE is an accurate model for efficiently predicting the energies[2–6] or associated properties[7–12] of different atomic configurations.

However, selecting the appropriate set of atomic clusters as descriptors is challenging: selections based on physical intuition are not robust, while those based on statistics are not physical. Initially, CE was largely applied to binary alloys[2,13–26]. Thereafter, it has been applied to more complex systems, including ternary to quinary alloys[5,6,12,27–29], semiconductors[7,30], battery materials[31,32], clathrates[33,34], magnetic alloys[35–37], and nanoscale alloys[3,4,11,38–42]. In complex systems, the reduced symmetry increases the number of symmetrically distinct clusters, exacerbating the cluster selection problem. With growing enthusiasm in applying CE to higher component systems, such as high-entropy alloys[12], it is timely to introduce an improved machine-learning procedure for creating reliable CEs with physically meaningful and robust ECIs.

Currently, there are two prevalent approaches for cluster selection. The first emphasizes using physical insights, such as via specific priors in the Bayesian framework[43] or via selection rules to incorporate smaller clusters before larger ones[44,45]. The second approach espouses using sparsity-driven regularization such as compressed sensing[6,33,34,46–49]. Fundamentally, CE is a standard linear regression problem $y = X\beta$—the response y_i is the first-principles energy of the ith structure in the training set $\{\sigma\}$, the coefficient β_j is the ECI of the jth cluster, and the component x_{ij} of the design matrix X is the correlation function $\Phi_j(\sigma_i)$ of structure i with respect to cluster j. Typically, the optimal $\hat{\beta}$ is given by the regularized least-squares solution

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|_2^2 + g(\beta), \tag{2}$$

where the ℓ_p-norm is defined by $\|z\|_p = \left(\sum_i |z_i|^p\right)^{1/p}$. The penalty function $g(\beta)$ constrains β to reduce overfitting and is key to high prediction accuracy for structures outside the training set. In compressed sensing[50,51], the least absolute shrinkage and selection operator (lasso) $g(\beta) \propto \|\beta\|_1$ selects atomic clusters by favoring parsimonious models[48,49]; such models are more interpretable and simpler for quick computation, for example, in Monte-Carlo simulations.

In this paper, we present *group lasso* regularization[52] as an efficient method for obtaining reliable CEs of multicomponent systems. As an example of structured sparsity in machine learning, group lasso combines sparsity-driven regularization with physical insights to select atomic clusters as descriptors. We show that even with the large parameter space of ternary alloys and beyond, the resulting truncated CE remains sparse and robust with interpretable ECIs. With a specially constructed convex penalty $g(\beta)$, group lasso imposes the physical insight that a cluster is selected only after all its subclusters. These selection rules avoid spuriously large fitting
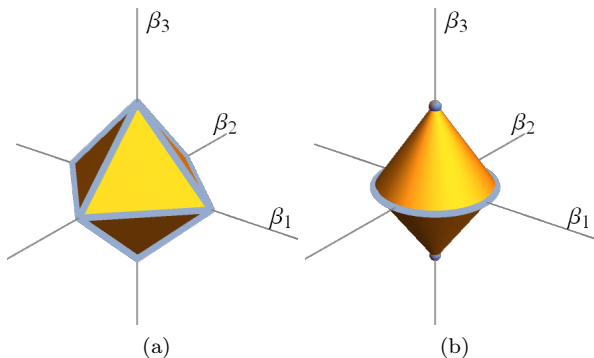
Figure 1: The constraints on $\{\beta_1, \beta_2, \beta_3\}$ in (a) lasso and (b) group lasso regularizations. The corners/edges (in light blue) of these constraints correspond to sparse solutions. In (b), coefficients β_1 and β_2 are grouped, while β_3 remains a singleton. This grouping favors solutions with β_1, β_2 either both zero or both nonzero.

parameters by redistributing them among lower order terms, resulting in more physical, accurate, and robust CEs. We will demonstrate these features of group lasso via the CE of ternary bcc alloy Mo-V-Nb.

II. METHODS

A. Group lasso

Group lasso is an extension of the well-known lasso regularization[53,54]. Using the nonanalyticity of the penalty functions, both methods favor sparse solutions to the linear regression problem $y = X\beta$. For example, the lasso penalty is $g(\beta) = \lambda \|\beta\|_1$ with hyperparameter λ, which has been studied in the context of compressed sensing CE[48,49]. In this case, the sparsity of the regularized solution from Eq. 2 can be understood in the dual picture

$$\hat{\beta} = \operatorname*{argmin}_{\beta} \|y - X\beta\|_2^2, \text{ with } \|\beta\|_1 < \tau, \quad (3)$$

where τ is inversely related to λ. Fig. 1a illustrates the constraint $\|\beta\|_1 < \tau$ for $\beta \in \mathbb{R}^3$. This constraint shrinks the least-squares solution to one that tends to lie on the corners/edges of the constraint highlighted in Fig. 1a. The resulting regularized solution is therefore sparse with some $\hat{\beta}_i$ vanishing.

In conventional lasso, the sparse solution is determined from a statistical fit, with little room for incorporating pertinent physical insights. In contrast, group lasso seeks a more physically meaningful solution by ensuring that physically-related coefficients are either all zero or all nonzero together as a group. For example, when applied to gene expression data for the diagnosis of diseases in bioinformatics, group lasso ensures that genes with coordinated functions are either all excluded or all included

in the model[55]. For CE, we will use group lasso to impose physical cluster selection rules.

In group lasso, the coefficients β are partitioned into J groups $\theta_1, \ldots, \theta_J$, where $\theta_j \in \mathbb{R}^{p_j}$ is a group of p_j coefficients. Let Z_j be the matrix formed by the columns of X corresponding to the group θ_j. Then, the regularized solution is

$$\hat{\beta} = \operatorname*{argmin}_{\beta} \frac{1}{2} \left\| y - \sum_{j=1}^{J} Z_j \theta_j \right\|_2^2 + \lambda \sum_{j=1}^{J} \sqrt{p_j} \|\theta_j\|_2, \quad (4)$$

with hyperparameter λ. Notice that unlike in the least-squares term, the ℓ_2-norm in the penalty is not squared and is therefore nonanalytic. It is this nonanalyticity that imposes sparsity.

In the dual picture, the unregularized least-squares solution is now constrained by

$$\sum_{j=1}^{J} \sqrt{p_j} \|\theta_j\|_2 < \tau. \quad (5)$$

Fig. 1b illustrates this group-lasso constraint for the case with three coefficients and the groups $\theta_1 = (\beta_1, \beta_2)$ and $\theta_2 = \beta_3$. In this case, Eq. 4 simplifies to

$$\hat{\beta} = \operatorname*{argmin}_{\beta} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \left(\sqrt{2}\sqrt{\beta_1^2 + \beta_2^2} + |\beta_3| \right) (6)$$

Compared to the lasso case in Fig. 1a, sharp corners/edges (representing sparse solutions) are now at $\beta_1, \beta_2 \neq 0, \beta_3 = 0$ and $\beta_1 = \beta_2 = 0, \beta_3 \neq 0$. Group lasso thus favors solutions with β_1, β_2 either both zero or both nonzero. In general, coefficients in the same group θ_j are either all zero or all nonzero.

When each group in Eq. 4 is a singleton, that is $p_j = 1$ for all j, the regularized solution reduces to that of lasso

$$\hat{\beta}_{\text{lasso}} = \operatorname*{argmin}_{\beta} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1. \quad (7)$$

We will use this to benchmark the performance of group lasso. Since the penalty terms for both lasso and group lasso are convex, the regularized solutions can be efficiently obtained by convex optimization. Note that the weights $\sqrt{p_j}$ in the penalty term of Eq. 4 ensure that groups of different sizes are penalized equally. Without these weights, a group with many coefficients will unfairly dominate the penalty term. We next discuss the cluster selection rules we wish to impose using group lasso.

B. Hierarchical cluster selection rules

In CE, the energy of an atomic configuration is expanded in terms of the atomic clusters and their associated ECIs. In general, since a cluster b is a higher

order correction to its subcluster $a \subset b$, the ECI $\beta_b \neq 0$ only if the subcluster ECI $\beta_a \neq 0$. I.e., a CE should include a cluster only if all its subclusters are also included. This is the hierarchical cluster selection rule we adopt here. Similar rules have been used for the CEs of binary systems[3,4,41,43–45,56,57]. Here, we extend such rules to alloy systems with more components.

Without vacancies, an m-component system requires the tracking of $m - 1$ independent atomic species. For $m \geq 3$, the key distinction from binaries is that for a given cluster, multiple decorations (of independent atomic species) need to be accounted for when considering subcluster relations. For a given independent decoration, the correlation function in Eq. 1 is defined as the number of clusters present in the atomic structure. For example, Fig. 2a shows three decorated clusters of a ternary system on a bcc lattice. The pair a, triplet b, and quadruplet c are related by $a \subset b$, $a \subset c$ and $b \not\subset c$. These relations are represented graphically in Fig. 2b, where each bubble contains a cluster (shown as a 2D schematic) with lines connecting it to its subclusters with one fewer atom. The three clusters in Fig. 2a correspond to those in the dashed box in Fig. 2b. The set of highlighted clusters (bubbles with yellow background) is an example satisfying the hierarchical cluster selection rules, while the set with red borders does not. Our work aims to use group lasso to obtain cluster sets that obey the hierarchical rules.

C. Cluster selection with group lasso

Imposing the cluster selection rules using group lasso is a subtle but important point. This is because the hierarchical rules require overlapping groups of ECIs, which are incompatible with how group lasso is formulated in Sec. II A. The solution is to use a variant of group lasso known as *overlap group lasso*[58].

To show how this variant of group lasso can impose the cluster selection rules, we consider just two clusters $c_1 \subset c_2$ and the corresponding ECIs β_1 and β_2. To have $\beta_2 \neq 0$ imply $\beta_1 \neq 0$ (as per the selection rules), we first write $\beta_1 = \theta_{11} + \theta_{21}$ and $\beta_2 = \theta_{22}$. Then, grouping together θ_{21} and θ_{22}, we apply group lasso using Eq. 4 to find the optimal θ_{11}, θ_{21}, and θ_{22}:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{2} \|y - x_1 (\theta_{11} + \theta_{21}) - x_2 \theta_{22}\|_2^2$$
$$+ \lambda \left(|\theta_{11}| + \sqrt{2} \sqrt{\theta_{21}^2 + \theta_{22}^2} \right). \quad (8)$$

As discussed, the form of group lasso's penalty ensures that θ_{21} and θ_{22} are either both zero or both nonzero. Consequently, $\beta_2 \neq 0$ implies that $\beta_1 \neq 0$ (almost surely), but we can still have $\beta_2 = 0$ with $\beta_1 \neq 0$. This is precisely the selection rule corresponding to the subcluster relation $c_1 \subset c_2$.

For a general set of p clusters $\{c_1, \ldots, c_p\}$, group lasso can similarly impose the selection rules. First, we write
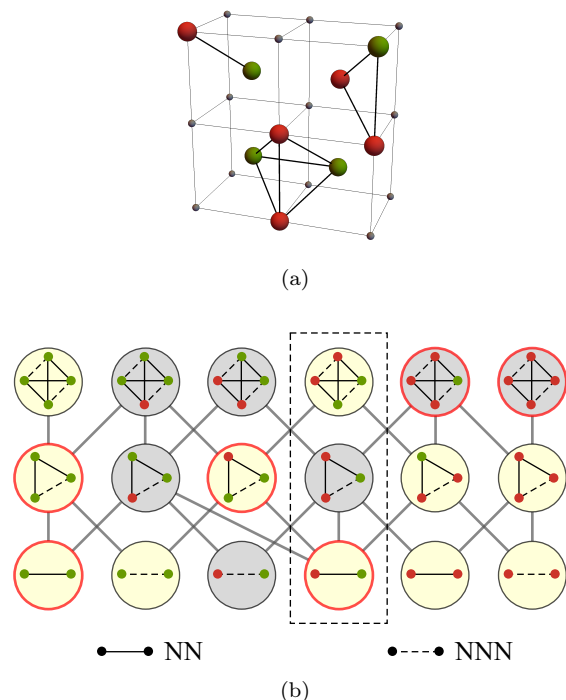

(a)

●——● NN   ●---● NNN

(b)

Figure 2: Atomic clusters of a bcc ternary system, with atomic species distinguished by colors. (a) Examples of the smallest pair, triplet, and quadruplet. (b) A graphical representation of the subcluster relations. Each bubble contains a cluster shown as a 2D schematic, with lines connecting it to all its subclusters with one fewer atom. The clusters in the dashed box correspond to those in (a). The highlighted vertices form a set of clusters obeying the hierarchical selection rules, while those with a red border do not.

the ECIs $\beta = (\beta_1, \ldots, \beta_p)^T$ as a sum of p groups of coefficients: $\beta = \sum_{j=1}^{p} \nu_j$ where $\nu_j \in \mathbb{R}^p$ is a vector constrained to be zero everywhere except in positions corresponding to c_j and its subclusters. That is, we fix $v_{j,k} = 0$ for all k such that $c_k \not\subseteq c_j$. Then, the group lasso solution for the unconstrained components is analogous to Eq. 8:

$$\hat{\nu} = \underset{\nu}{\operatorname{argmin}} \frac{1}{2} \left\| y - X \sum_{j=1}^{p} \nu_j \right\|_2^2 + \lambda \sum_{j=1}^{p} \sqrt{p_j} \|\nu_j\|_2, \quad (9)$$

where p_j is the number of subclusters of c_j (including c_j itself). That is, p_j is the number of unconstrained components in ν_j.

Here, we verify that Eq. 9 works as intended: the selection of a cluster c_j should imply the selection of its subcluster $c_l \subset c_j$. Given $\beta_j \neq 0$, we have $\nu_{k,j} \neq 0$ for some k such that $c_j \subseteq c_k$. Then, for a subcluster $c_l \subset c_j$ (and hence $c_l \subset c_k$), the $\|v_k\|_2$ term in the penalty ensures that $\nu_{k,l} \neq 0$. Consequently, $\beta_l \neq 0$, as required.
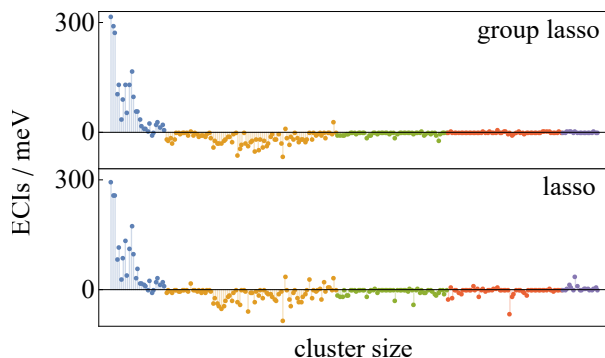
Figure 3: The values of 239 ECIs based on 800 training structures. Pairs, triplets, quadruplets, 5-bodies, and 6-bodies are colored blue, orange, green, red, and purple, respectively. The ECIs from group lasso are well-behaved—larger clusters generally have smaller ECIs—while for lasso, several isolated spikes corresponding to large ECIs are observed among the higher-order clusters (quadruplets and beyond).
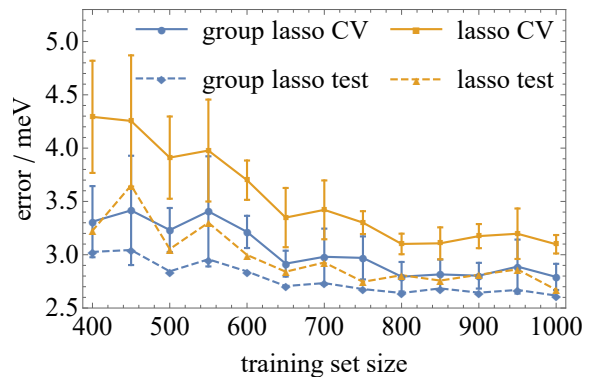

Figure 4: Five-fold cross-validation (CV) scores and test errors for group lasso and lasso versus training set size. The errors of group lasso are consistently lower than lasso's. The error bars for the CV scores correspond to one standard deviation among the five folds.

III. RESULTS

We showcase the features of group lasso via the CE of bcc ternary alloy Mo-V-Nb, whose constituent elements are well-known refractory metals. Previously, CE has been used to study the ground states of binary alloys V-Nb[59] and Mo-Nb[60,61]. Here, we benchmark the performance of group lasso (Eq. 9) against lasso (Eq. 7). The former method imposes the hierarchical cluster selection rules, while the latter performs regularization based just on statistics. The value of the hyperparameter λ in each method is fixed by cross-validation (CV). Our training structures have small unit cells with up to six atoms. We use 239 clusters consisting of pairs, triplets, ..., and six-body clusters, with 1654 cluster selection rules. As we will see, group lasso tends to produce CEs that are more physical, accurate, and robust than those from lasso. The appendix contains further technical details about our implementation.

Physicalness: Fig. 3 shows the values of the ECIs based on 800 training structures. The group lasso ECIs, by construction, obey all the cluster selection rules, and they satisfy the physical intuition that ECIs generally weaken with increasing cluster size. This behavior suggests that the CE is converging, given our initial pool of clusters. In contrast, the lasso ECIs obey only $\sim 87\%$ of the rules, and numerous large clusters have abnormally large ECIs. Therefore, via the selection rules, group lasso redistributes these spurious spikes in lasso among lower-order terms. While this redistribution decreases sparsity (205 nonzero ECIs for group lasso vs 180 for lasso), CEs from group lasso have more physical trends in the ECIs than from lasso. These general behaviors are observed regardless of the training set choices.

Accuracy: In addition to the training structures, we also have 500 test structures with large 16-atom unit cells not used for training. For both lasso and group lasso, Fig. 4 shows the CV scores and test errors decreasing as the number of training structures increases, signifying the convergence of the CEs. For either method, the CV scores and test errors are comparable. These observations imply that the lasso class of methods are able to distill the essential physics from training with just small structures, reliably predicting the energies of larger structures not in the training set. This is advantageous for ternary alloys and beyond, because of the huge number of large structures in these systems. For all training set sizes, group lasso is consistently more accurate than lasso (smaller CV scores and test errors). Therefore, the incorporation of physical hierarchy improves not only the physical interpretability of the ECIs but also the predictive capability of the CE. Group lasso reduces overfitting by redistributing the contributions from unphysical spikes in lasso's ECIs among numerous smaller clusters that are more important.

Robustness: The ECIs of a robust CE should converge towards the true physical values when more training structures are used. As such, a lack of robustness is signified by ECIs wildly fluctuating with respect to the size of the training set. The degree of fluctuations can be concisely illustrated using the root-mean-square (rms) of the ECIs in each cluster category (pairs, triplets, ..., and six-bodies). Fig. 5 shows that the five rms ECIs from group lasso are largely stable with respect to the number of training structures. However, the ECIs from lasso tend to vary wildly for the higher order clusters. This distinction shows that group lasso produces CEs that are more robust; the ECIs are more physically interpretable for group lasso (especially for higher order clusters), as they tend to fluctuate less with different training sets.
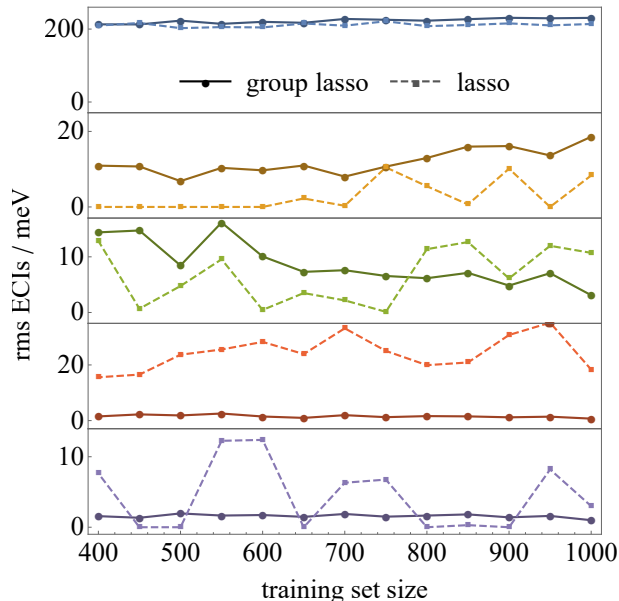
Figure 5: The root-mean-square ECIs with respect to the number of training structures for different category of clusters, namely, from top to bottom, pairs, triplets, quadruplets, 5-bodies, and 6-bodies. For higher order clusters, the ECIs of group lasso tend to fluctuate much less than those of lasso.

IV. DISCUSSIONS AND CONCLUSION

As mentioned in Sec. II B, similar hierarchical cluster selection rules have been used for CE[3,4,41,43–45,56,57]. Compared to previous works, the combination of these rules with sparsity-driven regularization in our work leads to more robust ECIs. This is because regularization shrinks the values of the selected ECIs to avoid spuriously large terms. Furthermore, since previous methods involve evaluating different combinations of clusters separately to find the optimal one, these methods become less computationally feasible for ternary systems and beyond, where many more combinations of clusters need to be explored. This is so unless the search space is shrunk by imposing additional selection criteria, for example, if an n-body cluster is included, then all n-body clusters of smaller spatial extent are also included[44,45]. We do not impose these additional criteria in our work; they might be too restrictive for ternaries and beyond because, for example, the inclusion of A-B pairs up to a certain spatial extent should not impact the spatial extent of pairs for other decorations (i.e., B-C, A-C).

In Ref.[56], the authors studied the invariance of CE under linear transformations of the site occupation variables. The authors showed that invariance is preserved only when the hierarchical cluster selection rules are obeyed. We emphasize that our group lasso implementation obeys the hierarchical rules, whereas standard lasso does not. Hence, our work presents a way for preserving the invariance of CE.

In conclusion, we presented *group lasso*[52] as an efficient method for producing reliable CEs of multicomponent alloys, resulting in accurate and robust surrogate models for predicting thermodynamic properties. A type of structured sparsity regularization, group lasso combines statistical learning with physical insights to select atomic clusters as descriptors for the CE model. Via convex optimization, group lasso imposes the cluster selection rules that a cluster is selected only after all its subclusters. These rules avoid spuriously large fitting parameters by redistributing them among numerous lower order terms, resulting in more physical, accurate, and robust CEs. These results are timely given the growing interests in applying CE to increasingly complex systems, where the larger parameter space demands a more reliable machine learning methodology to construct robust models. Furthermore, this work should inspire applying structured sparsity in modeling other physical systems.

APPENDIX

In this appendix, we present the technical details about our implementation of cluster expansion (CE) and group lasso.

1. First-principles calculations

The energies of the training and test structures are calculated based on density functional theory (DFT) with the Vienna Ab initio Simulation Package (VASP)[62,63]. We use the Perdew, Burke, and Ernzerhof exchange correlation based on the generalized gradient approximation[64,65]. The PAW potentials are used with the outer p semi-core states included in the valence states[66,67]. Plane-wave cutoffs are set to 520 eV and all atomic coordinates (including lattice vectors) were fully relaxed until the calculated Hellmann-Feynman force on each atom was less than 0.015 eV/Å. Calculations are non spin-polarized as Mo, Nb, and V are not known to be strongly magnetic. The k-point mesh is generated using a Gamma grid and density of 200 Å^{-3}.

2. Normalization choice for cluster correlations

The general expression for CE given by

$$E\left(\sigma\right) = \sum_{\alpha} \Phi_{\alpha}\left(\sigma\right) V_{\alpha}. \qquad (A.10)$$

can be rewritten to account for the degeneracy of the clusters in a specific lattice[68]. For any rescaling factor $\eta_{\alpha} > 0$, Eq. A.10 is invariant under the transformation $\Phi_{\alpha}\left(\sigma\right) \rightarrow \Phi_{\alpha}\left(\sigma\right)\eta_{\alpha}$ and $V_{\alpha} \rightarrow V_{\alpha}/\eta_{\alpha}$. The choice of η_{α} depends on whether degeneracy factors are subsumed into $\Phi_{\alpha}\left(\sigma\right)$ or V_{α}. Here, we choose η_{α} such that $\Phi_{\alpha} = N_{\alpha}/\widetilde{N}_{\alpha}$, where $N_{\alpha}\left(\widetilde{N}_{\alpha}\right)$ is the number of clusters in the structure that are symmetrically equivalent to cluster α, (without) taking into account the decorations. This normalization gives $0 \leq \Phi_{\alpha} \leq 1$ for all α's, which is convenient because the convergence of V_{α} with respect to cluster size would directly reflect the convergence of the CE.

In practice, we use occupation variables ξ to describe the atomic species at each lattice site of a structure: $\xi_{A}\left(\sigma_{j}\right)$ equals 1 (0) if site j in structure σ is (not) occupied by species $A \in \{Mo, V, Nb\}$. Note that this is distinct from the orthogonal basis in an alternate CE formalism[1]. Then, the correlation function of structure σ with respect to cluster α is computed using

$$\Phi_{\alpha}\left(\sigma\right) = \frac{1}{\widetilde{N}_{\alpha}} \sum_{c} \prod_{j \in c} \xi_{c_{j}}\left(\sigma_{j}\right), \qquad (A.11)$$

where the sum is over all clusters c symmetrically equivalent to α. The product is over all sites j in the cluster, with c_{j} giving the atomic species at site j. We reiterate that for ternary alloys and beyond, decorations need to be taken into account when considering symmetrically equivalent clusters.

3. Formation energy

In general, either the configuration energy $E\left(\sigma\right)$ or the formation energy $E_{F}\left(\sigma\right)$ could be used to train the CE. In this work, we use the latter, which is defined as

$$E_{F}\left(\sigma\right) = E\left(\sigma\right) - \sum_{A} \rho_{A}\left(\sigma\right) E\left(\sigma_{A}^{pure}\right), \quad (A.12)$$

where $\rho_{A}\left(\sigma\right)$ is the concentration of species A in the structure σ, and σ_{A}^{pure} is the pure system of species A. With the CE of $E\left(\sigma\right)$ from Eq. A.10, the formation energy can be expanded in terms of the ECIs:

$$E_{F}\left(\sigma\right) = \sum_{\alpha} \left[\Phi_{\alpha}\left(\sigma\right) - \sum_{A} \rho_{A}\left(\sigma\right) \Phi_{\alpha}\left(\sigma_{A}^{pure}\right)\right] V_{\alpha}. (A.13)$$
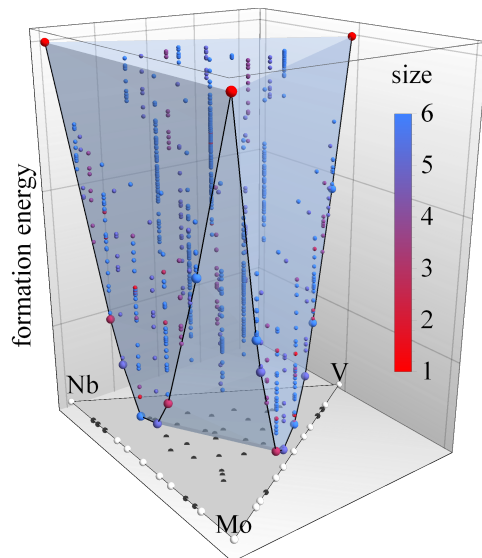

Figure 6: The DFT formation energies E_{F} of 1081 derivative structures with up to 6-atom unit cell in a bcc lattice, with respect to compositions. Structures with $E_{F} > 0$ are not shown. Redder (bluer) points are structures with smaller (larger) unit cells. The blue translucent surface is the ground state hull, with ground state structures represented by larger points. The ternary plot shows the compositions of the structures, with ground state structures highlighted in white.

Because the expression in the square bracket vanishes exactly for the empty cluster and singlets, the formation energy is expandable in terms of just pairs and larger clusters[57]. This form of the formation energy also naturally gives $E_{F} = 0$ for pure systems. Then, writing Eq. A.13 as the linear regression problem $y = X\beta$, we standardize the columns of X to have unit ℓ_{2}-norm before applying group lasso (or lasso), as per common practice[54]. That is, denoting the ith column of X by x_{i}, we apply the invariant rescaling $x_{i} \rightarrow x_{i}/\|x_{i}\|_{2}$ and $\beta_{i} \rightarrow \beta_{i}\|x_{i}\|_{2}$ such that $\|x_{i}\|_{2} = 1$ for all i's.

4. Generation of training and test structures

Ideally, the structures in a training set should be sufficiently varied to capture all important physics of the system. To cover a wide range of the configurational space, training structures can be selected either randomly[48,49] or systematically to maximize the covariance matrix of the correlation functions[69].

In practice, computational constraints limit the number of DFT calculations and favor training structures with smaller unit cells. This limitation is especially severe for ternary alloys and beyond, because the configurational space grows combinatorially with the number of atomic species. Therefore, we select our training structures from a pool of 1081 derivative structures, system-
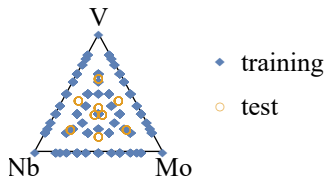
Figure 7: A ternary plot showing the compositions of the 1081 training structures and 500 test structures.

atically generated up to 6-atom unit cell[70,71]. Fig. 6 shows the DFT formation energies and compositions of these structures. Notably, lower energy structures tend to have smaller unit cells. Following the smallest-first algorithm[7], structures with smaller unit cells are chosen first. We exclude the three pure systems because their formation energies given by Eq. A.13 are identically zero.

To verify that such training sets suffice for ternary systems, we test the CE trained using small structures against a test set (holdout set) of larger structures. The test set consists of 500 randomly selected 16-atom derivative structures; this set is not used to train our CE model, but it serves to determine the testing/prediction error. The ternary plot in Fig. 7 shows the compositions of these test structures compared to those of the training set.

5. Initial set of clusters

In our CE model, we treat V and Nb as the independent species, while Mo is treated as dependent. As such, only clusters formed by Mo and V atoms are required. In the bcc lattice, we consider up to the 9th-nearest-neighbor (9NN) pairs, triplets with a 5NN cutoff, and four-body to six-body clusters with a 3NN cutoff. These correspond to an initial pool of 239 symmetrically distinct clusters, consisting of 27 pairs, 84 triplets, 54 four-body clusters, 56 five-body clusters, and 18 six-body clusters. Among these clusters are 1654 subcluster relations, which group lasso uses to derive the final truncated CE based on the cluster selection rules.

6. Tuning of hyperparameter λ

Using the DFT formation energies of the training structures, we use group lasso to select a properly trun-

cated CE set from the initial 239 distinct clusters. The group lasso minimization problem is efficiently solved using a block coordinate descent algorithm[54], which reduces the multidimensional minimization problem to a sequence of root-finding problems in 1D. Overfitting (underfitting) happens when the hyperparameter λ is too small (large). The optimal λ is selected based on a five-fold cross validation (CV) with the one-standard error rule[54], as illustrated in Fig. 8 (top). I.e., the optimal
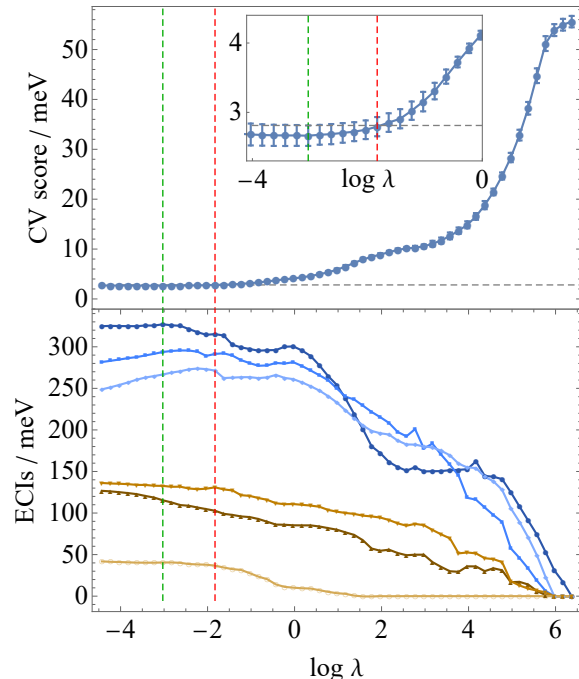

Figure 8: Top: The five-fold cross validation (CV) score of group lasso with respect to the hyperparameter λ, using 800 training structures. The green vertical line denotes the minimum CV score. The red vertical line is one-standard error away from the minimum and gives the optimal λ. Inset: a closeup of the same plot. Bottom: The six ECIs of the nearest-neighbor (blue) and next-nearest-neighbor (yellow) pairs with respect to λ.

λ corresponds to the most regularized model with CV score within one standard error of the minimal CV score. The bottom plot of Fig. 8 shows coefficient shrinkage and cluster selection in group lasso. As λ decreases, the model becomes less regularized and the ECIs generally increase; the solution is also less sparse as more ECIs become nonzero.

[*] leong_zhidong@ihpc.a-star.edu.sg
[†] Corresponding author: tantl@ihpc.a-star.edu.sg
[1] J. Sanchez, F. Ducastelle, and D. Gratias, Physica A: Statistical Mechanics and its Applications **128**, 334 (1984).

[2] V. Blum and A. Zunger, Physical Review B **69** (2004), 10.1103/PhysRevB.69.020103.
[3] T. L. Tan, L.-L. Wang, D. D. Johnson, and K. Bai, Nano Letters **12**, 4875 (2012).

[4] M.-F. Ng and T. L. Tan, Nano Letters **13**, 4951 (2013).

[5] J. S. Wróbel, D. Nguyen-Manh, M. Y. Lavrentiev, M. Muzyk, and S. L. Dudarev, Physical Review B **91** (2015), 10.1103/PhysRevB.91.024108.

[6] S. B. Maisel, M. Höfler, and S. Müller, Physical Review B **94** (2016), 10.1103/PhysRevB.94.014116.

[7] L. Ferreira, S.-H. Wei, and A. Zunger, The International Journal of Supercomputing Applications **5**, 34 (1991).

[8] A. Van der Ven, G. Ceder, M. Asta, and P. D. Tepesch, Physical Review B **64**, 184307 (2001).

[9] M. K. Y. Chan, J. Reed, D. Donadio, T. Mueller, Y. S. Meng, G. Galli, and G. Ceder, Physical Review B **81**, 174303 (2010).

[10] S. B. Maisel, M. Höfler, and S. Müller, Nature **491**, 740 (2012).

[11] L.-L. Wang, T. L. Tan, and D. D. Johnson, Physical Review B **86** (2012), 10.1103/PhysRevB.86.035438.

[12] A. Fernández-Caballero, J. S. Wróbel, P. M. Mummery, and D. Nguyen-Manh, Journal of Phase Equilibria and Diffusion **38**, 391 (2017).

[13] J. W. D. Connolly and A. R. Williams, Physical Review B **27**, 5169 (1983).

[14] Z. W. Lu, S.-H. Wei, and A. Zunger, Physical Review Letters **66**, 1753 (1991).

[15] C. Wolverton, G. Ceder, D. de Fontaine, and H. Dreysse, Physical Review B **45**, 13105 (1992).

[16] A. Zunger, in *Statics and Dynamics of Alloy Phase Transformations*, Vol. 319, edited by P. E. A. Turchi and A. Gonis (Springer US, Boston, MA, 1994) pp. 361–419.

[17] G. D. Garbulsky and G. Ceder, Physical Review B **49**, 6327 (1994).

[18] D. D. Fontaine, in *Solid State Physics*, Vol. 47 (Elsevier, 1994) pp. 33–176.

[19] Z. W. Lu, B. M. Klein, and A. Zunger, Superlattices and Microstructures **18**, 161 (1995).

[20] G. D. Garbulsky and G. Ceder, Physical Review B **51**, 67 (1995).

[21] C. Wolverton and A. Zunger, Physical Review B **52**, 8813 (1995).

[22] V. Ozoliņš, C. Wolverton, and A. Zunger, Physical Review B **57**, 6427 (1998).

[23] A. F. Kohan, P. D. Tepesch, G. Ceder, and C. Wolverton, Computational Materials Science **9**, 389 (1998).

[24] S. Müller and A. Zunger, Physical Review Letters **87**, 165502 (2001).

[25] A. Zunger, L. G. Wang, G. L. W. Hart, and M. Sanati, Modelling and Simulation in Materials Science and Engineering **10**, 685 (2002).

[26] A. van de Walle and M. Asta, Metallurgical and Materials Transactions A **33**, 735 (2002).

[27] H. Ji and Y. Jung, The Journal of Chemical Physics **146**, 064103 (2017).

[28] R. Feng, P. K. Liaw, M. C. Gao, and M. Widom, npj Computational Materials **3** (2017), 10.1038/s41524-017-0049-4.

[29] M. C. Nguyen, L. Zhou, W. Tang, M. J. Kramer, I. E. Anderson, C.-Z. Wang, and K.-M. Ho, Physical Review Applied **8** (2017), 10.1103/PhysRevApplied.8.054016.

[30] B. P. Burton, S. Demers, and A. van de Walle, Journal of Applied Physics **110**, 023507 (2011).

[31] A. Van der Ven and G. Ceder, Electrochemistry Communications **6**, 1045 (2004).

[32] K. Persson, Y. Hinuma, Y. S. Meng, A. Van der Ven, and G. Ceder, Physical Review B **82**, 125416 (2010).

[33] M. Ångqvist, D. O. Lindroth, and P. Erhart, Chemistry of Materials **28**, 6877 (2016).

[34] M. Ångqvist and P. Erhart, Chemistry of Materials **29**, 7554 (2017).

[35] R. Drautz and M. Fähnle, Physical Review B **69**, 104404 (2004).

[36] R. Drautz and M. Fähnle, Physical Review B **72**, 212405 (2005).

[37] M. Y. Lavrentiev, D. Nguyen-Manh, and S. L. Dudarev, Physical Review B **81**, 184202 (2010).

[38] J. Kang, S. Tongay, J. Li, and J. Wu, Journal of Applied Physics **113**, 143703 (2013).

[39] L. Cao and T. Mueller, The Journal of Physical Chemistry C **119**, 17735 (2015).

[40] L. Cao and T. Mueller, Nano Letters **16**, 7748 (2016).

[41] T. L. Tan, H. M. Jin, M. B. Sullivan, B. Anasori, and Y. Gogotsi, ACS Nano **11**, 4407 (2017).

[42] L. Cao, C. Li, and T. Mueller, Journal of Chemical Information and Modeling **58**, 2401 (2018).

[43] T. Mueller and G. Ceder, Physical Review B **80** (2009), 10.1103/PhysRevB.80.024103.

[44] A. van de Walle and G. Ceder, Journal of Phase Equilibria **23**, 348 (2002).

[45] N. A. Zarkevich and D. D. Johnson, Physical Review Letters **92** (2004), 10.1103/PhysRevLett.92.255702.

[46] G. L. W. Hart, V. Blum, M. J. Walorski, and A. Zunger, Nature Materials **4**, 391 (2005).

[47] R. Drautz and A. Díaz-Ortiz, Physical Review B **73**, 224207 (2006).

[48] L. J. Nelson, G. L. W. Hart, F. Zhou, and V. Ozoliņš, Physical Review B **87** (2013), 10.1103/PhysRevB.87.035125.

[49] L. J. Nelson, V. Ozoliņš, C. S. Reese, F. Zhou, and G. L. W. Hart, Physical Review B **88** (2013), 10.1103/PhysRevB.88.155105.

[50] E. J. Candes, J. Romberg, and T. Tao, IEEE Transactions on Information Theory **52**, 489 (2006).

[51] E. J. Candes and M. B. Wakin, IEEE Signal Processing Magazine **25**, 21 (2008).

[52] M. Yuan and Y. Lin, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**, 49 (2006).

[53] R. Tibshirani, Journal of the Royal Statistical Society. Series B (Methodological) **58**, 267 (1996).

[54] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability (CRC Press, 2015).

[55] S. Ma, X. Song, and J. Huang, BMC Bioinformatics **8**, 60 (2007).

[56] M. H. F. Sluiter and Y. Kawazoe, Physical Review B **71**, 212201 (2005).

[57] N. A. Zarkevich, T. L. Tan, L.-L. Wang, and D. D. Johnson, Physical Review B **77** (2008), 10.1103/PhysRevB.77.144208.

[58] L. Jacob, G. Obozinski, and J.-P. Vert, in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09 (ACM, New York, NY, USA, 2009) pp. 433–440.

[59] C. Ravi, B. K. Panigrahi, M. C. Valsakumar, and A. van de Walle, Physical Review B **85**, 054202 (2012).

[60] V. Blum and A. Zunger, Physical Review B **72**, 020104 (2005).

[61] W. P. Huhn and M. Widom, JOM **65**, 1772 (2013).

[62] G. Kresse and J. Furthmüller, Computational Materials Science **6**, 15 (1996).

[63] G. Kresse and J. Furthmüller, Physical Review B **54**, 11169 (1996).

[64] J. P. Perdew, K. Burke, and M. Ernzerhof, Physical Review Letters **77**, 3865 (1996).

[65] J. P. Perdew, K. Burke, and M. Ernzerhof, Physical Review Letters **78**, 1396 (1997).

[66] P. E. Blöchl, Physical Review B **50**, 17953 (1994).

[67] G. Kresse and D. Joubert, Physical Review B **59**, 1758 (1999).

[68] N. A. Zarkevich, T. L. Tan, and D. D. Johnson, Physical Review B **75**, 104203 (2007).

[69] A. Seko, Y. Koyama, and I. Tanaka, Physical Review B **80** (2009), 10.1103/PhysRevB.80.165122.

[70] G. L. W. Hart and R. W. Forcade, Physical Review B **77** (2008), 10.1103/PhysRevB.77.224115.

[71] G. L. Hart, L. J. Nelson, and R. W. Forcade, Computational Materials Science **59**, 101 (2012).