Tutorial

# Construction and Sampling of Alloy Cluster Expansions—A Tutorial

Pernilla Ekborg-Tanner, Petter Rosander, Erik Fransson, and Paul Erhart*
*Department of Physics, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden*

(Received 23 May 2024; revised 12 September 2024; published 17 October 2024)

Crystalline alloys and related mixed systems make up a large family of materials with high tunability that have been proposed as the solution to a large number of energy-related material design problems. Because of the presence of chemical order and disorder in these systems, neither experimental efforts nor *ab initio* computational methods alone are sufficient to span the inherently large configuration space. Therefore, fast and accurate models are necessary. To this end, cluster expansions have been widely and successfully used for the past few decades. Cluster expansions are generalized Ising models designed to predict the energy of any atomic configuration of a system after training on a small subset of the available configurations. Constructing and sampling a cluster expansion consists of multiple steps that have to be performed with care. In this tutorial, we provide a comprehensive guide to this process, highlighting important considerations and potential pitfalls. The tutorial consists of three parts, starting with cluster expansion construction for a relatively simple system, continuing with strategies for more challenging systems such as surfaces, and closing with examples of Monte Carlo sampling of cluster expansions to study order-disorder transitions and phase diagrams.

DOI: 10.1103/PRXEnergy.3.042001

## CONTENTS



*Contact author: erhart@chalmers.se



## I. INTRODUCTION

Materials engineering proposes a solution to many of the energy-related problems our society is faced with. Tremendous research efforts are currently being made towards optimizing green technologies such as solar cells, batteries, catalysts, and fuel cells. The performance of these technologies is dictated by the properties of the materials involved, meaning that optimization is achieved by improving the design of materials, which in turn requires tunability of the materials. A common strategy to expand the tunability is to consider multicomponent systems, since the larger chemical space enables tailoring via chemical composition and ordering. Recent examples can be found in the fields of photovoltaics [1–5], batteries [6–8], catalysis [9–11], fuel cells [12,13], nanophotonics

and nanoplasmonics [14,15], construction and manufacturing [16–19], and two-dimensional materials such as MXenes [20–23].

The increased tunability comes at the cost of added complexity in the material design process, and computational methods are often needed to efficiently span the composition space. While computational efforts are ideally based on *ab initio* methods such as density-functional theory (DFT) calculations, such methods are typically computationally too expensive for sampling the relevant composition space. To exemplify this consider that binary system consisting of $N$ atoms corresponds to approximately $2^N$ indistinguishable atomic configurations. This means that already at system sizes of 100 atoms one would need to examine roughly $10^{30}$ atomic configurations. For this reason, more effective models are necessary. For crystalline materials, cluster expansions (CEs) are ideal candidate models as indicated by the large number of successful applications, including phase-diagram prediction for metals [24–28] and semiconductors [29–33], ordering phenomena [34–40], and the properties of surfaces [41–51] and nanoparticles [52–59].

CEs are generalized Ising models that can, in principle, predict the energy (or any other function of the configuration) of any atomic configuration after training on only a small subset of the available atomic configurations (Fig. 1). They are typically sampled in Monte Carlo (MC) simulations to extract thermodynamic information about the system and study thermodynamic observables such as chemical ordering, free energy, and heat capacity. While we focus on energy prediction in this tutorial, we note that CEs have also been used to model, for example, activation barriers [60], vibrational properties [61,62], chemical expansion [38], and transport properties [37,63], which can serve as suitable starting points for interested readers. There are various software packages that implement the CE framework and thermodynamic sampling via MC sampling, including, for example, ATAT [64], UNCLE [65], CLEASE [66], CASM [67], ICET [68], and SMOL [69].

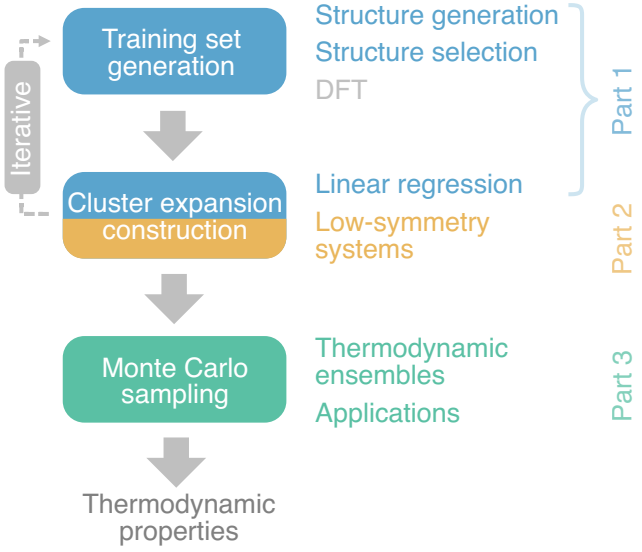


FIG. 1. Constructing and sampling a CE. Overview of the typical procedure for constructing and sampling a CE and connection to the three main parts of this tutorial.

In this tutorial, we present an overview of practical aspects to be considered when one is constructing and sampling CEs. It is accompanied by a set of Jupyter notebooks available online [70,71] that use the ICET package [68]. The tutorial is organized as follows. First, the CE formalism is briefly outlined in Sec. II, followed by practical considerations for constructing a CE in Sec. III. Then we review the CE construction process for a relatively simple system ($Mo_{1-x}V_xC_{1-y}\square_y$) in Sec. IV, with emphasis on training set generation and the training procedure. Next we discuss CE construction for a low-symmetry system—namely, a $Au_xPd_{1-x}$ surface—in Sec. V, which allows us to discuss the use of local symmetries, Bayesian priors, and constraints for improving CE performance. Lastly, in Sec. VI we introduce sampling of CEs via MC simulations to obtain various thermodynamic observables for $Au_3Pd$ as well as the $Ag_xPd_{1-x}$ alloy system.

## II. CLUSTER EXPANSION FORMALISM

The theory behind CEs has been discussed at length elsewhere [30,72–77]. In the present context, we present a short overview and focus on practical considerations.

A CE predicts the value of some observable $E$ as a function of the atomic configuration, represented by the occupation vector $\boldsymbol{\sigma}$, for a crystalline material, on the basis of all involved atomic clusters $\boldsymbol{\alpha}$. A cluster is a set of $k$ atomic sites on the crystalline lattice, where $k = 0, 1, 2, \ldots$ is the order of the cluster. The observable can be expressed as

$$E(\boldsymbol{\sigma}) = \sum_{\alpha} J_{\alpha} \Pi_{\alpha}(\boldsymbol{\sigma}), \tag{1}$$

where $J_{\alpha}$ is the contribution of cluster $\boldsymbol{\alpha}$, called the effective cluster interaction (ECI), and $\Pi_{\alpha}$ are orthogonal basis functions spanning the space of atomic configurations [72,76]. All symmetrically equivalent clusters have the same ECI and can be grouped into what we refer to as an "orbit" (originating from group theory and used in Ref. [68]), which is also known as a "representative cluster" in the CE literature. Use of this fact simplifies Eq. (1) to

$$E(\boldsymbol{\sigma}) = \sum_{\beta} m_{\beta} J_{\beta} \langle \Pi_{\alpha}(\boldsymbol{\sigma}) \rangle_{\beta}, \tag{2}$$

where $m_{\beta}$ is the multiplicity of orbit $\boldsymbol{\beta}$ and the $\langle \ldots \rangle_{\beta}$ bracket indicates that the basis function $\Pi_{\alpha}$ is averaged

over all clusters $\boldsymbol{\alpha}$ in the orbit $\boldsymbol{\beta}$. The sum in Eq. (2) runs over all orbits up to some cutoff criterion, typically defined by a cutoff radius $r$ for each cluster order $k$.

Training a CE involves finding the optimal ECIs on the basis of a set of atomic structures for which the value of the observable $E$ is known, called the "training set." In most cases, the observable of interest $E$ is the energy or a variant thereof—say, the mixing energy or a migration barrier—with reference data obtained from from DFT calculations. Equation (2) can be cast as a linear problem,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}, \tag{3}$$

where $\boldsymbol{y}$ is a vector comprising the target values of the observable $E$ for the training set, $\boldsymbol{w}$ is an unknown vector containing the ECIs, and $\boldsymbol{X}$ is the so-called design matrix, where each row corresponds to the values of the basis functions $\Pi_{\boldsymbol{\alpha}}$ for a structure in the training set. Note that in this form the multiplicities are absorbed into either $\boldsymbol{w}$ or $\boldsymbol{X}$. The number of ECIs, i.e., the size of $\boldsymbol{w}$ and accordingly the number of columns in $\boldsymbol{X}$, are determined by the cutoff radii used to select the allowed orbits of different order.

The optimal ECIs $\boldsymbol{w}_{\text{opt}}$ can be found with the use of linear regression techniques (see Sec. III). Once the optimal ECIs are found, the CE can be used to predict $E$ for any atomic configuration representable by a supercell of any size and concentration.

The CE formalism applies to ideal lattices. In most cases, however, relaxed structures are more thermodynamically relevant than structures with atoms residing precisely on the lattice sites. CEs are therefore often used to model the energy of relaxed structures mapped to the closest corresponding occupation on the ideal lattice. This means that the ECIs effectively include variations in the interactions due to volume changes and atomic relaxations.

The validity of a CE used over a concentration range has been debated in the literature [73,78,79]. In practice, for rare cases, it can become difficult to model the full concentration range with a single CE, especially if the system undergoes sharp transitions at some specific concentrations. This can (but need not) occur, for example, in materials with electronic band gaps if the charge state of a species changes with composition (e.g., $Mn^{2+} \rightarrow Mn^{3+} \rightarrow Mn^{4+}$ [80]) or if the species that are being mixed are aliovalent (e.g., $Si^{4+}$ and $Al^{3+}$ [81]). In these situations, one needs to consider the overall charge balance as the total energy becomes, in principle, dependent on the Fermi level (i.e., the electron chemical potential). This issue has been circumvented, for example, in the case of zeolites by the construction of CEs exclusively on charge-balanced configurations and with the use of MC trial moves that maintain charge balance [81], while a more general approach for ionic materials is described in Ref. [80]. There are also cases where CEs may not be sufficient to capture all relevant interactions in the system, such as long-ranged strain interactions, which requires extension of the CE formalism to reciprocal space [25,82,83].

## III. CONSTRUCTING CLUSTER EXPANSIONS

A CE model should be both accurate and transferable. In practice this means that we aim to find an optimal set of ECIs, $\boldsymbol{w}$ in Eq. (2). This entails choosing a regression method along with related hyperparameters and cutoff parameters (i.e., the size of the model) as well as composing a set of training structures. In the following, we first briefly review regression methods (Sec. III A), followed by a short discussion of how to assess model performance (Sec. III B) as well as the role of cutoffs and hyperparameters (Sec. III C). The impact of the composition of the training structure set is discussed through practical examples in Sec. IV B.

### A. Regression methods

There are many approaches to solving the linear regression problem in Eq. (3) [68,84–87]. Here we provide a short introduction to several common methods.

The solution of the linear problem, Eq. (3), with some typical regularization terms can be written as

$$\boldsymbol{w}_{\text{opt}} = \underset{\boldsymbol{w}}{\text{argmin}} \left\{ \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \boldsymbol{w}^T\boldsymbol{\Lambda}\boldsymbol{w} + \|\boldsymbol{M}\boldsymbol{w}\|_1 \right\}, \tag{4}$$

where $\boldsymbol{w}_{\text{opt}}$ is the solution vector, $\boldsymbol{\Lambda}$ is the $\ell_2$ regularization matrix, and $\boldsymbol{M}$ is the $\ell_1$ regularization matrix. Setting $\boldsymbol{\Lambda} = \boldsymbol{M} = \boldsymbol{0}$ yields the ordinary least squares (OLS) solution. OLS is, however, prone to overfitting, meaning that the model captures noise in the training data and therefore performs poorly on new unseen data. For $\boldsymbol{M} = 0$ and $\boldsymbol{\Lambda} = \alpha\boldsymbol{1}$, Eq. (4) reduces to ridge regression with regularization parameter $\alpha$, and with $\boldsymbol{\Lambda} = 0$ and $\boldsymbol{M} = \alpha\boldsymbol{1}$, one obtains the expression used in the least absolute shrinkage and selection operator (LASSO).

The regularized regression methods generally assume that the design matrix, $\boldsymbol{X}$, is standardized. Therefore, it is common practice to rescale the columns of $\boldsymbol{X}$ to have unit variance before solving the problem, and afterwards apply the inverse procedure to obtain the unscaled parameters. This rescaling needs to be taken into consideration if one is manually choosing values for the regularization matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{M}$ (for instance, for Bayesian CEs, see Sec. V B). Additionally, one commonly trains CEs for the mixing (or formation) energy, rather than the total energy, to avoid very large target values, $\boldsymbol{y}$.

Feature selection techniques can be used to reduce the number of nonzero ECIs in the solution vector. This can lead to more transferable models and faster predictions (e.g., in MC simulations). Recursive feature elimination (RFE) is an iterative algorithm where in each iteration one solves the linear problem (typically with OLS) and the

least important (smallest) ECIs are pruned (set to zero). This is done iteratively until a desired number of nonzero ECIs is reached. Automatic relevance detection regression (ARDR) is based on the Bayesian ridge regression technique where the regularization matrix is diagonal with elements $\Lambda_{ii} = \lambda_i$. The individual regularization strengths $\lambda_i$ for each parameter are updated throughout the optimization, and parameters are pruned (set to zero) if $\lambda_i$ increases above a threshold ($\lambda$ threshold) [88]. These linear regression methods can readily be used for CEs with the use of ICET via SCIKIT-LEARN [89] or even more directly via the TRAINSTATION interface to the former [86].

### B. Model performance and learning curves

In general, one wants to construct models that require as few training data and as few nonzero ECIs as possible. Fewer training data means fewer (usually computationally demanding) reference calculations, while fewer nonzero ECIs translates to reduced model complexity, a feature that is often associated with better transferability, i.e., such models perform more reliably on unseen data. These aspects need to be taken into account when one is constructing CEs.

Techniques such as ridge regression, RFE, ARDR, or LASSO involve hyperparameters, e.g., the regularization parameter $\alpha$ in ridge regression, the number of features in RFE, or the $\lambda$ threshold in the case of ARDR. In addition, one must make select the cutoffs that determine the range of the summation in Eq. (2). These parameters directly affect model performance in terms of accuracy, transferability, and data efficiency, i.e., the amount of reference data needed to obtain a well-converged model.

In practice one usually determines optimal parameters through so-called *learning curves*, which show a suitable performance indicator (see below) as a function of, for example, hyperparameters, cutoff values, or training set size. We exemplify this approach throughout this tutorial (see, e.g., Figs. 2, 8, and 10).

The most widely used measure for model performance is the root-mean-square error (RMSE) score calculated over a *validation* set, since the RMSE over the *training* set [i.e., the first term on the right-hand side of Eq. (4)] is a poor estimate of how a model will perform on unseen data points. The validation RMSE is commonly calculated via cross-validation.

As we already noted above, simpler models tend to exhibit better transferability. This principle is approximately represented through information criteria such as the Akaike information criterion and the Bayesian information criterion [90–93], which weigh validation RMSE versus model size. These measures can hence be useful when one is selecting between models that have similar RMSE values. We note, however, that in our experience these information criteria are often not sufficiently conclusive as a general measure when one is constructing CE models.

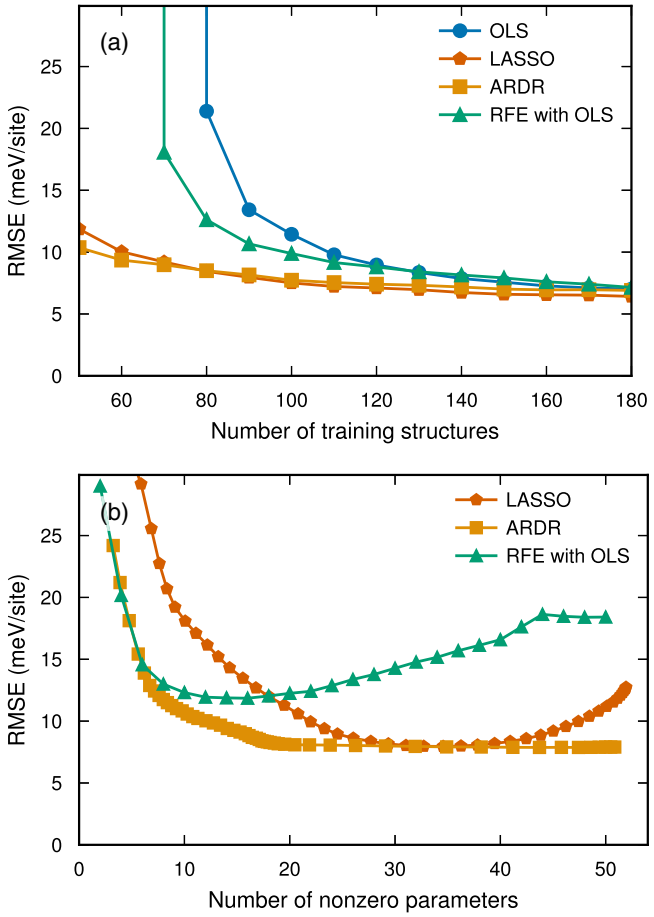


FIG. 2. RMSE obtained with different regression methods in the $Mo_{1-x}V_xC_{1-y}\square_y$ system. (a) Validation RMSE based on the random training set obtained with different regression methods as a function of the number of structures included in the training set. (b) Validation RMSE when 80 training structures are used as a function of the number of nonzero parameters obtained after regression and feature selection. Here the number of nonzero parameters obtained is controlled by variation of the hyperparameter controlling the sparsity of the solution continuously. For ARDR, this is the $\lambda$-threshold parameter, for LASSO it is $\alpha$ [see Eq. (4)], and for RFE it is the number of features.

Lastly, we emphasize that none of these performance measures are perfect when it comes to assessing the general performance of a CE and that they will always reflect the choice of training data to some degree. We therefore recommend that one should always study the property or properties of interest (e.g., phase diagram, heat capacity, or order parameters) throughout the CE construction process to ensure that convergence is achieved.

### C. Cutoff selection versus regularization

When it comes to selecting cutoffs, the conventional approach, especially when one is using OLS [and thus

no regularization, i.e., $\mathbf{\Lambda} = \boldsymbol{M} = \mathbf{0}$ in Eq. (4)], is to start with a set of small cutoffs and low orders and iteratively increase the size (and complexity) of the model. Too-small cutoffs (and thus a small number of ECIs) lead to underfitting, whereas too-large cutoffs (and thus a large number of ECIs) lead to overfitting. A good starting point for the length of the cutoff is on the order of the lattice parameter, and cutoffs larger than three lattice parameters are very rarely needed. The interaction strength decreases with the order of the cluster, meaning that inclusion of terms up to third or fourth order is sufficient in most cases.

When using regularization and feature selection approaches, one can, in principle, choose a large initial number of orbits using both larger cutoffs and higher-order orbits, which is then reduced by the regression method of choice. ARDR usually performs very well in such situations; see, e.g., Refs. [33,86] and examples below. This means that with regularization, the importance of cutoff selection decreases, as long as the cutoffs are large enough. In practice, however, one can run into problems if cutoffs are selected too large. In our experience, the best performance is therefore often found when ARDR is combined with cutoff selection.

Lastly, we note that one can also use a Bayesian method, where physical intuition is encoded in a regularization matrix $\mathbf{\Lambda}$ (see, e.g., Ref. [94] and Sec. V B), to, for instance, enforce higher importance of low-order and/or short-ranged orbits. This approach could be useful for certain complex systems, but requires significantly more work to set up than, for instance, ARDR.

## IV. PART 1: A FIRST EXAMPLE

### Key takeaways

(1) Simple linear regression techniques such as OLS often lead to overfitting and large validation errors. Use regularization and/or feature selection to improve model performance.
(2) Similarly, naive structure selection schemes such as randomization can lead to poorly performing models, while more advanced schemes generally yield better models and require fewer training structures.
(3) We recommend generating structures via either condition number minimization or uncertainty maximization in conjunction with ARDR.

In the first part of this tutorial, we illustrate the construction of a CE with emphasis on the choice of regression techniques (Sec. IV A) and approaches for structure selection (Sec. IV B). To this end, we consider a simple carbide $Mo_{1-x}V_xC_{1-y}\square_y$ with two sublattices on a rocksalt lattice. On the metal sublattice, we consider mixing between Mo and V, and on the carbon sublattice, we consider C atoms and vacancies ($\square$), with a maximum of 30% carbon vacancies. We use cutoffs of 9 and 5 Å for two-body and three-body clusters, respectively. This yields a total of 52 ECIs, including two singlets, 25 pairs, 24 triplets, and a constant, sometimes referred to as the "zerolet." A detailed analysis of the cutoff selection can be found in the notebooks accompanying this tutorial [70,71]. We use reference data from DFT calculations available online [71]. The computational details concerning these calculations can be found in Appendix A.

### A. Comparison of regression methods

We now explore some of the above-mentioned regression methods for solving Eq. (3) so as to illuminate some of their differences. For simplicity, we consider a set of 200 "random" training structures, where each structure has a randomized supercell size (up to a maximum of 50 atoms), random Mo/V and C/vacancy concentrations, and random occupation of the lattice. This set of structures is referred to as the "random set" of structures from here on. The resulting validation error as a function of the number of training structures is shown in Fig. 2(a). First, we note that all four regression methods yield similar validation errors when a large training set is used. For OLS the validation error increases rapidly when the number of training structures decreases, which is also to lesser extent the case for RFE, due to overfitting. Both ARDR and LASSO perform significantly better than OLS and RFE when a small number of training structures is used.

ARDR, RFE, and LASSO all have one hyperparameter that controls the sparsity (number of nonzero parameters in the solution), which needs to be chosen. This is typically done by scanning of the hyperparameter value and selection of the value yielding the smallest validation error. Here we use this procedure each time a model is trained.

Figure 2(b) shows an example for the variation of the validation error with the resulting number of features during a hyperparameter scan. LASSO has a minimum at about 35 nonzero parameters, whereas ARDR achieves the same validation error with only 20 parameters. The tendency of LASSO to overselect is known [95], and we have observed this behavior previously for both CEs [68] and in other applications, such as force constant expansions [86]. RFE has a minimum at about 15 nonzero parameters but with slightly higher validation error compared with the other two regression methods.

The trends described above are, in our experience, general, and we have found ARDR to be the best performing approach in most cases. Therefore, we recommend using ARDR as a starting point for CE construction.

### B. Training set generation methods

For simplicity in Sec. IV A we used randomized structures. We now discuss more informed approaches to generating or selecting training structures.

Recall that we are trying to solve the linear problem (3), with the design matrix $\boldsymbol{X}$ operating on the solution vector $\boldsymbol{w}$, i.e., the ECIs. Commonly we would like to find solutions in large cluster spaces (large cutoffs and more bodies), expecting though that only relatively few ECIs are significant due to the near-sightedness of physical interactions. This means $\boldsymbol{w}$ should be sparse. Under these circumstances it can be shown that optimal design matrices $\boldsymbol{X}$ should be nearly orthonormal, a feature that is characterized by the restricted isometry property [96]. Heuristically, this can be thought of as setting up the training set with as much variability in the cluster vectors as possible.

At first it might be tempting to use randomized structures, since two random configurations will very rarely be identical and the cluster vector is a function of the configuration. If one considers how the cluster vector is composed [Fig. 3(a)], it becomes, however, quickly evident that this is a poor choice. Let us imagine a simple binary system. After the zerolet, the first cluster vector item is the singlet, which reflects the overall composition. Then we have the pairs that reflect the proportion of pairs between alike (*A*-*A*, *B*-*B*) and unalike (*A*-*B*) pairs. For a large random structure, there is no tendency for ordering, which means that the proportion of alike and unalike pairs will be determined by the overall composition and take on the same value for all orders [Fig. 3(b)]. Smaller random structures allow some variation around the large-structure limit, but these variations are small compared with what can be achieved with other generation methods. A similar argument can be made for higher-order clusters. This example demonstrates that while random structures are rarely identical, they have similar cluster vectors. For contrast, we can compare the cluster vectors of large and small random structures as well as enumerated structures [Fig. 3(b)], which shows that by going beyond random structures, we can achieve a much larger spread of cluster vectors.

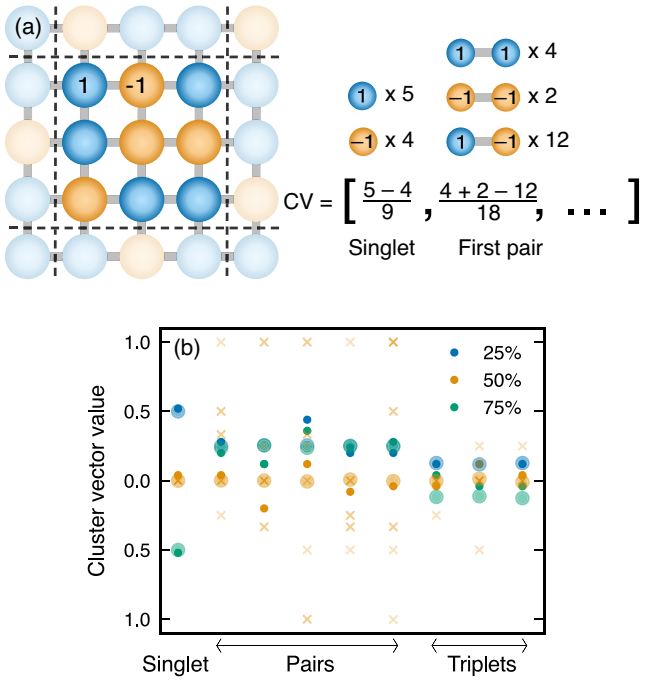


FIG. 3. Cluster vectors for random structures. (a) Example of how the first elements of a cluster vector are calculated for a 2D structure. (b) Cluster vectors for randomized atomic configurations of a 2D square binary system. The concentration is indicated by the color. The large and small circles represents large ($1\times10^4$ atoms) and small (25 atoms) random structures, respectively. The crosses represents ten randomly selected structures from a pool of enumerated structures with up to eight atoms (and concentration 50%) for comparison.

There are many methods for generating a set of training structures, each having their pros and cons. Strictly speaking, one should discriminate between structure *generation* and structure *selection*:

In this tutorial, we consider four ways of *generating* structures: enumeration of all possible structures up to some maximum structure size, randomized structures (in terms of structure size, composition, and/or configuration), MC sampling (see Sec. VI) of existing CEs, and selecting one or a set of target cluster vectors and finding the closest-matching structure (structures).

We also consider four ways of *selecting* structures. The most straightforward choice is to select all the generated structures, which leaves little control of the quality of the training set. More advanced selection methods can be applied with the aim of optimizing some aspect of the training set, such as minimizing the condition number of the design matrix, selecting structures with the largest uncertainty, or trying to achieve a set of structures with orthogonal cluster vectors.

These generation and selection schemes can be combined and modified to create a sheer endless number of different training sets. Here we consider five approaches that we refer to as "uncertainty maximization," "condition number minimization," "structure orthogonalization," "structure enumeration," and the "random set" discussed in Sec. IV A (Fig. 4). In the following, we present and discuss these approaches.

*Uncertainty maximization* is a form of *active learning*, which is a common training set generation approach for model construction in general, including CE construction [97–99]. Here the model is iteratively trained and new structures are selected at each iteration on the basis of the model uncertainty prediction for a pool of structures. The uncertainty of structures can be estimated from, for example, an ensemble of CEs from bootstrap sampling [68,99]. This means that a collection of CEs trained in identical fashion but based on different training sets is generated. The different training sets are generated by resampling of the original training set with replacement (meaning one structure can appear multiple times in a training set). Structures can, for example, be generated by MC simulations in

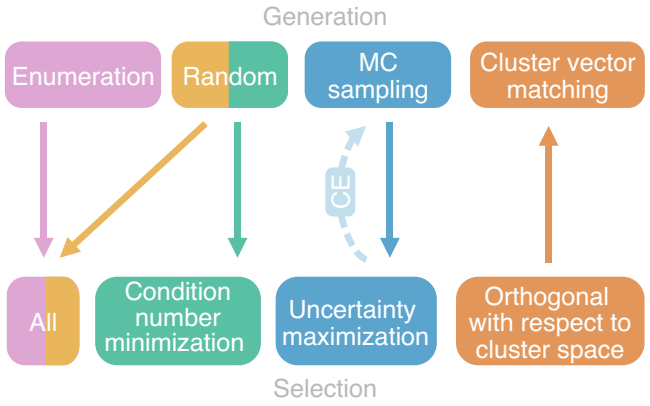


FIG. 4. Structure generation and selection schemes. The process of generating and selecting a set of structures for the training set can be designed in multiple ways. In this tutorial, we focus on the five approaches represented by the arrows in this schematic overview.

the canonical ensemble (see Sec. VI), and the ones with the largest uncertainties are selected to be included in the training set. The benefit of this structure generation approach is that the training structures added during each iteration are structures for which the current ensemble of models predicts large uncertainty, and therefore will lead to large accuracy increases in each iteration. Another benefit of this approach is that structures are generated from MC simulations with the desired conditions (concentrations, temperatures, etc.), meaning the structures selected will be thermodynamically relevant ones. On the other hand, this also limits the pool of structures to select from in terms of how they span the cluster vector space. Furthermore, this approach requires a relatively large effort due to the iteration process and the fact that one needs to define a set of structures for which the uncertainty is predicted.

"*Condition number minimization*" refers to the process of selecting structures such that the condition number of the linear problem in Eq. (3) is minimized. The condition number $c$ is defined as [100]

$$c = \frac{\max(\Gamma)}{\min(\Gamma)}, \tag{5}$$

where $\Gamma$ denotes the singular values of the design matrix. The condition number describes how well conditioned the linear problem is, with smaller values indicating better conditioning. It can be thought of as a measure of how sensitive the fitting result is with respect to changes or errors in the input data.

This approach starts by generating a large pool containing on the order of millions of randomly generated structures. An initial training set of $N$ structures is randomly drawn from the pool, where $N$ is on the order of hundreds. Next, a simulated annealing MC simulation (see Sec. VI) is conducted where structures from the large random pool are randomly swapped in and out of the training set with probability $P = \mathrm{e}^{-\Delta c/\Theta}$, where $\Theta$ is an artificial temperature and $\Delta c$ is the change in condition number when two structures are swapped. The resulting training set will be an approximate solution to the problem of selecting a subset of $N$ structures from the large pool with the lowest condition number.

*Structure orthogonalization* aims at providing a set of structures with cluster vectors orthogonal to each other [84,101]. It is related to the condition number approach described above in the sense that both methods aim to span the cluster vector space. The procedure starts from a structure with a random cluster vector. New structures are then added iteratively by finding a cluster vector that is orthogonal to the rest and identifying the closest-matching structure.

A benefit with this approach is that, in principle, one can quickly produce a training set without the need for any iterative models (as in the uncertainty maximization) or a large pool of structures (as in the condition number minimization). The latter depends, however, on the ability to find a matching structure for a target cluster vector without, for example, a pool of structures. In ICET, this is possible due to an implemented method based on simulated annealing.

There are, however, two major drawbacks to this approach. First, the entire cluster vector space is not available due to correlations between the cluster vector elements (Fig. 5). For instance, the number of possible *A*-*B* nearest-neighbor pairs is limited by the value of the singlet. We show this in Fig. 5 for the entire available space (represented by the large random pool used for condition number minimization) and the dataset produced with this approach. Clearly, both the fact that the concentration range for the carbon sublattice is restricted and the fact that the singlet and pairs are correlated significantly limit the available space. The orthogonalization procedure will more often than not ask for a structure outside this space, and the resulting structure (with the closest-matching cluster vector) will not be orthogonal to the other structures.

Second, the number of orthogonal cluster vectors is limited by the number of elements in the cluster vector (which in turn is determined by the cutoffs). This means that the training set size is limited by the cluster cutoffs when this approach is used.

*Structure enumeration* is a method that generates all symmetrically inequivalent structures that are permissible given a certain lattice, under the constraint that the number of atoms in each structure must be smaller than some given number [102]. The benefit of structure enumeration is that it requires no input other than the maximum number of atoms in a supercell and will systematically generate high-symmetry and ground-state structures. One drawback is that the maximum number of atoms needs to be quite small (typically fewer than 15 atoms) for the number of

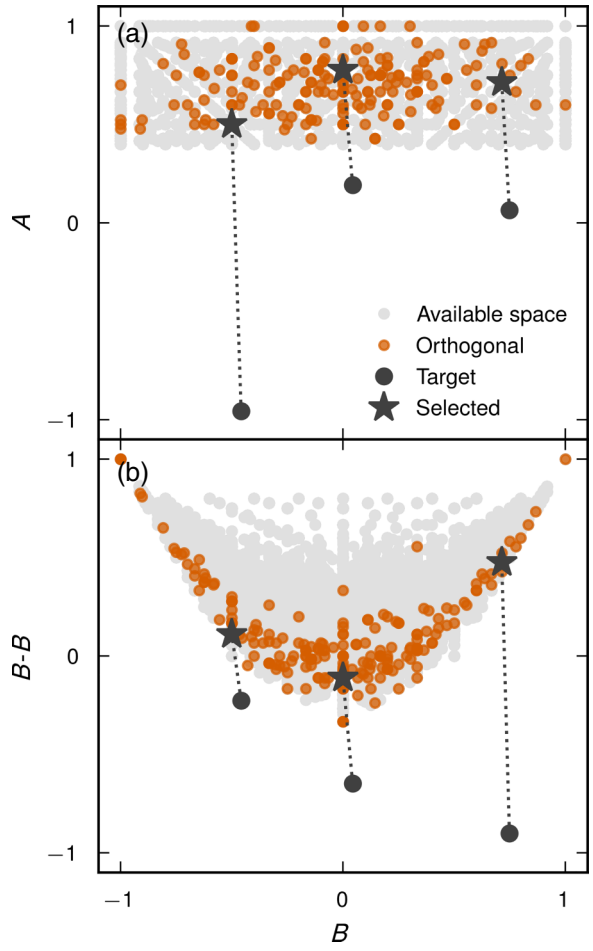


FIG. 5. Correlations between orbits. Correlations between (a) the singlets of the carbon ($A$) and metal ($B$) sublattices and (b) the metal singlet and nearest neighbor. The light-gray area indicates the available space represented by all the structures from the large random pool used for condition number minimization. The orthogonal dataset is shown in orange. The orthogonalization procedure is visualized by the initial (circle) and final (star) positions of three example structures.

structures to be computationally feasible, and this can lead to the training set not spanning long-ranged interactions. For our model system, $Mo_{1-x}V_xC_{1-y}\square_y$ (with the chosen cutoffs), enumeration up to 12 atoms leads to about 650 structures, but produces an ill-conditioned design matrix and should thus not be used for training. For this reason, the enumerated training set is used here only for testing purposes. Additionally, for systems with larger and/or complex primitive cells, it may be unfeasible to use altogether as the number of enumerated structures grows exponentially with the size of the primitive cell.

Before we compare these approaches, we note that the validation error is not a suitable measure to evaluate the quality of a training set generation scheme, because a procedure generating very similar (or identical) structures would lead to models with low validation errors

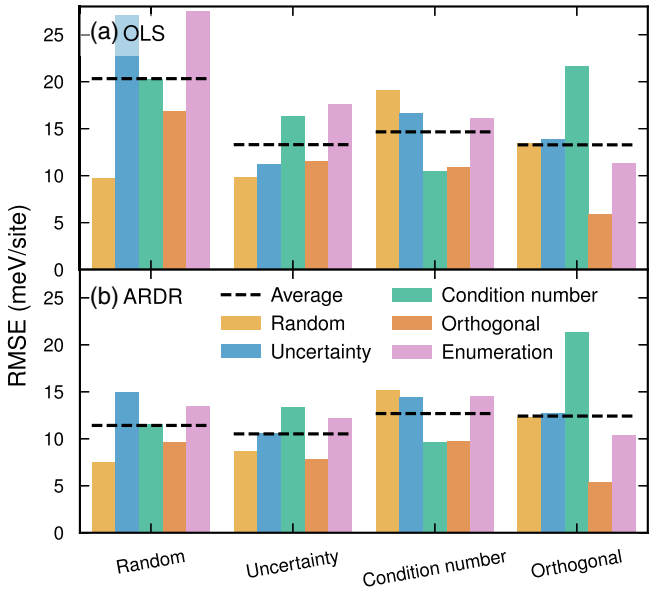


FIG. 6. Comparison of structure generation schemes for the $Mo_{1-x}V_xC_{1-y}\square_y$ system. Here CEs were constructed with the use of four different training sets (as indicated on the $x$ axis) consisting of 110 structures and trained with either (a) OLS or (b) ARDR. The RMSEs were evaluated with use of the different training sets as test sets (as indicated by the bar color), including a set of enumerated structures. The dashed line indicates the average over the five test sets.

but poor predictions on structures outside the training set. Therefore, we evaluate each CE on all the structure sets generated (Fig. 6).

In Fig. 6(a) the validation error is shown over all sets of structures for CEs trained with each training set using OLS. Here we see that the random training set leads to larger errors across all structure sets, whereas training with the three other structure sets yields lower validation errors across the board. In Fig. 6(b) the same analysis is done for training with ARDR. Here we see that the validation errors obtained with a random training set are on the same level as those obtained with the other training sets, demonstrating the efficacy of ARDR even with a poor choice of training structures. We also note that training with the structure orthogonalization structures leads to a large error over the condition number structures, yet very small validation error, indicating poor transferability.

### C. Conclusions

In the first part of this tutorial, we have demonstrated the CE construction process for a relatively simple system, focusing on linear regression and structure generation and selection. We have found that OLS with a training set consisting of randomized structures leads to large validation errors, and that this can be prevented by the use of better structure selection schemes as well as regularization and

feature selection. We note that the material considered here represents a rather simple system with high symmetry, and for more complex cases, optimization of the training set and regression method typically yields a larger benefit.

On the basis of these findings, our general recommendation is to generate structures by either uncertainty minimization (if it seems worthwhile to put in the effort) or condition number minimization (if one wants to avoid an iterative process) in conjunction with ARDR when one is solving the linear problem (3). Lastly, we again emphasize that the training set generation and selection methods can be combined in multiple ways, and the procedures outlined here do not need to be followed strictly. It might, for instance, be beneficial to start with a fast method (enumeration or orthogonalization) and extend the training set with an iterative method. In practice, it is also often beneficial to include thermodynamically relevant structures, in particular ground-state structures, in the training set.

## V. PART 2: LOW-SYMMETRY SYSTEMS

### Key takeaways

(1) Low-symmetry system generally have a large number of ECIs, which makes CE construction challenging.
(2) Local symmetries and Bayesian inference can be used to couple similar orbits.
(3) Weighted constraints can be applied to ensure certain properties are represented more accurately by the CE.

In the second part of this tutorial, we consider CE construction for a low-symmetry system, specifically a surface, which comes with new challenges. The number of orbits for a given set of cutoffs grows with the number of symmetrically inequivalent sites that make up the material. For simple bulk systems, such as the one in Sec. IV, the unit cell typically only comprises few atoms which implies a small number of inequivalent sites. For *low-symmetry* systems such as surfaces and nanoparticles, on the other hand, the unit cell is generally larger. For instance, the number of inequivalent sites for a surface slab is at least the number of layers divided by 2 (Fig. 7). Consequentially, the number of orbits increases rapidly with the number of layers, and the CE construction procedure outlined in Sec. IV can be insufficient.

In the following, we discuss three approaches to improve CE construction for low-symmetry systems. The first two approaches are based on grouping similar orbits and either explicitly merging them or coupling them in a Bayesian framework. The third approach consists of adding weighted constraints to ensure that specific properties are more accurately predicted. These approaches can be combined and are not restricted to low-symmetry systems.

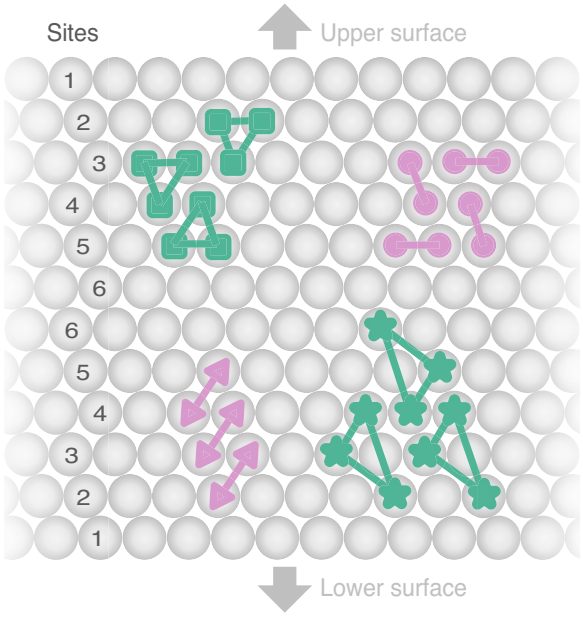


FIG. 7. Merging of orbits. Side view of a 12-layer face-centered cubic (fcc) (111) surface slab. The numbers on the left row of atoms indicate the symmetrically inequivalent sites. Examples of orbits that can be merged if only site 1 is considered a surface site are marked in pink for pairs and in green for triplets and are marked with different symbols for inequivalent orbits (note the difference in length between sets of pairs).

To showcase these approaches we construct CEs for a 12-layer $Au_xPd_{1-x}$ fcc (111) surface slab. The cutoffs used for CE construction are 6 and 3 Å for two-body and three-body clusters, respectively, resulting in 76 orbits. The training set consists of the pure Au and Pd slabs and 201 structures generated with the orthogonal cluster space approach described in Sec. IV B (with the use of slightly larger cutoffs to avoid underdetermined systems during training). Details of the DFT calculations can be found in Ref. [50].

We begin by training CEs as in Sec. IV, using plain OLS and ARDR to fit the ECIs. As expected, ARDR outperforms OLS in terms of the validation error as well as with respect to the number of structures necessary to reach convergence (Fig. 8). The number of structures needed to reach convergence is, however, relatively large for both fitting methods.

The validation error is not always a sufficient measure for CE accuracy. In the present example of a surface, we find that although the validation energy has converged, the segregation energy prediction is associated with large errors, which will ultimately result in erroneous predictions of the surface segregation. The segregation energy is the energy difference caused by the moving of, for example, a single Pd atom in an otherwise pure Au slab from the middle of the slab (i.e., the bulk) to a site in the surface region.

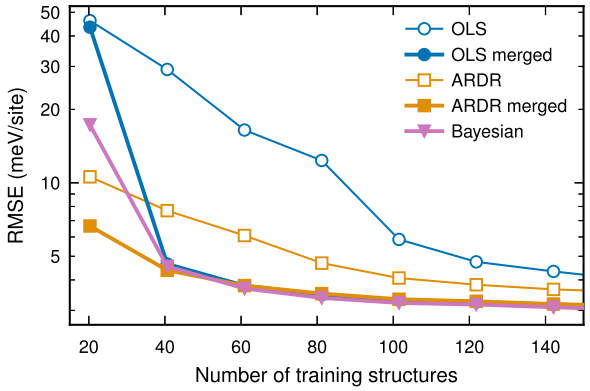


FIG. 8. Learning curves for a surface CE. The validation error as a function of the training set size for different linear regression methods, with and without merging of orbits.

(Here we construct $2\times2\times12$ slabs for this purpose.) This calculation requires the prediction of two configurations, which means it is the relative error between these two configurations that matters. In Figs. 9(a) and 9(d), we find that the segregation energies predicted by the plain OLS and ARDR CEs have large errors and systematically overestimate the energy gain of moving Pd (in Au) to the surface and the energy cost of moving Au (in Pd) to the surface. There are also unphysical oscillations of the segregation energy in the inner layers, whereas the DFT results indicate that the segregation energy varies only in the outer two to three layers. This suggest that the models predict too-large differences between the atoms in the inner layers, as discussed further in the following section.

### A. Merging of orbits

At a sufficient distance from the surface, the atomic interactions should reach the bulk limit. For example, the energetic contribution from a nearest-neighbor pair in layer 5 should be very similar to the one in layer 6. We can use this idea to define so-called *local* symmetries. Orbits that belong to the same local symmetry should have the same ECI, which means that they can be merged into a single orbit. This process is analogous to when the individual clusters are grouped into orbits in Eq. (2). Note that in this approach we no longer rely on the *global* symmetries, which can be rigorously derived from the underlying lattice. Instead the specification of the local symmetries is up to the person constructing the CE, and is based on additional *physical* knowledge of the system, such as the range of the interactions and the similarity of the local environments.

Note that it is important to be aware of the treatment of multiplicities when one is merging orbits. In many implementations of the CE method, including ICET, the default behavior is to include the multiplicities in the ECI vector $\boldsymbol{w}$. This will lead to incorrect results if any of the merged orbits have different multiplicities. Instead, the multiplicities should be included in the design matrix $\boldsymbol{X}$, which will lead to correct averaging of the multiplicities when merging is done.

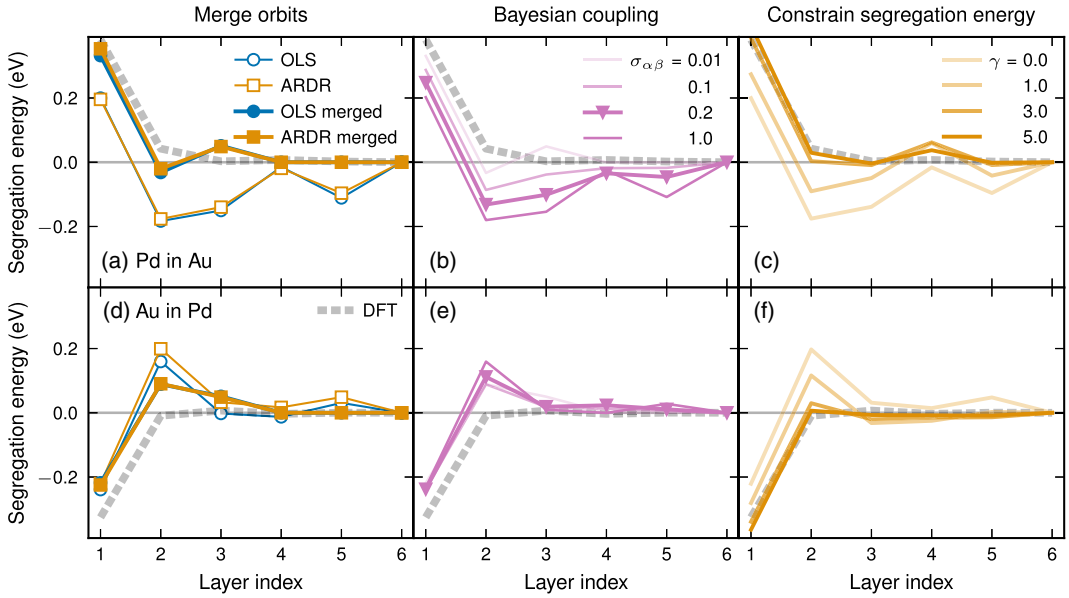


FIG. 9. Segregation energy predictions for surface CEs. Segregation energies for a single Pd atom in Au [(a)–(c)] and a single Au atom in Pd [(d)–(f)] positioned in each atomic layer calculated by DFT and predicted by different CEs. The layer index is counting from the surface, as in Fig. 7. (a),(d) CEs trained with OLS and ARDR with and without merging of orbits. (b),(e) CEs trained using Bayesian coupling of similar orbits. The CE with the optimal coupling parameter $\sigma_{\alpha\beta}$ from Fig. 10 is highlighted with triangles. (c),(f) CEs trained with (unmerged) ARDR and an added constraint enforcing correct segregation energy prediction with increasing weight $\gamma$.

In the following, we choose to consider only the outermost layer as the surface and to treat the remaining layers as the bulk. We then merge all orbits with the same order and radius that consist exclusively of bulk sites (Fig. 7), reducing the number of orbits from 76 to 20. Other options include extending the surface region to several layers, introducing a subsurface region treated differently, and treating the local symmetries differently depending on cluster order.

The present, rather aggressive, merging strategy leads to a significant reduction of the number of training structures necessary to reach a certain accuracy, while reducing the final validation error (Fig. 8). Except for the smallest training set considered, there is almost no difference between OLS and ARDR. This is because the lack of regularization in OLS often leads to overfitting, which is avoided here due to the small number of features. We also find that the segregation energy predictions are significantly better for the merged CEs [Figs. 9(a) and 9(d)]. This is a result of our restricting the differences between the ECIs in the inner layers.

### B. Bayesian coupling of orbits

The assumption that similar orbits should have similar energetic contribution to the total energy is an example of a physical insight about the system. If such insights can be formulated in the form of Bayesian priors, they can be used to construct improved CEs within a Bayesian framework [84,94,103]. Here we follow the approach first presented by Mueller and Ceder [94].

We assume Gaussian priors for the ECIs such that

$$P(\boldsymbol{w}|\boldsymbol{X}) \propto \prod_{\alpha} e^{-w_\alpha^2/2\sigma_\alpha^2} \prod_{\alpha,\beta\neq\alpha} e^{-(w_\alpha - w_\beta)^2/2\sigma_{\alpha\beta}^2}, \quad (6)$$

where $P(\boldsymbol{w}|\boldsymbol{X})$ is the posterior. Here the first product controls the magnitude of the ECIs via $\sigma_\alpha$, which is the standard deviation of the prior and should be chosen to roughly correspond to the expected value of the respective ECI. If one, for instance, wanted to achieve smaller ECIs for large clusters, one could define $\sigma_\alpha$ such that it decreases with orbit size. The second product controls the coupling between orbits via $\sigma_{\alpha\beta}$, which is the inverse coupling strength between orbits $\alpha$ and $\beta$.

Going back to the linear problem formulated in Eq. (4), the Bayesian priors are introduced via the $\ell_2$ regularization matrix $\Lambda$, which has diagonal elements $\Lambda_{\alpha\alpha} = (\sigma^2/\sigma_\alpha^2) + \sum_{\beta\neq\alpha}(\sigma^2/\sigma_{\alpha\beta}^2)$ and off-diagonal elements $\Lambda_{\alpha\beta} = \Lambda_{\beta\alpha} = -(\sigma^2/\sigma_{\alpha\beta}^2)$, where $\sigma$ reflects the typical error of the model. The maximum posterior estimate for the ECIs is then given by

$$\boldsymbol{w}_{\text{opt}} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Lambda})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \quad (7)$$

As in the merging approach, one has to be careful with the treatment of multiplicities when using Bayesian coupling of similar orbits. Coupling orbits means that we expect the values of their ECI to be similar. It is therefore crucial that the multiplicities are included in the design matrix and not the ECI vector when this approach is applied. In addition, in linear regression one typically uses standardization, i.e., rescaling of the design matrix, to improve the linear regression. If this is the case, one has to use the same scaling of the Bayesian regularization matrix.

In this tutorial, we apply a rather simple Bayesian approach using the same definition of local symmetries as in Sec. V and couple the orbits belonging to the same local symmetry using a single coupling parameter $\sigma_{\alpha\beta}$ for all coupled orbits. The ECI magnitude is controlled by a single value of $\sigma_\alpha$ for all orbits as well. We emphasize that the Bayesian priors can have a much more intricate design than this to allow greater tunability [94]. For example, one can take into account the order and size of orbits and define complex criteria for the similarity of orbits.

In Fig. 10 we show how the validation error and the training error vary with the coupling parameter $\sigma_{\alpha\beta}$ while keeping $\sigma_\alpha = 10$. For large $\sigma_{\alpha\beta}$, the unmerged CE is obtained with a high validation error but low training error (indicating overfitting). For small $\sigma_{\alpha\beta}$, the merged CE is recovered, and the training and validation errors approach each other, with a significant reduction of the validation error compared with the validation error in the large-$\sigma_{\alpha\beta}$ limit. The latter is particularly prominent for smaller training sets. In between these to extremes, at $\sigma_{\alpha\beta} \approx 0.2$, the validation error has a minimum. This indicates that the optimal compromise between the freedom of unmerged orbits and reduced feature space of merged

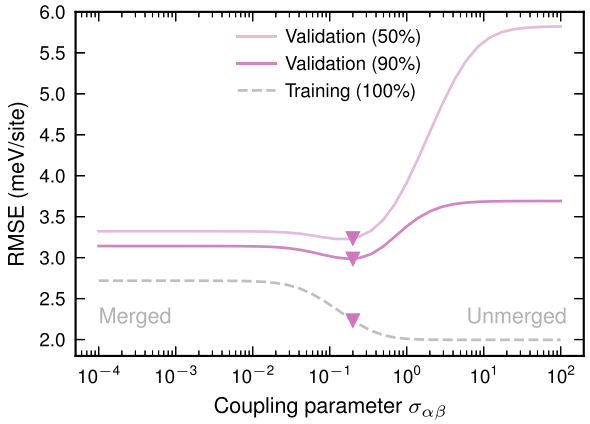


FIG. 10. Performance of Bayesian surface CEs with differing coupling strength. Validation error and training error with differing coupling parameter $\sigma_{\alpha\beta}$ for similar orbits. The validation error is shown for two different training set sizes (50% and 90%), and the training error is obtained with the full training set (100%). The optimal coupling parameter value, $\sigma_{\alpha\beta} = 0.2$, is highlighted with triangles.

orbits is found. Figures 9(b) and 9(e) show the segregation energy prediction for various values of $\sigma_{\alpha\beta}$.

The identified optimal $\sigma_{\alpha\beta} = 0.2$ provides some improvement in segregation energy prediction compared with the unmerged CEs, but decrease of $\sigma_{\alpha\beta}$ further towards the merged limit results in significantly better segregation energy prediction, in particular for the case of Pd in Au. This again shows that the validation RMSE is not always sufficient to find the most physically sound CE.

### C. Adding constraints and weights

A CE can also be manipulated by introducing constraints and/or weights to ensure that certain properties are reproduced more accurately. The property of interest depends on the purpose of the model. If, for example, the model is going to be used to study surface segregation phenomena, a natural choice is the segregation energy. Calculation of the segregation energy requires DFT calculations of surface slabs with a single Pd atom in a Au slab, respectively positioned in each atomic layer, and vice versa, resulting in a total of 12 structures. A straightforward approach to improve the segregation energy prediction is to include these structures in the training set. Additionally, one could give these structures a higher weight by simply multiplying the corresponding rows of the design matrix $\boldsymbol{X}$ and elements in the solution vector $\boldsymbol{y}$ by a suitably chosen factor. By one doing so, the prediction error will shrink for these specific structures, effectively reducing the segregation energy error.

Another option, which we demonstrate in this tutorial, is to explicitly enforce better predictions of the segregation energy as a constraint. This entails reformulating the linear problem as

$$\boldsymbol{w}_{\text{opt}} = \underset{\boldsymbol{w}}{\text{argmin}}\left\{\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \gamma\|(\boldsymbol{X}_{\text{S}} - \boldsymbol{X}_{\text{B}})\boldsymbol{w} - \boldsymbol{E}_{\text{seg}}\|_2^2\right\}, \tag{8}$$

excluding any regularization terms [see Eq. (4)]. Here $\boldsymbol{X}_{\text{S}}$ contains the cluster vectors of the structures with a single atom positioned in different atomic layers. Similarly, $\boldsymbol{X}_{\text{B}}$ contains the cluster vectors of the corresponding structures with a single atom in the bulk position (i.e., the innermost atomic layer). $\boldsymbol{E}_{\text{seg}}$ represents the corresponding segregation energies and $\gamma$ is a weight factor that dictates how strongly the constraint is enforced. In practice, this is achieved by one adding rows to the design matrix $\boldsymbol{X}$ corresponding to $\gamma(\boldsymbol{X}_{\text{S}} - \boldsymbol{X}_{\text{B}})$ and the target vector $\boldsymbol{y}$ corresponding to $\boldsymbol{E}_{\text{seg}}$.

In Figs. 9(c) and 9(f), we show the segregation energy predictions for different weight factors $\gamma$. As expected, the segregation energy predictions move closer to the DFT line with increasing weight. The improvement in segregation energy prediction comes at the cost of a slightly increased validation error, from 2.8 meV/site for $\gamma = 0$ to 3.4 meV/site for $\gamma = 5$.

The strategy of introducing constraints and/or weights is not restricted to low-symmetry systems and can be applied to any property that can be expressed as a function of the cluster vector. For example, one could constrain the mixing energy of the pure Au and Pd structures to be 0, give higher weight to structures close to some specific composition, or use weights to compensate for an uneven sampling of the configuration space.

### D. Conclusions

In this section, we have presented three ways to adapt the CE construction process so as to achieve accurate models for complex systems, such as surfaces. These approaches should be seen as tools that can be used on their own (as shown here) or in conjunction, and can be modified to provide a tailored solution for the problem at hand.

## VI. PART 3: MONTE CARLO SAMPLING

### Key takeaways

(1) MC simulations can be used to sample a CE in different thermodynamical ensembles.
(2) The variance-constrained semi-grand-canonical (VCSGC) ensemble can be used for sampling across miscibility gaps, but the semi-grand-canonical (SGC) ensemble cannot be used since the chemical potential maps to two different concentrations.
(3) Phase transitions can be identified from different thermodynamic properties, such as heat capacity as well as short-range and long-range order parameters.

In the third and last part of this tutorial, we show how the configuration space described by a CE can be sampled via MC simulations. This allows calculation of thermodynamic observables such as free energies, order parameters, and heat capacities (see Appendix B for how the latter two are obtained). We begin this section with a short introduction to MC simulations and the most common thermodynamic ensembles, and then illustrate these concepts using two examples.

### A. Monte Carlo simulations

Thermodynamic sampling is usually done via Metropolis MC simulations [104], which are generally executed as follows. Starting from some arbitrary initial state, a so-called trial step is suggested, which consists of a change in the atomic configuration. This step is either accepted (i.e., implemented) or rejected according to the Metropolis criterion with probability given by

$$\mathcal{P} = \min\{1, \exp(-\Delta\psi/k_{\text{B}}T)\},$$

where $\Delta\psi$ is the change in the relevant thermodynamic potential (which is introduced further in Sec. VI B). If the trial step is rejected, the system remains in its current state. The procedure is repeated until convergence is achieved.

### B. Thermodynamic ensembles

MC simulations can be executed in different thermodynamic ensembles. The choice of which depends on the goal of the simulation and the properties to be extracted.

The *canonical ensemble* models a situation where the total number of particles $N$ and their concentrations $c_i$ are kept constant along with the temperature $T$. It thus represents a system free to exchange heat with a reservoir at temperature $T$.

Strictly speaking, in the canonical ensemble (and similarly in the SGC and VCSGC ensembles; see below) the volume $V$ is constant. When constructing a CE, one, however, commonly trains CEs against configurations that have been relaxed with respect to both atomic positions and volume/cell shape. CE models trained in this fashion therefore incorporate the strain energy term that separates the canonical ensemble ($Nc_iVT$) from the isobaric-isothermal ensemble ($Nc_iTp$). It is therefore more sensible to interpret the results of such CE MC simulations in terms of the latter ensemble. In keeping with the literature, we, however, use the terms for the constant-volume ensembles in the following.

The trial steps used to sample the canonical ensemble need to preserve the conserved quantities, specifically the concentrations of the different species $c_i$. This can be accomplished by one (simultaneously) swapping the occupancy of two sites. The change of the thermodynamic potential $\psi$ is then given by the energy difference between the two states, i.e., $\Delta\psi = E_{\text{new}} - E_{\text{old}} = \Delta E$. In the context of this tutorial, $E$ refers to the energy predicted by the CE.

It is also possible to gradually reduce the temperature in an MC simulation in the canonical ensemble, an approach that is referred to as "*simulated annealing*." This procedure can be used as a general optimization algorithm, for example, to find the lowest-energy (ground-state) structures or in structure selection as implemented in Sec. IV B.

The "SGC ensemble" refers to a case where the total number of particles $N$ is fixed while the concentration of the different species is controlled via the relative chemical potentials $\Delta\mu_i$, again at constant temperature $T$. The SGC ensemble thus represents a system in connection with both a heat reservoir ($T$) and one or several particle reservoirs ($\Delta\mu_i$) under the constraint of a constant number of particles $N$.

In the CE literature, it is not uncommon for the SGC ensemble to be referred to as the "grand canonical ensemble," which is, however, a misnomer. The actual *grand canonical* ensemble represents an open system that has no constraint on the total number of particles, with the absolute chemical potentials $\mu_i$ as variables [105]. By contrast, the *semi-grand-canonical* ensemble represents a semiopen system, where the relative proportion of particles of different species but not their total number may change. In cases involving at least one sublattice with vacancies (such as the example discussed in Sec. IV), one can interpret CE-based MC simulations in the SGC ensemble or the VCSGC ensemble in terms of the grand canonical ensemble by imposing suitable thermodynamic boundary conditions. This approach is described and demonstrated in, for example, Ref. [27], which also discusses paraequilibrium and full equilibrium in this context.

As before, in the case of the canonical ensemble, trial moves used for sampling the SGC ensemble need to respect the respective constraints. A suitable trial move is therefore to change the occupation of a single site. The change in the thermodynamic potential is given by $\Delta\psi = \Delta E - N\sum_{i>1}\Delta c_i\Delta\mu_i$, where $N$ is the total number of sites, $c_i$ is the concentration of species $i$, and $\Delta\mu_i = \mu_i - \mu_1$ is the chemical potential difference of species $i$ relative to the first species.

A big advantage of the SGC ensemble compared with the canonical ensemble is that it directly connects to the derivative of the free energy per atom with respect to the concentration(s)

$$\frac{1}{N}\frac{\partial F}{\partial c_i} = -\Delta\mu_i,$$

which can be integrated along the concentration axis (or axes) to yield the free energy. Inside multiphase regions, the same chemical potentials map, however, to multiple different concentrations, which renders it impossible to perform integration over (and perform sampling inside) miscibility gaps.

The VCSGC ensemble can, in contrast to the SGC ensemble, be used for sampling inside and across miscibility gaps [106]. Similarly to the SGC ensemble, it imposes a flexible constraint on the mean concentrations, but unlike the SGC ensemble it also constrains their fluctuations. This is achieved via two variables, $\bar{\phi}$ and $\bar{\kappa}$, which control the mean and the variance of the concentration(s), respectively. The corresponding thermodynamic potential is given by $\Delta\psi = \Delta E + Nk_{\text{B}}T\bar{\kappa}(\Delta c + \bar{\phi}/2)^2$, and the ensemble can be sampled with the same kind of moves as those used for the SGC ensemble.

The free energy derivative per atom with respect to the concentration is given by

$$\frac{1}{N}\frac{\partial F}{\partial c} = -2k_{\text{B}}T\bar{\kappa}\left(\langle c\rangle + \bar{\phi}/2\right). \tag{9}$$

As a result of the constraint on the variance of the concentration, one can also stabilize concentrations inside

miscibility gaps. This allows one to obtain the free energy as a continuous function of concentration, and thus enables free energy integration.

The choice of $\bar{\kappa}$ affects the strength of the variance constraint, corresponding to the inverse variance of the expected concentration. Larger values enforce smaller fluctuations, which reduces the acceptance ratios, requiring more sampling. Smaller values, on the other hand, can cause the fluctuations to become too large to allow sampling of two-phase regions as the VCSGC ensemble then approaches the SGC ensemble [106]. Empirically, we have found that $\bar{\kappa} = 200$ strikes a good balance between sampling efficiency and sampling quality for most systems.

We can get an idea of how to select suitable values of $\bar{\phi}$ by considering Eq. (9). For typical temperatures and values of $\bar{\kappa}$, the terms in front of the left parenthesis on the right-hand side of Eq. (9) are much larger than the variation of the free energy on the left-hand side. As a result, we can assume that $2\langle c\rangle + \bar{\phi} \approx 0$. It then follows that the concentration limits $\langle c\rangle \to 0$ and $\langle c\rangle \to 1$ are reached by one setting $\bar{\phi} \approx -2 - \delta$ and $\bar{\phi} \approx \delta$, respectively. Here $\delta$ indicates that sampling needs to be done somewhat beyond those limits in order to cover the full composition range. Empirically, a value of $\delta$ of about 0.3 is sufficient in most cases.

### C. Ordering in $Au_3Pd$ using the canonical ensemble

As the first illustration we consider the order-disorder transition in AuPd3, which can be conveniently accessed via simulations in the canonical ensemble. Here we use a CE developed in Ref. 27 for $Au_{1-x}Pd_x$ on the fcc lattice. In Fig. 11 we show the resulting long-range order parameter and heat capacity as a function of temperature. The long-range order is represented here by the partial static structure factor calculated between Pd atoms at $\mathbf{q} = (2\pi/a_0)[1, 0, 0]$ [see Eq. (B2)], while the heat capacity is calculated via Eq. (B1).

The sharp peak in the heat capacity at approximately 175 K indicates that the material undergoes a continuous phase transition from a disordered phase to an ordered phase [see the inset in Fig. 11(b)] with decreasing temperature. The phase transition can also be seen in the long-range order parameter, as this increases sharply at the phase transition.

All simulations indicate a phase transition at around 175 K, but there the sharpness of the transition is sensitive to supercell size, with larger systems yielding sharper transitions. This behavior is typical for continuous phase transitions, as the correlation length of the fluctuations in the systems diverges at the critical temperature [107]. To achieve convergence one therefore needs to consider a range of system sizes, and extrapolate to the infinite size limit if possible.

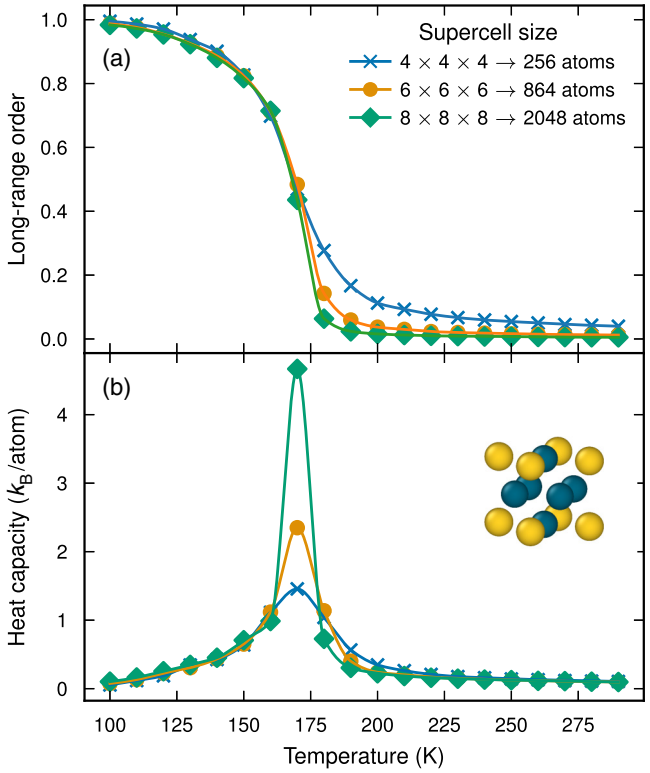


FIG. 11. Order-disorder transition in $AuPd_3$ from MC simulations. (a) Long-range order parameter and (b) heat capacity as a function of temperature and system size, where the solid lines are guides for the eye. The system size is given in multiples of the primitive (four-atom) unit cell. The long-range order is represented by the partial static structure factor between Pd atoms. The inset shows the ordered low-temperature phase.

### D. Phase diagram of $Ag_xPd_{1-x}$ via the SGC and VCSGC ensembles

We now illustrate the construction of a phase diagram for a system with a miscibility gap, namely fcc $Ag_xPd_{1-x}$, using a CE from Ref. [68], a $10\times10\times10$ supercell and MC simulations in the SGC and VCSGC ensembles.

In the case of the SGC MC simulations, the chemical potential was sampled with 105 evenly spaced points between $-1.04$ and 1.04 eV/atom. For the VCSGC MC simulations, the variance constraint parameter $\bar{\kappa}$ was set to 200, while the average constraint parameter $\bar{\phi}$ was varied between $-2.3$ and 0.3 with 105 evenly spaced points.

#### *1. Miscibility gap*

Figure 12(a) shows the free energy derivative obtained from sampling in both ensembles at a temperature below (400 K) and above (800 K) the miscibility gap. Here the miscibility gap separates a Pd-rich phase containing almost no Ag from a mixed phase with up to approximately 60%–70% Pd depending on temperature.

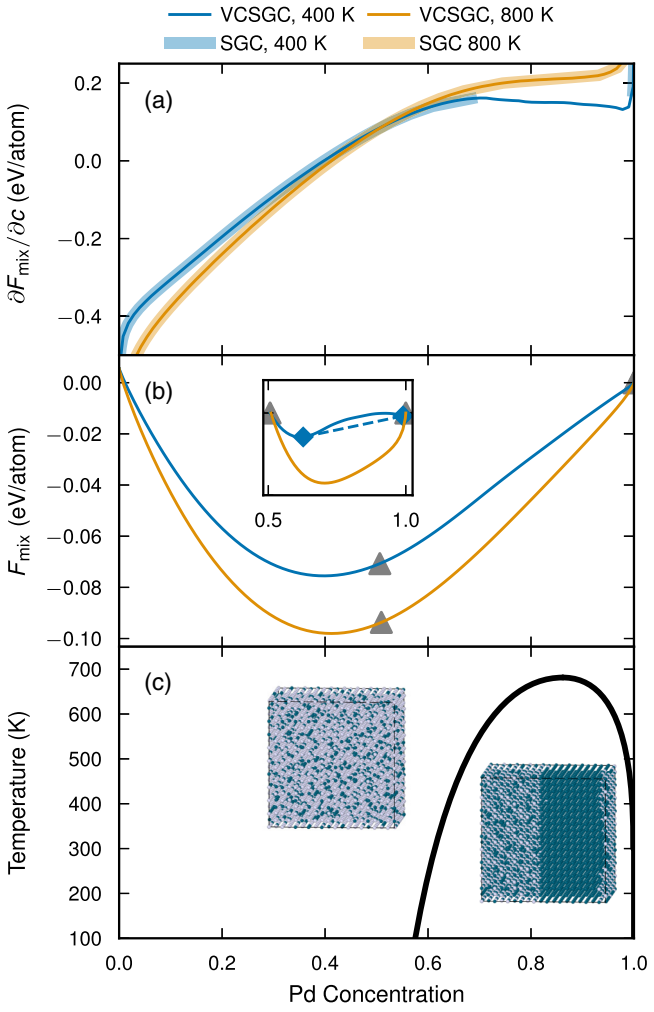


FIG. 12. MC sampling of the $Ag_xPd_{1-x}$ system. (a) Derivative of the free energy of mixing with respect to composition (i.e., the chemical potential), (b) free energy of mixing, and (c) phase diagram. The inset in (b) shows the enlarged and tilted free energy of mixing in the vicinity of the miscibility gap (indicated by the blue diamonds). Because of the concave free energy at 400 K, the SGC ensemble cannot form a stable solution. The miscibility gap is extracted from the free energy landscape of the VCSGC ensemble.

Outside the miscibility gap, simulations in the SGC and VCSGC ensembles yield numerically identical results. In the SGC MC simulations at 400 K it is, however, apparent that the miscibility gap, i.e., the two-phase region between approximately 60% and 99%, is not accessible as the mapping between $\partial F/\partial c$ and the concentration $c$ is one-to-many (i.e., noninjective), and the variance of the concentration diverges. Here the identification of the free energy derivative with the relative chemical potential breaks down, as the latter is meaningfully defined only in single-phase regions.

When SGC MC simulations are used, the free energy in the single-phase regions can still be obtained through thermodynamic integration starting from known limits such as in the low-temperature or high-temperature expansions [108]. Furthermore, one can obtain the phase diagram by tracing the phase boundaries [108]. To mitigate hysteresis effects, this requires iterative sampling across two-phase regions as well as careful consideration of the convergence with respect to the numerical resolution of the chemical potential.

To counteract the divergence of the concentration in miscibility gaps, in the VCSGC ensemble one includes a term representing a *constraint* on the variance in the thermodynamic potential. This allows one to access also compositions inside the miscibility gap and sample the free energy derivative as a continuous function of composition [Fig. 12(a)]. This, in turn, enables integration of the free energy over the entire concentration range [Fig. 12(b)]. Finally, via the convex hull construction [inset in Fig. 12(b)], one can then readily extract the phase boundaries [Fig. 12(c)].

Thanks to the possibility to compute the free energy across two-phase regions, one can furthermore extract *excess* free energies. Thereby, it is possible to compute, for example, interface free energies as a function of interface orientation. The interested reader can find more information on this topic in, for example, Refs. [83,106].

Finally, the VCSGC ensemble naturally enables an even sampling of the concentration axis as equidistant values of $\bar{\phi}$ lead to approximately equidistant concentrations. This avoids adaptive sampling in the vicinity of phase transitions that are necessary when the SGC ensemble is used.

### *2. Secondary phase transitions*

By mapping out the free energy, we were able to locate the miscibility gap in the Ag-Pd system. This phase boundary corresponds to a first-order phase transition as evident from the sudden (and discontinuous) transition in the derivative of the free energy with respect to composition. It is, however, very difficult (if not practically impossible for numerical reasons) to obtain information about continuous phase transitions. To this end, as we saw ín Sec. VI C, the heat capacity and order parameters are better suited.

Analysis of the heat capacity as a function of temperature and composition reveals a transition at around 25% Pd and 200 K [Fig. 13(a)]. This transition closely resembles the one in $Au_3Pd$ that we analyzed before, but here we also obtain the composition dependence of the transition. At 25% Pd the transition is very sharp, while both the heat capacity and the transition temperature drop with either decreasing or increasing composition. Note that the heat capacity shows no discernible features around the miscibility gap.

The order-disorder transition at around 25% Pd is also apparent from the short-range order parameter computed

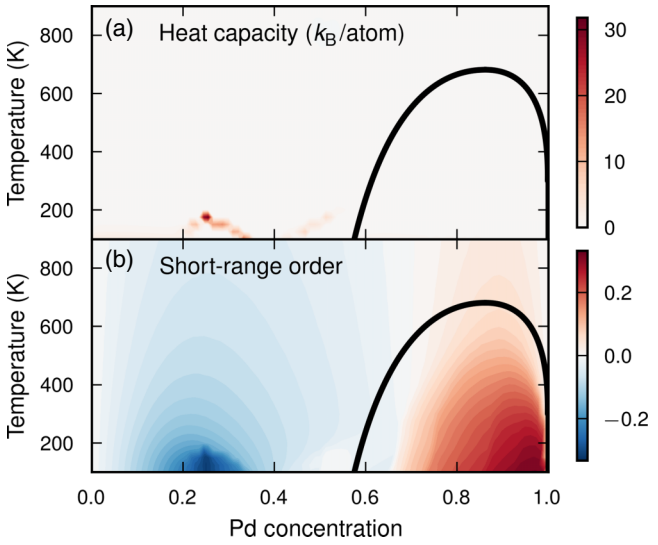


FIG. 13. Order-disorder transitions in $Ag_xPd_{1-x}$ from VCSGC MC simulations. (a) The heat capacity map reveals a continuous phase transition between a disordered phase and an ordered phase at around 25% Pd. (b) The short-range order parameter [Eq. (B3)] indicates the existence of a miscibility gap; however, it is hard to identify the exact boundaries. It also shows a sharp change at the order-disorder phase transition.

according to Eq. (B3) [Fig. 13(b)]. The short-range order map (unlike the heat capacity), however, also exhibits structure across the miscibility gap, which is the result of the variation in the proportions of the two different phases (Pd rich and mixed). It is, however, clearly apparent that the short-range order is not suited for identifying the actual phase boundaries as the map is smooth, and distinct features emerge only well inside the miscibility gap.

### E. Conclusions

In the final part of this tutorial, we demonstrated the utility of MC simulations in different thermodynamic ensembles for extracting thermodynamic observables and identifying phase boundaries. Which ensemble is best suited depends on the task at hand. Generally speaking, the canonical ensemble is simple to use (and understand) but limited insofar as it does not directly enable one to observe a first derivative of the free energy. It can still be used for thermodynamic integration (which was not covered here). The SGC and VCSGC ensembles both provide access to the first derivative of the free energy with respect to concentration and thereby can be used more readily for obtaining phase diagrams. The VCSGC ensemble additionally allows one to handily extract excess free energies. Here we did not discuss acceptance ratios, but in closing, we point out that while the canonical ensemble tends to yield higher acceptance ratios for small concentrations, the other two ensembles achieve higher ratios at higher concentrations [109].

## VII. OUTLOOK

In this tutorial we have provided a short practical introduction to the topic of constructing and sampling CEs for the study of many-component systems. These techniques can be used to investigate a huge range of materials, including, in particular, materials with relevance to energy applications. In the field of battery research, for example, CEs have been applied to study chemical ordering [110], voltage curves [111], and migration barriers [112,113], whereas with respect to photovoltaic materials, successful applications of CEs include studies of the phase stability of perovskites [32,114], band-gap engineering [115,116], and chemical order and the effect of vacancies [117]. Catalysis is another field where CEs have proven to be useful—for instance, when one is studying the effect of nanoparticle size and shape [118] and composition [119], as well as for comparing active sites [120] and performing high-throughput compositional screening of candidate materials [121]. In the field of thermoelectrics, CE have been used to study the impact of chemical order [37] and composition [39,122] on phase stability and even transport properties, for materials such as Si-Ge nanowires [122], clathrates [37,39,63], and skutterudites [123]. Lastly, CEs are of importance for studying various high-performance alloys, including superalloys [124], high-entropy alloys [125], and intermetallic compounds [126,127]. These materials often exhibit favorable thermal and mechanical properties and high tunability, and have applications in, for instance, lightweight construction and high-temperature environments.

Most applications of CEs consist of CE construction and MC sampling, which means that the contents of parts 1 and 3 of this tutorial are relevant in most cases. The structure selection and training strategies from part 1 are of particular importance for more complex systems, such as the multicomponent high-performance alloys mentioned above, where these aspects are especially challenging due to the large number of components and/or larger unit cells. Similarly, the content of part 3 is of particular interest for studies with an emphasis on phase diagrams, phase transitions, and chemical order, including in thermoelectrics and high-performance alloys. In addition, understanding the thermodynamic ensembles, in particular the SGC and VCSGC ensembles, is crucial when the chemical potential is of interest—for instance, for battery voltage curves and systems in contact with gases. Lastly, the strategies in part 2 are useful for studies of surfaces and nanoparticles, which is almost always the case in the field of catalysis, but can also be applied for other challenging systems where the strategies from part 1 are not sufficient.

Finally, we emphasize that there are various more advanced topics related to CEs that are beyond the scope of this tutorial that we invite so-inclined readers to pursue. With regard to the *construction* of CEs, this includes managing sublattices [27] and strain [25,82,83], handling thermodynamic constraints through nullspaces [27,128], or quadratic programming [129] as well as alternative regression approaches [130]. In terms of *applications*, one can mention, for example, ground-state finding [131, 132], materials under pressure [126], precipitate formation [133], the description of migration barriers [60], and CE-based kinetic MC simulations [136].

The data that support the findings of this article are openly available [135,136].

## ACKNOWLEDGMENTS

We gratefully acknowledge funding from the Swedish Research Council (Grants No. 2020-04935 and No. 2021-05072), the Swedish Energy Agency (Grant No. 45410-1), the Excellence Initiative Nano at Chalmers University of Technology, and the Chalmers Initiative for Advancement of Neutron and Synchrotron Techniques. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden partially funded by the Swedish Research Council through Grant Agreements No. 2022-06725 and No. 2018-05973.

## APPENDIX A: COMPUTATIONAL DETAILS

The DFT calculations for the simple carbide system, $Mo_{1-x}V_xC_{1-y}\square_y$, were performed with the projector-augmented method [137,138] as implemented in VASP [139–141]. We used the van der Waals density functional method with consistent exchange [142,143] for the exchange-correlation potential. The Brillouin zone was sampled with a Γ-centered **k**-point grid, with the smallest allowed spacing between two **k**-grid points being 0.25 $Å^{-1}$ and the plane-wave cutoff energy being 520 eV. Both the positions and the cell were allowed to relax; the largest allowed residual forces were set to 0.02 $eVÅ^{-1}$.

## APPENDIX B: THERMODYNAMIC PROPERTIES

In addition to the properties mentioned in Appendix A, one can obtain several other thermodynamic properties of interest from MC simulations.

The heat capacity is readily obtained via

$$C_V(T) = \frac{\mathrm{var}[E(T)]}{k_\mathrm{B}T^2}, \tag{B1}$$

where var$[E(T)]$ is the variance of the energy.

Long-range order can be assessed from the partial static structure factors, which are defined for atom types $A$ and $B$ as

$$S_AB(\boldsymbol{q}) \propto \sum_j^{N_A}\sum_k^{N_B} \exp\left[-\mathrm{i}\boldsymbol{q}\cdot(\boldsymbol{R}_j - \boldsymbol{R}_k)\right], \tag{B2}$$

where $\boldsymbol{q}$ is a reciprocal lattice point and $\boldsymbol{R}_k$ is the position of atom $k$.

Short-range order is often measured in terms of the Warren-Cowley short-range order parameter, which determines to which degree a binary ($A$-$B$) system mixes or segregates [144]. For an atom $i$ of type $A$, it is defined as

$$\alpha_i = 1 - \frac{Z_B}{Z_{\mathrm{tot}}c_B}, \tag{B3}$$

where $Z_B$ is the number of $B$ neighbors in the first-neighbor shell, $Z_{\mathrm{tot}}$ is the total number of neighbors in the first shell, and $c_B$ is the concentration of $B$ atoms. For a random mixture, we obtain $\alpha = 0$, for a phase-separated system, we get $\alpha > 0$, while $\alpha < 0$ indicates ordering.

[1] W. Ma, J. M. Luther, H. Zheng, Y. Wu, and A. P. Alivisatos, Photovoltaic devices employing ternary $PbS_xSe_{1-x}$ nanocrystals, Nano Lett. **9**, 1699 (2009).
[2] Z.-Y. Xiao, Y.-F. Li, B. Yao, R. Deng, Z.-H. Ding, T. Wu, G. Yang, C.-R. Li, Z.-Y. Dong, L. Liu, L.-G. Zhang, and H.-F. Zhao, Bandgap engineering of $Cu_2Cd_xZn_{1-x}SnS_4$ alloy for photovoltaic applications: A complementary experimental and first-principles study, J. Appl. Phys. **114**, 183506 (2013).
[3] N. Lu and I. Ferguson, III-nitrides for energy production: Photovoltaic and thermoelectric applications, Semicond. Sci. Technol. **28**, 074023 (2013).
[4] T. Feurer, P. Reinhard, E. Avancini, B. Bissig, J. Löckinger, P. Fuchs, R. Carron, T. P. Weiss, J. Perrenoud, S. Stutterheim, S. Buecheler, and A. N. Tiwari, Progress in thin film CIGS photovoltaics – Research and development, manufacturing, and applications, Prog. Photovolt.: Res. Appl. **25**, 645 (2016).
[5] W. Li, X.-F. Cai, N. Valdes, T. Wang, W. Shafarman, S.-H. Wei, and A. Janotti, $In_2Se_3$, $In_2Te_3$, and $In_2(Se, Te)_3$ alloys as photovoltaic materials, J. Phys. Chem. Lett. **13**, 12026 (2022).
[6] M.-S. Balogun, Y. Luo, W. Qiu, P. Liu, and Y. Tong, A review of carbon materials and their composites with alloy metals for sodium ion battery anodes, Carbon **98**, 162 (2016).
[7] M. Lao, Y. Zhang, W. Luo, Q. Yan, W. Sun, and S. X. Dou, Alloy-based anode materials toward advanced sodium-ion batteries, Adv. Mater. **29**, 1700622 (2017).
[8] K. Song, C. Liu, L. Mi, S. Chou, W. Chen, and C. Shen, Recent progress on the alloy-based anode for sodium-ion batteries and potassium-ion batteries, Small **17**, 1903194 (2019).
[9] A. K. Singh and Q. Xu, Synergistic catalysis over bimetallic alloy nanoparticles, ChemCatChem **5**, 652 (2013).

[10] R. T. Hannagan, G. Giannakakis, M. Flytzani-Stephanopoulos, and E. C. H. Sykes, Single-atom alloy catalysis, Chem. Rev. **120**, 12044 (2020).

[11] Y. Nakaya and S. Furukawa, Catalysis of alloys: Classification, principles, and design for a variety of materials and reactions, Chem. Rev. **123**, 5859 (2022).

[12] E. Antolini, Formation of carbon-supported PtM alloys for low temperature fuel cells: A review, Mater. Chem. Phys. **78**, 563 (2003).

[13] X. Ren, Q. Lv, L. Liu, B. Liu, Y. Wang, A. Liu, and G. Wu, Current progress of Pt and Pt-based electrocatalysts used for fuel cells, Sustain. Energy Fuels **4**, 15 (2020).

[14] T. Gong, P. Lyu, K. J. Palm, S. Memarzadeh, J. N. Munday, and M. S. Leite, Emergent opportunities with metallic alloys: From material design to optical devices, Adv. Opt. Mater. **8**, 2001082 (2020).

[15] I. Darmadi, S. Z. Khairunnisa, D. Tomeček, and C. Langhammer, Optimization of the composition of PdAuCu ternary alloy nanoparticles for plasmonic hydrogen sensing, ACS Appl. Nano Mater. **4**, 8716 (2021).

[16] T. Dursun and C. Soutis, Recent developments in advanced aircraft aluminium alloys, Mater. Des. (1980-2015) **56**, 862 (2014).

[17] J. García, V. Collado Ciprés, A. Blomqvist, and B. Kaplan, Cemented carbide microstructures: A review, Int. J. Refract. Met. Hard Mater. **80**, 40 (2019).

[18] J. M. Torralba and M. Campos, High entropy alloys manufactured by additive manufacturing, Metals **10**, 639 (2020).

[19] Z.-X. Zhang, J. Zhang, H. Wu, Y. Ji, and D. D. Kumar, Iron-based shape memory alloys in construction: Research, applications and opportunities, Materials **15**, 1723 (2022).

[20] B. Anasori, M. R. Lukatskaya, and Y. Gogotsi, 2D metal carbides and nitrides (MXenes) for energy storage, Nat. Rev. Mater. **2**, 16098 (2017).

[21] X. Tang, X. Guo, W. Wu, and G. Wang, 2D metal carbides and nitrides (MXenes) as high-performance electrode materials for lithium-based batteries, Adv. Energy Mater. **8**, 1801897 (2018).

[22] Y. Gogotsi and B. Anasori, The rise of MXenes, ACS Nano **13**, 8491 (2019).

[23] L. Yin, Y. Li, X. Yao, Y. Wang, L. Jia, Q. Liu, J. Li, Y. Li, and D. He, MXenes for solar cells, Nano-Micro Lett. **13**, 78 (2021).

[24] M. Asta, R. McCormack, and D. de Fontaine, Theoretical study of alloy phase stability in the Cd-Mg system, Phys. Rev. B **48**, 748 (1993).

[25] V. Ozoliņš, C. Wolverton, and A. Zunger, Cu-Au, Ag-Au, Cu-Ag, and Ni-Au intermetallics: First-principles study of temperature-composition phase diagrams and structures, Phys. Rev. B **57**, 6427 (1998).

[26] Y. S. Meng and M. E. Arroyo-de Dompablo, First principles computational materials design for energy storage materials in lithium ion batteries, Energy Environ. Sci. **2**, 589 (2009).

[27] J. M. Rahm, J. Löfgren, E. Fransson, and P. Erhart, A tale of two phase diagrams: Interplay of ordering and hydrogen uptake in Pd–Au–H, Acta Mater. **211**, 116893 (2021).

[28] M. Gren, E. Fransson, M. Ångqvist, P. Erhart, and G. Wahnström, Modeling of vibrational and configurational degrees of freedom in hexagonal and cubic tungsten carbide at high temperatures, Phys. Rev. Mater. **5**, 033804 (2021).

[29] A. Van der Ven, M. K. Aydinol, G. Ceder, G. Kresse, and J. Hafner, First-principles investigation of phase stability in $Li_xCoO_2$, Phys. Rev. B **58**, 2975 (1998).

[30] G. Ceder, A. Van der Ven, C. Marianetti, and D. Morgan, First-principles alloy theory in oxides, Model. Simul. Mater. Sci. Eng. **8**, 311 (2000).

[31] F. Zhou, T. Maxisch, and G. Ceder, Configurational electronic entropy and the phase diagram of mixed-valence oxides: The case of $Li_xFePO_4$, Phys. Rev. Lett. **97**, 155704 (2006).

[32] J. S. Bechtel and A. Van der Ven, First-principles thermodynamics study of phase stability in inorganic halide perovskite solid solutions, Phys. Rev. Mater. **2**, 045401 (2018).

[33] C. Linderälv, J. M. Rahm, and P. Erhart, High-throughput characterization of transition metal dichalcogenide alloys: Thermodynamic stability and electronic band alignment, Chem. Mater. **34**, 9364 (2022).

[34] H. Kim, M. Kaviany, J. C. Thomas, A. Van der Ven, C. Uher, and B. Huang, Structural order-disorder transitions and phonon conductivity of partially filled skutterudites, Phys. Rev. Lett. **105**, 265901 (2010).

[35] B. P. Burton, A. van de Walle, and H. T. Stokes, First principles phase diagram calculations for the octahedral-interstitial system $ZrO_x$, $0 \leq x \leq 1/2$, J. Phys. Soc. Jpn. **81**, 014004 (2011).

[36] B. P. Burton and A. van de Walle, First principles phase diagram calculations for the octahedral-interstitial system $ZrO_x$, $0 \leq x \leq 1/2$, Calphad **37**, 151 (2012).

[37] M. Ångqvist, D. O. Lindroth, and P. Erhart, Optimization of the thermoelectric power factor: Coupling between chemical order and transport properties, Chem. Mater. **28**, 6877 (2016).

[38] M. Ångqvist and P. Erhart, Understanding chemical ordering in intermetallic clathrates from atomic scale simulations, Chem. Mater. **29**, 7554 (2017).

[39] M. Troppenz, S. Rigamonti, and C. Draxl, Predicting ground-state configurations and electronic properties of the thermoelectric clathrates $Ba_8Al_xSi_{46-x}$ and $Sr_8Al_xSi_{46-x}$, Chem. Mater. **29**, 2414 (2017).

[40] N. S. H. Gunda, B. Puchala, and A. Van der Ven, Resolving phase stability in the Ti-O binary with first-principles statistical mechanics methods, Phys. Rev. Mater. **2**, 033604 (2018).

[41] R. Drautz, H. Reichert, M. Fähnle, H. Dosch, and J. M. Sanchez, Spontaneous $L1_2$ order at $Ni_{90}Al_{10}(110)$ surfaces: An X-ray and first-principles-calculation study, Phys. Rev. Lett. **87**, 236102 (2001).

[42] M. H. F. Sluiter and Y. Kawazoe, Cluster expansion method for adsorption: Application to hydrogen chemisorption on graphene, Phys. Rev. B **68**, 085410 (2003).

[43] P. Welker, O. Wieckhorst, T. C. Kerscher, and S. Müller, Predicting the segregation profile of the $Pt_{25}Rh_{75}$ (100)

surface from first-principles, J. Phys.: Condens. Matter **22**, 384203 (2010).
[44] J. A. Stephens, H. C. Ham, and G. S. Hwang, Atomic arrangements of AuPt/Pt(111) and AuPd/Pd(111) surface alloys: A combined density functional theory and Monte Carlo study, J. Phys. Chem. C **114**, 21516 (2010).
[45] W. Chen, D. Schmidt, W. F. Schneider, and C. Wolverton, Ordering and oxygen adsorption in Au–Pt/Pt(111) surface alloys, J. Phys. Chem. C **115**, 17915 (2011).
[46] L. Cao and T. Mueller, Rational design of $Pt_3Ni$ surface structures for the oxygen reduction reaction, J. Phys. Chem. C **119**, 17735 (2015).
[47] L. M. Herder, J. M. Bray, and W. F. Schneider, Comparison of cluster expansion fitting algorithms for interactions at surfaces, Surf. Sci. **640**, 104 (2015).
[48] E. Fransson, M. Gren, and G. Wahnström, Complexions and grain growth retardation: First-principles modeling of phase boundaries in WC-Co cemented carbides at elevated temperatures, Acta Mater. **216**, 117128 (2021).
[49] E. Fransson, M. Gren, H. Larsson, and G. Wahnström, First-principles modeling of complexions at the phase boundaries in Ti-doped WC-Co cemented carbides at finite temperatures, Phys. Rev. Mater. **5**, 093801 (2021).
[50] P. Ekborg-Tanner and P. Erhart, Hydrogen-driven surface segregation in Pd alloys from atomic-scale simulations, J. Phys. Chem. C **125**, 17248 (2021).
[51] J.-Z. Xie and H. Jiang, Revealing carbon vacancy distribution on $\alpha$-$MoC_{1-x}$ surfaces by machine-learning force-field-aided cluster expansion approach, J. Phys. Chem. C **127**, 13228 (2023).
[52] T. Mueller and G. Ceder, Effect of particle size on hydrogen release from sodium alanate nanoparticles, ACS Nano **4**, 5647 (2010).
[53] R. V. Chepulskii, W. H. Butler, A. van de Walle, and S. Curtarolo, Surface segregation in nanoparticles from first principles: The case of FePt, Scr. Mater. **62**, 179 (2010).
[54] K. Yuge, Concentration effects on segregation behavior of Pt-Rh nanoparticles, Phys. Rev. B **84**, 1 (2011).
[55] T. Mueller, Ab initio determination of structure-property relationships in alloy nanoparticles, Phys. Rev. B **86**, 144201 (2012).
[56] T. L. Tan, L.-L. Wang, D. D. Johnson, and K. Bai, A comprehensive search for stable Pt-Pd nanoalloy configurations and their use as tunable catalysts, Nano Lett. **12**, 4875 (2012).
[57] L.-L. Wang, T. L. Tan, and D. D. Johnson, Configurational thermodynamics of alloyed nanoparticles with adsorbates, Nano Lett. **14**, 7077 (2014).
[58] C. Li, D. Raciti, T. Pu, L. Cao, C. He, C. Wang, and T. Mueller, Improved prediction of nanoalloy structures by the explicit inclusion of adsorbates in cluster expansions, J. Phys. Chem. C **122**, 18040 (2018).
[59] L. Cao, C. Li, and T. Mueller, The use of cluster expansions to predict the structures and properties of surfaces and nanostructured materials, J. Chem. Inf. Model. **58**, 2401 (2018).
[60] A. Van der Ven, G. Ceder, M. Asta, and P. D. Tepesch, First-principles theory of ionic diffusion with nondilute carriers, Phys. Rev. B **64**, 184307 (2001).
[61] D. Morgan, A. van de Walle, G. Ceder, J. D. Althoff, and D. de Fontaine, Vibrational thermodynamics: Coupling of chemical order and size effects, Model. Simul. Mater. Sci. Eng. **8**, 295 (2000).
[62] A. van de Walle and G. Ceder, The effect of lattice vibrations on substitutional alloy thermodynamics, Rev. Mod. Phys. **74**, 11 (2002).
[63] J. Brorsson, Y. Zhang, A. E. C. Palmqvist, and P. Erhart, Order-disorder transition in inorganic clathrates controls electrical transport properties, Chem. Mater. **33**, 4500 (2021).
[64] A. van de Walle, M. Asta, and G. Ceder, The Alloy Theoretic Automated Toolkit: A user guide, Calphad **26**, 539 (2002).
[65] D. Lerch, O. Wieckhorst, G. L. W. Hart, R. W. Forcade, and S. Müller, UNCLE: A code for constructing cluster expansions for arbitrary lattices with minimal user-input, Model. Simul. Mater. Sci. Eng. **17**, 055003 (2009).
[66] J. H. Chang, D. Kleiven, M. Melander, J. Akola, J. M. Garcia-Lastra, and T. Vegge, CLEASE: A versatile and user-friendly implementation of cluster expansion method, J. Phys.: Condens. Matter **31**, 325901 (2019).
[67] B. Puchala, J. C. Thomas, A. R. Natarajan, J. G. Goiri, S. S. Behara, J. L. Kaufman, and A. Van der Ven, CASM — A software package for first-principles based study of multicomponent crystalline solids, Comput. Mater. Sci. **217**, 111897 (2023).
[68] M. Ångqvist, W. A. Muñoz, J. M. Rahm, E. Fransson, C. Durniak, P. Rozyczko, T. H. Rod, and P. Erhart, ICET—A Python library for constructing and sampling alloy cluster expansions, Adv. Theory Simul. **2**, 1900015 (2019).https://icet.materialsmodeling.org/.
[69] L. Barroso-Luque, J. H. Yang, F. Xie, T. Chen, R. L. Kam, Z. Jadidi, P. Zhong, and G. Ceder, SMOL: A Python package for cluster expansions and beyond, J. Open Source Softw. **7**, 4504 (2022).
[70] Cluster expansion tutorials, https://ce-tutorials.materialsmodeling.org, accessed 2024-04-23.
[71] The Jupyter notebooks associated with this tutorial as well as underlying data are available at doi:10.5281/zenodo.10997198.
[72] J. M. Sanchez, F. Ducastelle, and D. Gratias, Generalized cluster description of multicomponent systems, Phys. A: Stat. Mech. Appl. **128**, 334 (1984).
[73] J. M. Sanchez, Cluster expansion and the configurational theory of alloys, Phys. Rev. B **81**, 224202 (2010).
[74] D. de Fontaine, in *Solid State Physics*, Vol. 47, edited by H. Ehrenreich and D. Turnbull (Academic Press, San Diego, 1994), p. 33.
[75] A. Zunger, L. G. Wang, G. L. W. Hart, and M. Sanati, Obtaining Ising-like expansions for binary alloys from first principles, Model. Simul. Mater. Sci. Eng. **10**, 685 (2002).
[76] A. van de Walle, Multicomponent multisublattice alloys, nonconfigurational entropy and other additions to the Alloy Theoretic Automated Toolkit, Calphad Tools for Computational Thermodynamics **33**, 266 (2009).
[77] J.-Z. Xie, X.-Y. Zhou, and H. Jiang, Perspective on optimal strategies of building cluster expansion models for

configurationally disordered materials, J. Chem. Phys. **157**, 200901 (2022).
[78] T. Mueller, Comment on "Cluster expansion and the configurational theory of alloys", Phys. Rev. B **95**, 216201 (2017).
[79] J. M. Sanchez, Reply to "Comment on 'Cluster expansion and the configurational theory of alloys"', Phys. Rev. B **95**, 216202 (2017).
[80] L. Barroso-Luque, P. Zhong, J. H. Yang, F. Xie, T. Chen, B. Ouyang, and G. Ceder, Cluster expansions of multicomponent ionic materials: Formalism and methodology, Phys. Rev. B **106**, 144202 (2022).
[81] M. Fant, M. Ångqvist, A. Hellman, and P. Erhart, To every rule there is an exception: A rational extension of Loewenstein's rule, Angew. Chem. - Int. Ed. **60**, 5132 (2021).
[82] D. B. Laks, L. G. Ferreira, S. Froyen, and A. Zunger, Efficient cluster expansion for substitutional systems, Phys. Rev. B **46**, 12587 (1992).
[83] J. M. Rahm, J. Löfgren, and P. Erhart, Quantitative predictions of thermodynamic hysteresis: Temperature-dependent character of the phase transition in Pd–H, Acta Mater. **227**, 117697 (2022).
[84] L. J. Nelson, V. Ozoliņš, C. S. Reese, F. Zhou, and G. L. W. Hart, Cluster expansion made easy with Bayesian compressive sensing, Phys. Rev. B **88**, 155105 (2013).
[85] F. Zhou, W. Nielson, Y. Xia, and V. Ozoliņš, Lattice anharmonicity and thermal conductivity from compressive sensing of first-principles calculations, Phys. Rev. Lett. **113**, 185501 (2014).
[86] E. Fransson, F. Eriksson, and P. Erhart, Efficient construction of linear models in materials modeling and applications to force constant expansions, npj Comput. Mater. **6**, 135 (2020).
[87] Y. Cheng, L. Zhu, J. Zhou, and Z. Sun, pyGACE: Combining the genetic algorithm and cluster expansion methods to predict the ground-state structure of systems containing point defects, Comput. Mater. Sci. **174**, 109482 (2020).
[88] D. J. C. MacKay, Bayesian interpolation, Neural Comput. **4**, 415 (1992).
[89] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011).
[90] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Contr. **19**, 716 (1974).
[91] G. Schwarz, Estimating the dimension of a model, Ann. Stat. **6**, 461 (1978).
[92] K. Aho, D. Derryberry, and T. Peterson, Model selection for ecologists: The worldviews of AIC and BIC, Ecology **95**, 631 (2014).
[93] K. P. Murphy, *Machine Learning: A Probabilistic Perspective* (MIT Press, Cambridge, MA, 2013).
[94] T. Mueller and G. Ceder, Bayesian approach to cluster expansions, Phys. Rev. B **80**, 024103 (2009).
[95] H. Zou, The adaptive lasso and its oracle properties, J. Am. Stat. Assoc. **101**, 1418 (2006).
[96] E. Candes and T. Tao, Decoding by linear programming, IEEE Trans. Inf. Theory **51**, 4203 (2005).
[97] A. van de Walle and G. Ceder, Automating first-principles phase diagram calculations, J. Phase Equilib. **23**, 348 (2002).
[98] A. Seko, Y. Koyama, and I. Tanaka, Cluster expansion method for multicomponent systems based on optimal selection of structures for density-functional theory calculations, Phys. Rev. B **80**, 165122 (2009).
[99] D. Kleiven, J. Akola, A. A. Peterson, T. Vegge, and J. H. Chang, Training sets based on uncertainty estimates in the cluster-expansion method, J. Phys.: Energy **3**, 034012 (2021).
[100] G. H. Golub and C. F. Van Loan, *Matrix Computations* (Johns Hopkins University Press, Philadelphia, PA, 2013), 4th ed.
[101] L. J. Nelson, G. L. W. Hart, F. Zhou, and V. Ozoliņš, Compressive sensing as a paradigm for building physics models, Phys. Rev. B **87**, 035125 (2013).
[102] G. L. W. Hart and R. W. Forcade, Algorithm for generating derivative structures, Phys. Rev. B **77**, 224115 (2008).
[103] E. Cockayne and A. van de Walle, Building effective models from sparse but precise data: Application to an alloy cluster expansion model, Phys. Rev. B **81**, 012104 (2010).
[104] D. Frenkel and B. Smit, in *Understanding Molecular Simulation*, edited by D. Frenkel and B. Smit (Academic Press, San Diego, London, 2023), 3rd ed., Chap. 6, p. 181.
[105] R. Pathria and P. D. Beale, in *Statistical Mechanics (Fourth Edition)*, edited by R. Pathria and P. D. Beale (Academic Press, London, San Diego, 2022), 4th ed., p. 93.
[106] B. Sadigh and P. Erhart, Calculation of excess free energies of precipitates via direct thermodynamic integration across phase boundaries, Phys. Rev. B **86**, 134204 (2012).
[107] S.-K. Ma, *Statistical Mechanics* (World Scientific Publishing, Singapore, Singapore, 1985).
[108] A. van de Walle and M. Asta, Self-driven lattice-model Monte Carlo simulations of alloy thermodynamic properties and phase diagrams, Model. Simul. Mater. Sci. Eng. **10**, 521 (2002).
[109] B. Sadigh, P. Erhart, A. Stukowski, A. Caro, E. Martinez, and L. Zepeda-Ruiz, Scalable parallel Monte Carlo algorithm for atomistic simulations of precipitation in alloys, Phys. Rev. B **85**, 184203 (2012).
[110] A. Van der Ven and G. Ceder, Ordering in $Li_x(Ni_{0.5}Mn_{0.5})O_2$ and its relation to charge capacity and electrochemical behavior in rechargeable lithium batteries, Electrochem. Commun. **6**, 1045 (2004).
[111] T. Chen, G. Sai Gautam, W. Huang, and G. Ceder, First-principles study of the voltage profile and mobility of Mg intercalation in a chromium oxide spinel, Chem. Mater. **30**, 153 (2017).
[112] H.-C. Yu, C. Ling, J. Bhattacharya, J. C. Thomas, K. Thornton, and A. Van der Ven, Designing the next generation high capacity battery electrodes, Energy Environ. Sci. **7**, 1760 (2014).
[113] J. H. Chang, P. B. Jørgensen, S. Loftager, A. Bhowmik, J. M. G. Lastra, and T. Vegge, On-the-fly assessment of diffusion barriers of disordered transition metal oxyfluorides

using local descriptors, Electrochim. Acta **388**, 138551 (2021).
[114] K. Yamamoto, S. Iikubo, J. Yamasaki, Y. Ogomi, and S. Hayase, Structural stability of iodide perovskite: A combined cluster expansion method and first-principles study, J. Phys. Chem. C **121**, 27797 (2017).
[115] X. Xu and H. Jiang, Cluster expansion based configurational averaging approach to bandgaps of semiconductor alloys, J. Chem. Phys. **150**, 034102 (2019).
[116] G. Han, I. W. Yeu, K. H. Ye, C. S. Hwang, and J.-H. Choi, Atomistic prediction on the composition- and configuration-dependent bandgap of Ga(As, Sb) using cluster expansion and ab initio thermodynamics, Mater. Sci. Eng.: B **280**, 115713 (2022).
[117] K. Yu and E. A. Carter, Elucidating structural disorder and the effects of Cu vacancies on the electronic properties of $Cu_2ZnSnS_4$, Chem. Mater. **28**, 864 (2016).
[118] T. Eom, W. J. Kim, H.-K. Lim, M. H. Han, K. H. Han, E.-K. Lee, S. Lebègue, Y. J. Hwang, B. K. Min, and H. Kim, Cluster expansion method for simulating realistic size of nanoparticle catalysts with an application in $CO_2$ electroreduction, J. Phys. Chem. C **122**, 9245 (2018).
[119] C. Ai, J. H. Chang, A. S. Tygesen, T. Vegge, and H. A. Hansen, Impact of hydrogen concentration for $CO_2$ reduction on $PdH_x$: A combination study of cluster expansion and kinetics analysis, J. Catal. **428**, 115188 (2023).
[120] T. T. Yang, T. L. Tan, and W. A. Saidi, High activity toward the hydrogen evolution reaction on the edges of $MoS_2$-supported platinum nanoclusters using cluster expansion and electrochemical modeling, Chem. Mater. **32**, 1315 (2020).
[121] C. Ai, J. H. Chang, A. S. Tygesen, T. Vegge, and H. A. Hansen, High-throughput compositional screening of $Pd_xTi_{1-x}H_y$ and $Pd_xNb_{1-x}H_y$ hydrides for $CO_2$ reduction, ChemSusChem **17**, e202301277 (2023).
[122] M. K. Y. Chan, J. Reed, D. Donadio, T. Mueller, Y. S. Meng, G. Galli, and G. Ceder, Cluster expansion and optimization of thermal conductivity in SiGe nanowires, Phys. Rev. B **81**, 174303 (2010).
[123] E. B. Isaacs and C. Wolverton, Electronic structure and phase stability of Yb-filled $CoSb_3$ skutterudite thermoelectrics from first-principles, Chem. Mater. **31**, 6154 (2019).
[124] S. B. Maisel, M. Höfler, and S. Müller, Configurationally exhaustive first-principles study of a quaternary superalloy with a vast configuration space, Phys. Rev. B **94**, 014116 (2016).
[125] A. Sharma, P. Singh, D. D. Johnson, P. K. Liaw, and G. Balasubramanian, Atomistic clustering-ordering and high-strain deformation of an $Al_{0.1}CrCoFeNi$ high-entropy alloy, Sci. Rep. **6**, 31028 (2016).
[126] M. Alidoust, D. Kleiven, and J. Akola, Density functional simulations of pressurized Mg-Zn and Al-Zn alloys, Phys. Rev. Mater. **4**, 045002 (2020).
[127] A. R. Natarajan and A. Van der Ven, First-principles investigation of phase stability in the Mg-Sc binary alloy, Phys. Rev. B **95**, 214107 (2017).
[128] F. Eriksson, E. Fransson, and P. Erhart, The hiphive package for the extraction of high-order force constants by machine learning, Adv. Theory Simul. **2**, 1800184 (2019).
[129] W. Huang, A. Urban, Z. Rong, Z. Ding, C. Luo, and G. Ceder, Construction of ground-state preserving sparse lattice models for predictive materials simulations, npj Comput. Mater. **3**, 30 (2017).
[130] P. Zhong, T. Chen, L. Barroso-Luque, F. Xie, and G. Ceder, An $\ell_0\ell_2$-norm regularized regression model for construction of robust cluster expansions in multicomponent systems, Phys. Rev. B **106**, 024203 (2022).
[131] P. M. Larsen, K. W. Jacobsen, and J. Schiøtz, Rich ground-state chemical ordering in nanoparticles: Exact solution of a model for Ag-Au clusters, Phys. Rev. Lett. **120**, 256101 (2018).
[132] W. Huang, D. A. Kitchaev, S. T. Dacek, Z. Rong, A. Urban, S. Cao, C. Luo, and G. Ceder, Finding and proving the exact ground state of a generalized Ising model by convex optimization and max-sat, Phys. Rev. B **94**, 134424 (2016).
[133] D. Kleiven and J. Akola, Precipitate formation in aluminium alloys: Multi-scale modelling approach, Acta Mater. **195**, 123 (2020).
[134] Z. Deng, T. P. Mishra, W. Xie, D. A. Saeed, G. S. Gautam, and P. Canepa, kMCpy: A Python package to simulate transport properties in solids with kinetic Monte Carlo, Comput. Mater. Sci. **229**, 112394 (2023).
[135] Zenodo record https://doi.org/10.5281/zenodo.10997197.
[136] https://ce-tutorials.materialsmodeling.org.
[137] P. E. Blöchl, Projector augmented-wave method, Phys. Rev. B **50**, 17953 (1994).
[138] G. Kresse and D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B **59**, 1758 (1999).
[139] G. Kresse and J. Hafner, Ab initio molecular dynamics for liquid metals, Phys. Rev. B **47**, 558 (1993).
[140] G. Kresse and J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B **54**, 11169 (1996).
[141] G. Kresse and J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci. **6**, 15 (1996).
[142] M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, and B. I. Lundqvist, Van der Waals density functional for general geometries, Phys. Rev. Lett. **92**, 246401 (2004).
[143] K. Berland and P. Hyldgaard, Exchange functional that tests the robustness of the plasmon description of the van der Waals density functional, Phys. Rev. B **89**, 035412 (2014).
[144] J. M. Cowley, An approximate theory of order in alloys, Phys. Rev. **77**, 669 (1950).