# Comparison of Cluster Expansion Fitting Algorithms for Interactions at Surfaces

Laura M. Herder[a], Jason M. Bray[b], William F. Schneider[a,b,1]

[a]*Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, IN, USA*
[b]*Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, IN, USA*

**Abstract**

Cluster expansions (CEs) are Ising-type interaction models that are increasingly used to model interaction and ordering phenomena at surfaces, such as the adsorbate-adsorbate interactions that control coverage-dependent adsorption or surface-vacancy interactions that control surface reconstructions. CEs are typically fit to a limited set of data derived from density functional theory (DFT) calculations. The CE fitting process involves iterative selection of DFT data points to include in a fit set and selection of interaction clusters to include in the CE. Here we compare the performance of three CE fitting algorithms—the MIT Ab-initio Phase Stability code (MAPS, the default in ATAT software), a genetic algorithm (GA), and a steepest descent (SD) algorithm—against synthetic data. The synthetic data is encoded in model Hamiltonians of varying complexity motivated by the observed behavior of atomic adsorbates on a face-centered-cubic transition metal close-packed (111) surface. We compare the performance of the leave-one-out cross-validation score against the true fitting error available from knowledge of the hidden CEs. For these systems, SD achieves lowest overall fitting and prediction error independent of the underlying system complexity. SD also most accurately predicts cluster interaction energies without ignoring or introducing extra interactions into the CE. MAPS achieves good results in fewer iterations, while the GA performs least well for these particular problems.

*Keywords:* cluster expansion, adsorbate interactions, surface ordering, genetic algorithm, MAPS, steepest descent

*corresponding author
*Email address:* wschneider@nd.edu (William F. Schneider)

## 2. Introduction

A cluster expansion (CE) is a lattice-based model Hamiltonian of a multi-component system in which the energy is expressed as a series expansion of site interactions, or "clusters," of increasing order and size [1]. The occupancy of a particular site is specified by a spin variable ($\sigma$) that, within the Ising convention, takes takes a value of $\pm 1$ in a binary system.[2] A particular arrangement of spins is referred to as a configuration, $\boldsymbol{\sigma}$, and the energy of a particular configuration (or structure), $E_{\mathrm{CE}}$, is expanded in polynomials of the spin variables weighted by an effective cluster interaction ($J$), or ECI:

$$E_{\mathrm{CE}}(\boldsymbol{\sigma}) = J_0 + \sum_{i}^{\text{sites}} J_i \sigma_i + \sum_{ij}^{\text{pairs}} J_{ij} \sigma_i \sigma_j + \sum_{ijk}^{\text{triplets}} J_{ijk} \sigma_i \sigma_j \sigma_k + ... \tag{1}$$

The untruncated CE exactly represents the energy of any binary system [1, 3], and the formalism can be expanded to ternaries and beyond [4]. In practical application, a limited number of configurations and associated energies are typically available to parametrize the CE, often from from first-principles methods such as density functional theory (DFT). Once parametrized, the CE energy can be evaluated very rapidly with accuracy approximately equal to the underlying first-principle calculations used in the fit [5]. The CE can be used further to screen for structures having a particular property, to conduct statistical sampling (e.g., via Monte Carlo simulation) to arrive at thermodynamic averages, or even to drive kinetic simulations [6–14].

CEs were originally developed to predict stable structures and phase diagrams of bulk metal alloy and metal oxide systems [15–19]. Recently, surface reconstructions [10, 20], surface alloys [20–22], and adsorbate adlayers [6, 11, 23] have been modeled using two-dimensional CEs [3, 15, 16, 19, 24–31]. For cases of adsorption or surface reconstruction, the spin variable, $\sigma$, signifies the presence or absence of adsorbates [8–10, 20, 32], whereas in surface alloys $\sigma$ indicates an element of one type or another [3, 20–22, 27, 33–38]. It remains unclear whether fitting conventions and procedures originally designed for bulk alloy systems are appropriate for surface adsorbate problems.

Our interest in two-dimensional CEs derives primarily from their relevance to surface adsorbate phenomena in heterogeneous catalysis [9, 24, 39–42]. Non-zero adsorbate-adsorbate interactions influence the adsorbate coverages, adsorption energies, adsorbate orderings, and the finite-temperature distribution of adsorbates at a surface. Adsorbate interactions have been observed to exhibit some regular scaling behavior across different metals [28, 29, 43–45], but in general no universal interaction coefficients apply to all adsorbate systems. Since

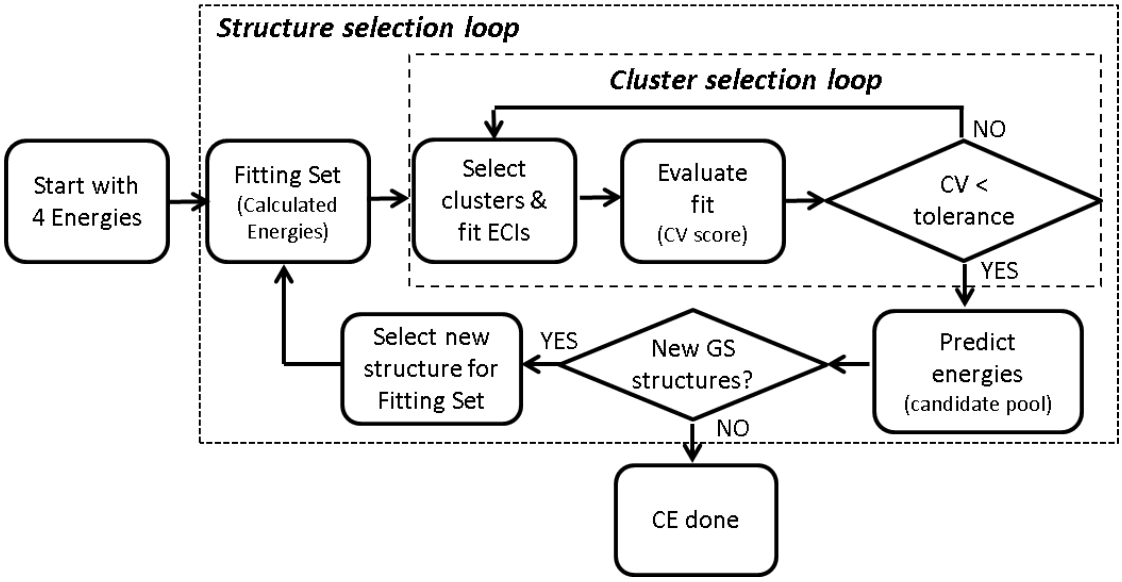


Figure 1: Schematic overview of the cluster expansion (CE) construction process.

ECIs are an *effective* fit to a particular set of data and do not necessarily represent intrinsic material interactions, a new CE must be created for each new system studied [19].

Figure 1 illustrates a typical CE parametrization process, commonly referred to as the Structure Inversion Method (SIM) or the Connolly-Williams method [3, 46, 47]. One begins with an initial set of ordered structures, called the fitting set, of known energies, $E_k$. A predetermined algorithm, represented in the figure by the "cluster selection loop" box and discussed further below, identifies some set of clusters to include in the series expansion, and least-squares regression analysis of the fitting set determines the corresponding ECI values. The predictive error of the CE is evaluated using a leave-one-out cross-validation (CV) score [48]:

$$(CV)^2 = n^{-1} \sum_{k}^{n} (E_k - E'_k)^2 \tag{2}$$

Here $n$ equals the number of structures used to fit the CE, $E_k$ is the known energy and $E'_k$ is the CE-fitted energy of structure $k$ obtained from a fit to the $n-1$ other structures in the fitting set. If the predictive error is acceptable, the chosen clusters are retained and the loop exited; if not, the cluster set is updated using some algorithm and the evaluation repeated.

It is common to start with a small fitting set and to use a series of intermediate CEs to bootstrap the CE construction process, in the "structure selection loop." The CE obtained from the cluster selection loop is used to predict the energies of some large candidate pool of structures, some of which are selected to be evaluated using the true (e.g., DFT) model and added to the fitting set. Common practice in fitting a CE selects potential new ground states, or ordered structures with predicted energies that fall below the minimum energy hull of the current set of fitting structures.[49] The cluster selection loop is then applied to the now enlarged fitting set to create a new CE. Iterations through both loops continue

until the known and predicted formation energy hulls are consistent and the CV score meets some acceptability criterion.

Both first-principles calculations of the fitting set structures and the iterative fitting process of Figure 1 must be carried out efficiently to obtain maximum CE predictive reliability with minimum computational cost. While first-principles calculations are generally the slowest step, the algorithms driving structure selection and cluster selection are arguably more critical to the overall success of the CE. Early CEs relied on intuition to manually select clusters and structures, while recent approaches incorporate automated optimization algorithms removing this human element. Several automated computer codes facilitate this optimization [3, 47, 50–53]. In this work, we examine the merits of several different cluster selection algorithms, including the default "MAPS" algorithm implemented in the Alloy Theoretic Automated Toolkit (ATAT)[3, 34, 35], a stochastic genetic algorithm (GA) [54–58] implemented as an add-on to ATAT, and a "steepest descent" (SD) approach implemented in-house [8, 24].

To circumvent the computational cost of generating extensive, high-quality DFT data sets, we apply these three algorithms to three model Hamiltonians defined in terms of two-dimensional, "hidden" CEs. These hidden CEs include a finite number of prescribed clusters and associated ECI values. The three are constructed to be of increasing level of complexity and to be representative of the interactions of adsorbates at a surface. For each case, we initialize the fitting set with the same four ordered structures, execute the procedure illustrated in Figure 1, and observe the evolution of the generated CEs with each step through the structure selection loop. We evaluate both CV scores within the available fitting set of structures and actual performance against the much larger set of energies easily accessible from the hidden CEs. We show that the SD provides the best overall fit independent of the underlying system complexity and highlight the short-comings of the leave-one out CV score as single criterion for evaluating the quality of a CE.

## 3. Cluster Selection Algorithms

We used the Alloy Theoretic Automated Toolkit (ATAT), version 2.86 [3, 34, 35], to facilitate cluster selection. ATAT by default uses the MIT ab-initio Phase Stability (MAPS) cluster selection algorithm, which favors compact clusters and requires that all the subclusters of a many-body cluster are included in a CE.[59] The genetic algorithm (GA) was implemented as an add-on to ATAT by Dalach et al.[21, 33], although application of the method to cluster expansions in general was pioneered previously [54–58]. Briefly, GA mimics the idea of biological evolution, incorporating a stochastic series of combinations and modifications of several candidate clusters, treating individual clusters as building blocks for the larger CE "genomes," and using survival of the fittest to optimize the CV score.

The steepest descent (SD) cluster selection algorithm is implemented in-house by wrapping the standard tools available in ATAT within a series of loops to systematically change the clusters in the CE, evaluate the CV score, and keep track of clusters needed for an optimal CE. While conceptually simple, this approach has, to our knowledge, never been

reported for cluster selection prior to its development for generation of a Pt(321)−O CE [24, 32] and, in a slightly modified form, a Pt(111)−O CE [8].

SD adds or removes clusters from an extensive list of candidates one at a time to seek the greatest decrease in CV score. Once the CV score reaches a (local) minimum, the cluster selection is considered complete. With no analytical function describing the gradient of the CV score with respect to inclusion of specific clusters, the steepest descent move is determined simply by testing the CV score for every possible addition (or removal) of a single cluster to (or from) the existing CE. The description of this algorithm as "steepest descent" is somewhat unconventional, as the steps are determined by discrete changes in cluster number and type rather than from a continuous, analytical gradient. This brute force method of testing every possible cluster is more amenable to two-dimensional problems that have a relatively small number of possible clusters. In theory, restricting changes to individual clusters could prevent SD from finding a global minimum in CV score, but in practice the single-cluster-swap is capable of reaching a very low CV score, and this restriction is implemented primarily for computational efficiency. In contrast to GA or other stochastic approaches, SD is completely deterministic and always provides the same optimized CE for a given starting CE and list of candidate clusters. For consistency, the default starting CE in this work includes the empty and point clusters (in multi-site CE's, one point cluster would be included for each unique site).

Prior to executing SD, a candidate cluster list must be generated and should contain an adequate number of cluster types extending to a sufficient size in order to minimize sensitivity of the results to this list. In this work, a pool of 114 candidate clusters including pair, triplet and 4-body terms was generated. SD steps through two nested loops to select clusters. The outer loop tracks CV score, ensuring that it decreases with every step, and ends when the CV score no longer decreases (reaching a local minimum) or the entire list of candidate clusters is exhausted. The inner loop starts with the current CE and cycles every cluster in the candidate list on (or off) one at at time, calculating the CV score each time. After testing all clusters, the change resulting in the lowest CV score is retained. Clusters are not required to remain part of the final CE once added, rather clusters may be removed on subsequent inner loop iterations if this yields the best improvement to the CV score.

## 4. Results

### *4.1. Model Hamiltonians*

We consider a triangular lattice inspired by the (111) facet of a hexagonal close-packed face-centered cubic metal. The energy of a structure $\boldsymbol{\sigma}$ is expressed as a per site formation energy, $E_F$, relative to the fully vacant and the fully occupied structures:

$$E_F(\boldsymbol{\sigma}) = E(\boldsymbol{\sigma}) - (1-\theta)E_{\theta=0} - \theta E_{\theta=1} \tag{3}$$

Here $\theta(\boldsymbol{\sigma}) = \sum_i(\sigma_i+1)/N$ equals the fraction of occupied sites in structure $\boldsymbol{\sigma}$ that contains $N$ sites. We define a site as the fcc 3-fold hollow site on the (111) surface (shown in Figure 2), and full coverage to be one mono-layer (1 ML).

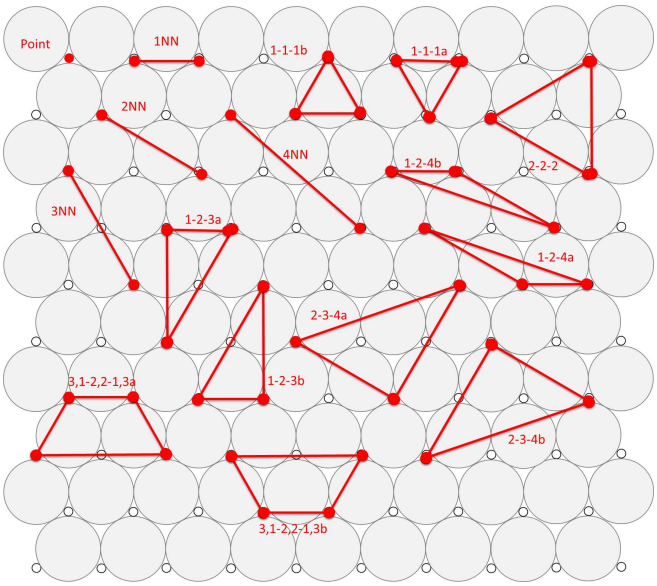


Figure 2: Clusters included in the hidden CEs.

We constructed three hidden CEs of increasing degrees of complexity. Figure 2 shows the clusters used including empty and point terms, four two-body clusters, eight three-body clusters, and two four-body clusters. Assigned ECI values for each cluster are in Table 1. The simplest hidden CE, CE-A, includes only two-body terms up to third-nearest-neighbor; CE-A corresponds to a weakly binding adsorbate system with weak interactions. The intermediate hidden CE, CE-B, adds the fourth-nearest-neighbor and five three-body terms. CE-B is more representative of behavior typical of O and similar adsorbates on a late transition metal [13]. CE-C includes non-zero ECIs for all clusters shown in Figure 2. CE-C does not correspond to a particular adsorbate system but tests fitting behavior in an extreme case of interaction. ECI magnitudes were varied between the hidden CEs ensuring fitting errors were independent of ECI magnitude. To avoid an unfair bias against the MAPS algorithm, all subclusters of a given cluster are included in all the hidden CEs.

Formation energies of 23,468 structures (up to 15 adsorption sites per unit cell) were calculated. We assume that these structures contain all relevant information about the underlying hidden CEs and may be referred to as the "full set," defining the "true" ground states and serving as the definitive ruler for evaluating predictive accuracy of any trial CEs. A subset of 3,164 structures (up to 12 adsorption sites per unit cell) served as the candidate pool available for CE construction, and we confirmed the inclusion of all ground state structures in this subset. The small size of the candidate pool reduced the computational cost of CE construction and reserved external structures for calculating predictive error. These set sizes were chosen to be consistent with sizes typical in fitting a DFT-based adsorbate system. Figure 3 shows the relationship between the full set and candidate pool as well as the fitting set, a further subset that grows with each iteration of the structure selection loop.

Formation energies (eq. 3) of candidate pool structures are plotted vs. adsorbate coverage in Figure 4. The pairwise-only CE-A hidden CE is necessarily symmetric about the 50%

Table 1: Effective Cluster Interaction ($J$) values of the clusters shown in Figure 2 for each of the hidden CEs

| Cluster | CE-A | CE-B | CE-C |
|---|---|---|---|
| Clean | −0.29 | −0.51 | −9.00 |
| Point | 0.00 | −0.15 | −4.80 |
| 1NN | 0.06 | 0.10 | 0.90 |
| 2NN | 0.03 | 0.05 | 0.50 |
| 3NN | 0.01 | 0.01 | 0.40 |
| 4NN | . | 0.01 | 0.20 |
| 1-1-1a | . | 0.08 | 0.80 |
| 1-1-1b | . | 0.03 | 0.54 |
| 1-2-3a | . | 0.01 | 0.32 |
| 1-2-3b | . | . | 0.21 |
| 1-2-4a | . | . | 0.09 |
| 1-2-4b | . | . | 0.08 |
| 2-2-2 | . | . | 0.06 |
| 2-3-4a | . | . | 0.03 |
| 2-3-4b | . | . | 0.02 |
| 3,1-2,2-1,3a | . | . | 0.60 |
| 3,1-2,3-1,3b | . | . | 0.30 |

coverage point, and inclusion of higher-order clusters breaks the symmetry, evident in CE-B and CE-C. These constructed energy hulls are similar in nature to literature reports of adsorbate interactions on a (111) metal surface [8, 21, 25, 60–62].

*4.2. Iterative CE Construction*

We initialized the fitting sets for each algorithm and each hidden CE using the 0.5 ML p(2×1), 0.25 ML p(2×2) structures and reference the empty (0 ML) and full (1 ML) lattices with an energy of zero. For MAPS and GA, ATAT was allowed to select the next structure added for each iteration of the structure selection loop using an internal algorithm, and for SD, the structure with a predicted energy farthest below the ground state energy hull was added. In all cases, the structure selection loop was terminated once no predicted energies fell below the known ground state formation energy hull of the hidden CE.

Besides the computational benefit of using contrived CEs instead of DFT, knowledge of the underlying CE also allows the actual predictive error of all energies in the full set to be assessed. After each iteration of the structure selection loop, the root mean square error between the trial CE and hidden CE is calculated:

$$\%\ \text{Fitting Error} = \sqrt{M^{-1}\sum_{k}^{M}\left(\frac{E_k - E'_k}{E_k}\right)^2} \times 100\% \tag{4}$$

The sum includes $M = 23,468$ structures in the full set with known formation energies $E_k$ and predicted energies $E'_k$ for structure $k$. This percent fitting error is plotted vs. structure

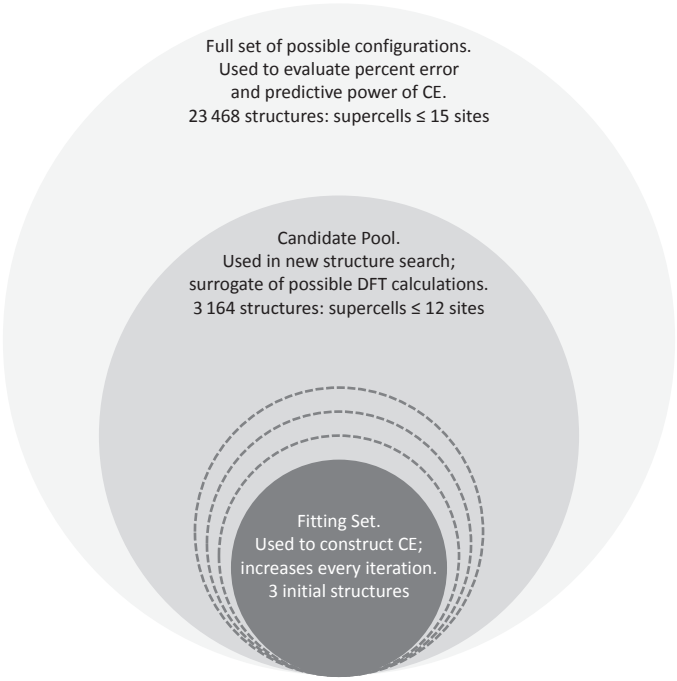


Figure 3: Schematic relationship between structure sets.

selection loop iteration in Figure 5. Although not a requirement for CE completion, MAPS, GA, and SD predicted the correct ground state structures for all hidden CEs.

Figure 5(a) compares the number of clusters vs. iterations/size of fitting set for the three methods applied to CE-A, which contains four non-zero ECI. The MAPS method, shown as red circles, explores CEs of various sizes as structures are added to the fitting set before finally settling on four clusters after 38 iterations. As shown in Figure 5(b) the CV score decreases to zero after only a few iterations. Subsequent iterations are spent refining the fit coefficients to capture all the ground states. As shown in Figure 5(c) the actual fitting error also quickly attains a small value. In comparison, the GA, shown in Figure 5 as green diamonds, uses only 18 iterations to identify the four non-zero clusters. The GA CV score drops to almost zero by the third iteration and never increases. In contrast, the GA fitting error sits above 10% through 10 iterations, beyond which it decreases to less than 1%. The SD results are shown as blue squares in Figure 5. SD explores CEs of various sizes up to 14 terms across the course of 32 iterations before arriving at a CE that predicted no energies below the formation hull and included all four non-zero clusters. As with the other two methods, the CV score achieves a small value after only a handful of iterations and the percent fitting error declines to less than 1% by the eleventh iteration. In all three fitting methods the fitting error never vanished completely due to small numerical errors in the fitted ECIs. The ultimate CV score across all three methods was on the order of $10^{-6}$ eV/site, negligibly small.

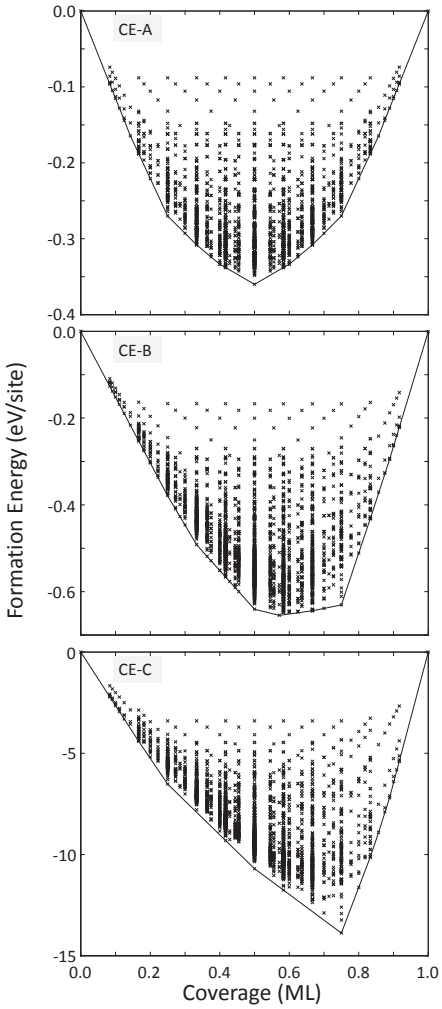


Figure 4: Formation energies vs. coverage for the three hidden CE's. Points are the formation energies of the 23,468 structures included in the full set (up to 15 adsorption sites per unit cell). Black lines represents the hull of lowest total formation energies at a given coverage.

The top three sub-panels in Figure 6 plot the predicted ECIs from each fit of CE-A against the actual hidden values (Table 1). Correspondence between hidden and predicted values fall along the diagonal. Because the trial CEs are not restricted to use the same clusters present in the hidden CEs, the respective methods occasionally incorporate additional clusters or leave out the original clusters. Such cases are represented in the figure by ECIs with a value of zero. As listed in Table 1, the point cluster for CE-A is zero. The MAPS, GA, and SD algorithms ultimately recover the same clusters, include no extraneous clusters, and recover ECI values the same as the hidden model within numerical precision. However, each algorithm did incorporate other clusters at intermediate steps along the cluster selection process.

The same general pattern is seen when fitting the hidden CE-B with MAPS. As shown in red circles in Figure 5(d), MAPS discovers eight of the nine non-zero clusters within 8

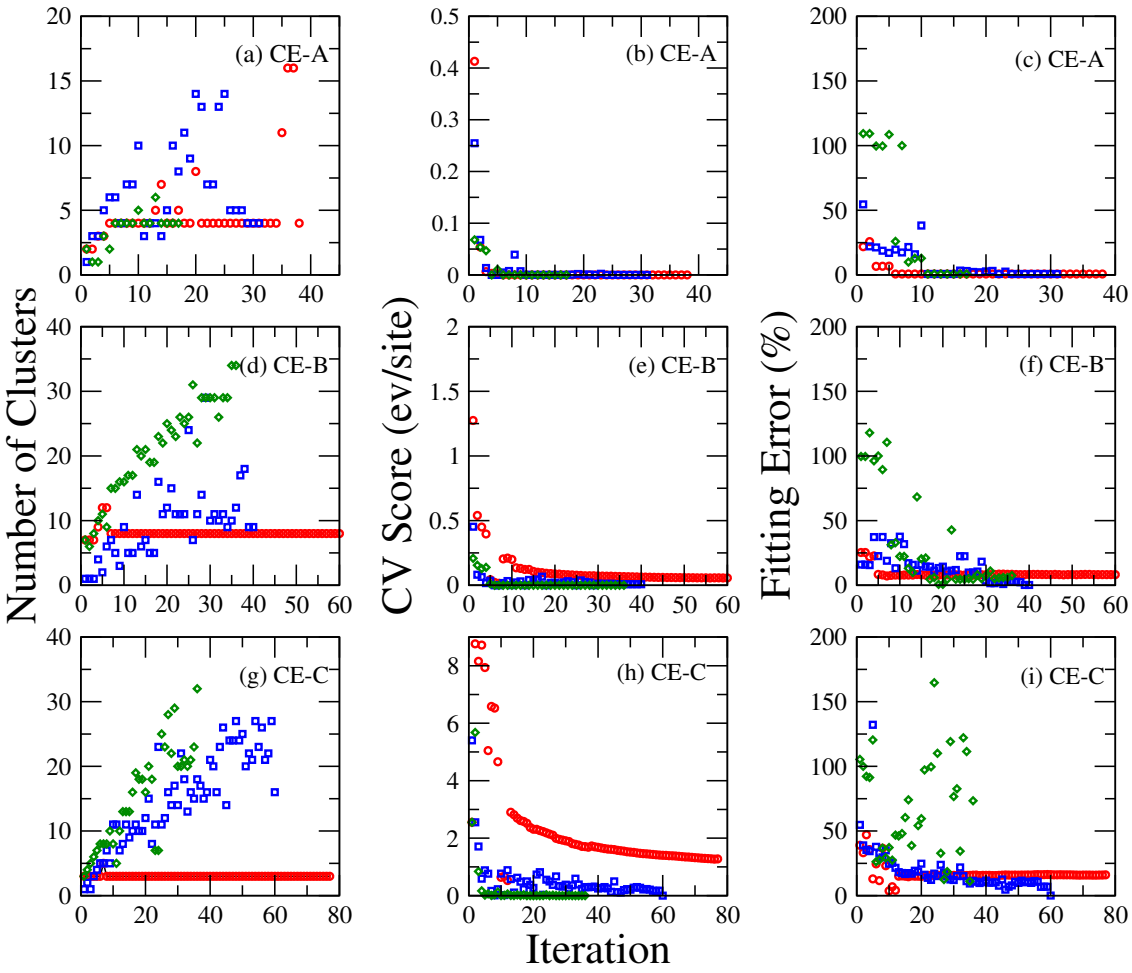


Figure 5: Left: Number of clusters in trial CE vs. iteration. Center: Cross-validation score vs. iteration. Right: Percent fitting error vs. number of iterations in the structure selection loop. Results are shown for each fitting method (blue squares = SD, red circles = MAPS, green diamonds = GA) for the three known CEs of varying complexity (from top to bottom: CE-A, CE-B, and CE-C).

iterations but continues to predict structures below the energy hull for another 50 iterations. The MAPS CV score (Figure 5(e)) declines steeply to about 0.2 eV/site in the first 8 iterations and declines slowly after that, ultimately reaching $10^{-2}$ eV/site. The fitting error (Figure 5(f)) follows a similar pattern, reaching about 8% after 8 iterations and not declining beyond. The GA, shown in green diamonds, ceases to predict new ground states after only 37 iterations and ended with a CE containing 34 clusters. The corresponding CV score declines to a very small value after only a few iterations; in contrast, the fitting error is much more substantial and variable with iteration, finally attaining a value of 8%. The SD algorithm (blue squares) continued 41 iterations before ceasing to predict new ground states and ended with the same nine clusters as in the hidden CE-B. The CV score, as with the GA, attains quite small values after only a handful of iterations. The corresponding fitting error declined more slowly but ultimately reached approximately 1%, the lowest of the three

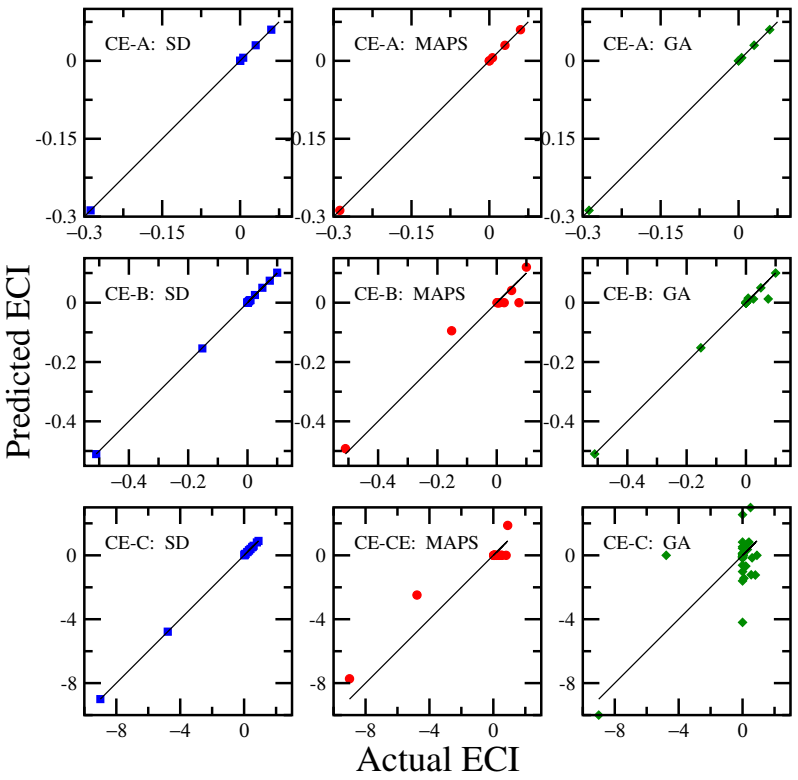


Figure 6: Predicted vs. known effective cluster interaction coefficients for each hidden CE. Results are shown for only the final iteration in each fitting process (blue circles = SD, red squares = MAPS, green diamonds = GA) for all three hidden CEs (from top to bottom: CE-A, CE-B, CE-C).

methods for CE-B.

On the final iteration for CE-B, SD reproduced the correct clusters and corresponding ECI values. MAPS missed one cluster and was therefore unable to correctly predict the other ECI values. GA also missed one cluster, but included twice as many clusters as the original CE-B. However, most of the extraneous clusters had ECI values near zero and are therefore not discerned in Figure 6.

The differences between the MAPS, GA, and SD algorithms are most evident in the more extreme CE-C. As shown across the bottom row of Figure 5(g), MAPS used only three clusters through the entire course of its 78 iterations. The CV score declined gradually as more structures were added to its database (Figure 5(h)). The corresponding fitting error (Figure 5(i)) reached low values after about twelve iterations and rises to a higher, constant value when more structures were added.

Table 2 reports the ultimate values of CV score and percent fitting error at the final iteration of CE-C. We performed multiple trials with the stochastic GA algorithm and recovered similar results in each case; results for a representative run are reported here. The GA ceased predicting new ground states after 36 iterations. GA again tends to be generous in including extra clusters beyond those in the hidden CE-C. The CV score from the GA is smaller than from MAPS and converges to a very small value after only a handful

of iterations. Fitting errors were however quite substantial with GA. The SD algorithm stopped after 59 iterations. As shown in Figure 6, SD was the only one of the three methods to recover the same clusters as in the underlying model. Both CV score and fitting error decreased in a more regular fashion across iteration.

Overall, CV scores tended to be larger as the underlying CE contained more and larger terms. There was little correlation between CV score and the fitting error extracted from structures not used in the fit. The data extracted from the final iterations in fitting CE-C shown in Table 2 illustrates this absence of correspondence between percent fitting error and CV score.

Table 2: Percent fitting error and CV scores for the three different fitting methods applied to the last iteration of CE-C.

| Method | % Fitting Error | CV score (eV/site) | Iterations |
|---|---|---|---|
| SD | 1% | $10^{-3}$ | 59 |
| MAPS | 15% | 1.27 | 78 |
| GA | 80% | $10^{-6}$ | 36 |

The three algorithms selected new structures in a different order and did not necessarily include the same structures in their final fitting sets. In any given fit, up to 50% of the included structures might be different from those included using a different algorithm. To determine whether the superior performance of the SD in reproducing the hidden CE arose from a fortuitous compilation of structures in its fitting set, we used the identical fitting set to re-fit clusters and ECI values using MAPS and GA to the final SD fitting set for CE-C. MAPS had a CV score of 1.7 and had a fitting error of 10.1% whereas GA's CV score was on the order of $10^{-5}$ and had a fitting error of 18.0%.

The percent fitting error represents an average of all structures in the full set but could be dominated by certain regions of configuration space. To explore the distribution of this error as a function of coverage, we report in Figure 7 the difference between known and CE-predicted energy vs. the coverage (in ML) of all structures in the full set. Results are shown from the final iterations on CE-C. Small differences in real and predicted energies fall along the $y = 0$ line. The SD error is small across coverages. GA errors are largest and generally coverage independent. MAPS errors are intermediate and, for unknown reasons, tend to be negative at low coverage and positive at high. Results are similar but less dramatic for CE-B. Errors are small for all methods within CE-A.

## 5. Discussion

In this work, we compared several cluster selection algorithms with respect to (a) their ability to identify CEs that satisfactorily reproduce the energies of an underlying hidden Hamiltonian, (b) their efficiency in identifying those CEs, and (c) their ability to recover the functional form of the Hamiltonian. While correct ECI values and accurate prediction energies are not completely independent, we would be satisfied with accurate energy pre-

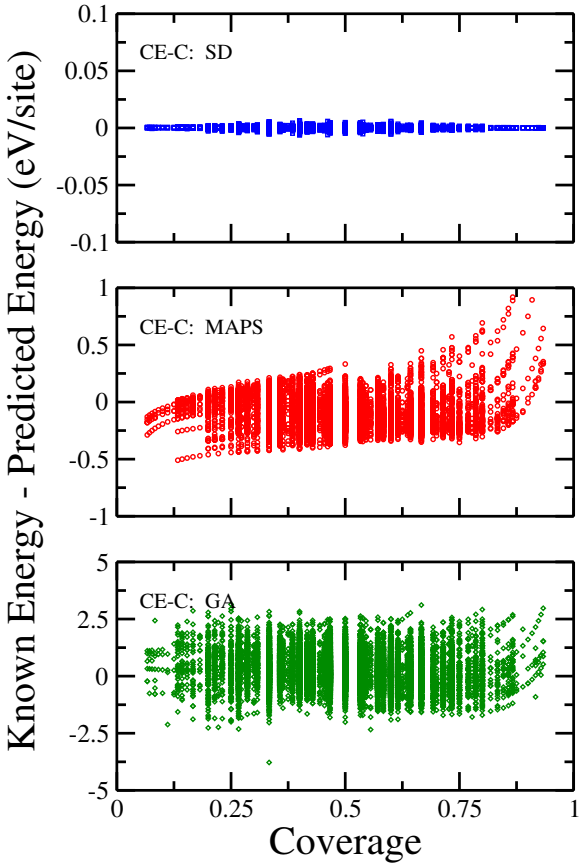


Figure 7: Coverage dependent error of the known formation energies minus predicted formation energies for all configurations in the full set. Data is from the last iteration of the fit to CE-C; from top to bottom: SD, MAPS, GA. Note different scales on y axes.

dictions regardless of the CE used. However, in this work we were unable to find a case of accurate predictions without a correct choice of clusters and ECI values.

SD uniformly attains the lowest fitting error and is most successful in recovering the hidden CEs. It also provides the most uniform errors vs. coverage. It does so at the expense of a relatively larger number of iterations through the structure selection loop. As implemented here, SD only considered clusters in a user-generated list of candidates, thereby making its results potentially user dependent. The SD approach could be susceptible to over-fitting, especially if both the candidate cluster list and the fitting set are large. Here the hidden CEs were not truly unknown to the user, so that we could choose an appropriately sized cluster candidate list. Alternatively, if the candidate cluster list missed critical clusters, the SD would not be able to recover an optimal CE. In spite of these shortcomings, SD gave the lowest and most uniform fitting error.

Fitting errors from the MAPS algorithm were generally larger, but MAPS reaches a constant fitting error within the fewest structure selection steps. The MAPS fit used a slightly different structure selection algorithm than did the SD fit, and this difference may

have contributed to the faster convergence of MAPS. In limited tests on CE-C, though, the SD-selected fit structures was superior to the MAPS-selected set. The MAPS algorithm requires that all sub-clusters of a give cluster be included in the CE. This constraint likely both helps MAPS to converge quickly to a reasonable CE and fitting error and hinders its ability to ultimately minimize the fitting error. The most complex CE-C contains many contributions but is dominated by a few large ECIs. Once MAPS (and GA) find these two or three important clusters, it struggles to discern the remaining contributions to the underlying CE. SD seems not to suffer from this problem for reasons that are unclear.

The GA algorithm typically produced the largest fitting errors even as its CV score was low. The GA approach is the least constrained and more clusters of increasing size will continue to be added to the CE indefinitely. Thus for CE-C, the GA approach led to severe over-fitting of the fitting set. Extra clusters were added to the CE, providing an excellent fit to the particular structures in the fitting set and a low CV score, but these clusters were inappropriate for describing the entire configuration space of the full set and resulted in large fitting errors. Of the three algorithms, the GA as applied here provided the worst trade-off between speed and quality of fit. We did run multiple replicates of the GA against the hidden CEs to ensure no bias in any particular fit. Of course these results should be generalized with caution, as they could depend on the optimization problem and the particular GA implementation.

As seen in all cases, the CV score was a poor measure of predictive accuracy. Given an essentially unlimited number of cluster terms in the CE, many different CEs could describe the fitting set equally well but grossly mispredict energies outside the fitting set. While over-fitting can mislead by causing unrealistically small CV scores, CV scores may also be deceptive in the opposite extreme when one or two structures fit very poorly and lead to an inflated CV score that is also not representative of the actual CE quality. Therefore measuring prediction error against an independent set of known energies (i.e., a testing set) is important. The testing set here included over 20,000 structures, but smaller sets of 20 or 30 randomly selected structures would provide a better error estimate than the CV score alone.

The computational cost of MAPS, GA, and SD scale differently with the size of the hidden system. MAPS attempts to test every possible CE within the size constraints (i.e., the number of clusters cannot exceed the size of the fitting set). It is trivial to test every possibility when fitting sets are small, but as the fitting set grows large, the computational cost of the MAPS fitting grows rapidly. GA never "finishes" fitting; rather the user terminates the fit. Although simple systems reach an optimum quickly, for more complicated systems the GA introduces and tests numerous, higher-order clusters that only marginal improve the CE. Additionally, because the GA is stochastic, it should be executed several times to ensure it does not lead to a local minimum or other suboptimal solutions. SD scales linearly with the number of candidate clusters (user-specified) and number of steps required to minimize the CV score. Every step of the SD optimization takes approximately the same amount of time, unlike MAPS and GA which slow dramatically after the initial iterations. The single-cluster-swap implementation keeps the code both simple and computationally efficient, but it does not guarantee a global minimum. Yet allowing the swapping of even

two clusters at a time would change the code to scale as $N^2$ with the number of candidate clusters. In general, results of the single-pass algorithm more than adequately describe the fitting set, so the marginal improvement in quality of the CE or the potential for reaching the final CE in fewer steps would not make up for the additional computational time of each step.

Finally, we note that recently "compressive sensing" [63, 64] has shown promise as an efficient and unbiased approach to structure and cluster selection. To our knowledge the approach has yet to be applied to two-dimensional problems.

## 6. Conclusion

The cluster expansion is a useful approach for efficient modeling of problems involving non-ideal mixing at a surface, such as those involving adsorbate-adsorbate interactions. The quality of a CE depends both on the data that it is fit to and the algorithm used to identify terms to include in the CE. In real applications in which the fit is to a limited amount of data typically generated by expensive DFT calculations, it is difficult to test the performance of algorithms. Here we circumvent this problem by testing three different CE fitting algorithms against three different hidden CEs of increasing complexity. Performance depends on system details and complexity and becomes a trade-off between accuracy and computational cost. For the three contrived systems here, representative of atomic adsorbates on close-packed transition metal surfaces, the steepest descent (SD) method is the preferred CE fitting method. SD provided the lowest overall fitting error independent of the underlying system complexity. SD most accurately reproduced the contrived hidden clusters and effective interaction coefficients as well without including spurious clusters. The computational cost of SD scales predictably unlike ATAT default structure inversion method, MAPS, and ATAT add-on genetic algorithm, GA. Results from all tests show that the leave-one-out CV score is a poor indicator of predictive accuracy. Comparisons against an independent test set are necessary to quantify error.

## 7. Acknowledgments

We acknowledge funding provided by U.S. Department of Energy, Office of Basic Energy Sciences, under grant DE-FG02-06ER15830. LH acknowledges support from an Eiler's Graduate Fellowship from the Center for Sustainable Energy at Notre Dame. Computing resources and technical support were provided by the Notre Dame Center for Research Computing. We also thank Dr. David Schmidt and Dr. Kurt Frey for helpful advice and discussions.

## 8. References


[1] J. Sanchez, F. Ducastelle, D. Gratias, Generalized cluster description of multicomponent systems, Physica A: Statistical Mechanics (1984) 334–350.

[2] E. Ising, Beitrag zur Theorie des Ferromagnetismus, Zeitschrift fur Physik 31 (1) (1925) 253–258.

[3] A. van de Walle, G. Ceder, Automating first-principles phase diagram calculations, Journal of Phase Equilibria 23 (4) (2002) 348–359.

[4] C. Wolverton, D. de Fontaine, Cluster expansions of alloy energetics in ternary intermetallics, Phys. Rev. B 49 (1994) 8627–8642. doi:10.1103/PhysRevB.49.8627.
URL http://link.aps.org/doi/10.1103/PhysRevB.49.8627

[5] A. V. Ruban, I. A. Abrikosov, Configurational thermodynamics of alloys from first principles: effective cluster interactions, Reports on Progress in Physics 71 (4) (2008) 046501.

[6] M. Stamatakis, D. G. Vlachos, Unraveling the complexity of catalytic reactions via kinetic Monte Carlo simulation: current status and frontiers, ACS Catalysis 2 (12) (2012) 2648–2663.

[7] M. Stamatakis, Y. Chen, D. G. Vlachos, First-principles-based kinetic Monte Carlo simulation of the structure sensitivity of the water-gas shift reaction on platinum surfaces, The Journal of Physical Chemistry C 115 (2011) 24750–24762.

[8] D. J. Schmidt, W. Chen, C. Wolverton, W. F. Schneider, Performance of cluster expansions of coverage-dependent adsorption of atomic oxygen on Pt(111), Journal of Chemical Theory and Computation 8 (1) (2012) 264–273.

[9] C. Wu, D. J. Schmidt, C. Wolverton, W. F. Schneider, Accurate coverage-dependence incorporated into first-principles kinetic models: Catalytic NO oxidation on Pt (111), Journal of Catalysis 286 (2) (2012) 88–94.

[10] W. Chen, D. J. Schmidt, W. F. Schneider, C. Wolverton, Ordering and oxygen adsorption in AuPt/Pt (111) surface alloys, The Journal of Physical Chemistry C 115 (2011) 17915–17924.

[11] J. Nielsen, M. D'Avezac, J. Hetherington, M. Stamatakis, Parallel kinetic Monte Carlo simulation framework incorporating accurate models of adsorbate lateral interactions., The Journal of chemical physics 139 (22) (2013) 224706.

[12] D.-J. Liu, J. W. Evans, Realistic multisite lattice-gas modeling and KMC simulation of catalytic surface reactions: Kinetics and multiscale spatial behavior for CO-oxidation on metal (100) surfaces, Progress in Surface Science 88 (4) (2013) 393–521. doi:10.1016/j.progsurf.2013.10.001.
URL http://linkinghub.elsevier.com/retrieve/pii/S0079681613000361

[13] K. Frey, D. J. Schmidt, C. Wolverton, W. F. Schneider, Implications of coverage-dependent o adsorption for catalytic no oxidation on the late transition metals, Catal. Sci. Technol. (2014) in press.

[14] M. Stamatakis, Kinetic Modelling of Heterogeneous Catalytic Systems, Journal of Physics: Condensed Matter 27 (2015) 013001.

[15] G. Ceder, A derivation of the Ising model for the computation of phase diagrams, Computational Materials Science 1 (2) (1993) 144–150.

[16] G. Ceder, G. Garbulsky, D. Avis, K. Fukuda, Ground states of a ternary fcc lattice model with nearest- and next-nearest-neighbor interactions, Physical Review B 49 (1).

[17] R. McCormack, M. Asta, D. de Fontaine, G. Garbulsky, G. Ceder, hcp Ising cluster-variation approximations, Physical Review B 48 (10) (1993) 6767.

[18] T. Mueller, G. Ceder, Bayesian approach to cluster expansions, Physical Review B 80 (2) (2009) 024103.

[19] T. Mueller, G. Ceder, Exact expressions for structure selection in cluster expansions, Physical Review B 82 (18) (2010) 184107.

[20] W. Chen, D. J. Schmidt, W. F. Schneider, C. Wolverton, First-principles cluster expansion study of missing-row reconstructions of fcc (110) surfaces, Physical Review B 83 (7) (2011) 075415.

[21] W. Chen, P. Dalach, W. F. Schneider, C. Wolverton, Interplay between subsurface ordering, surface segregation, and adsorption on Pt-Ti(111) near-surface alloys., Langmuir : the ACS journal of surfaces and colloids 28 (10) (2012) 4683–93.

[22] J. F. Weaver, J.-J. Chen, A. L. Gerrard, Oxidation of Pt(111) by gas-phase oxygen atoms, Surface Science 592 (1-3) (2005) 83–103.

[23] P. Valentini, T. E. Schwartzentruber, I. Cozmuta, ReaxFF Grand Canonical Monte Carlo simulation of adsorption and dissociation of oxygen on platinum (111), Surface Science 605 (23-24) (2011) 1941–1950.
[24] J. M. Bray, J. L. Smith, W. F. Schneider, Coverage-Dependent adsorption at a low symmetry surface: DFT and statistical analysis of oxygen chemistry on kinked Pt(321), Topics in Catalysis 57 (1-4) (2013) 89–105.
[25] H. Tang, A. Van der Ven, B. L. Trout, A. V. D. Ven, Lateral interactions between oxygen atoms adsorbed on platinum (111) by first principles, Molecular Physics 102 (3) (2004) 273–279.
[26] H. Tang, A. Van der Ven, B. L. Trout, A. V. D. Ven, Phase diagram of oxygen adsorbed on platinum (111) by first-principles investigation, Physical Review B 102 (3) (2004) 045420.
[27] R. Riedinger, H. Dreysse, G. Ceder, Electronic structure of disordered alloy described with a reduced set of configurations, Solid state 80 (7) (1991) 489–492.
[28] S. D. J. Miller, J. R. Kitchin, Relating the coverage dependence of oxygen adsorption on Au and Pt fcc(111) surfaces through adsorbate-induced surface electronic structure effects, Surface Science 603 (5) (2009) 794–801.
[29] S. D. J. Miller, N. G. Inoglu, J. R. Kitchin, Configurational correlations in the coverage dependent adsorption energies of oxygen atoms on late transition metal fcc(111) surfaces., The Journal of chemical physics 134 (10) (2011) 104709.
[30] J. R. Kitchin, Correlations in coverage-dependent atomic adsorption energies on Pd(111), Physical Review B 79 (20) (2009) 205412.
[31] D.-J. Liu, J. W. Evans, Interactions between oxygen atoms on Pt(100): implications for ordering during chemisorption and catalysis., Chemphyschem : a European journal of chemical physics and physical chemistry 11 (10) (2010) 2174–81.
[32] J. M. Bray, W. F. Schneider, First-principles thermodynamic models in heterogeneous catalysis, in: A. Asthagiri, M. J. Janik (Eds.), Computational Catalysis, The Royal Society of Chemistry, 2014, pp. 59–115.
[33] P. Dalach, D. E. Ellis, A. van de Walle, First-principles thermodynamic modeling of atomic ordering in yttria-stabilized zirconia, Physical Review B 82 (14) (2010) 144117.
[34] A. van de Walle, Multicomponent multisublattice alloys, nonconfigurational entropy and other additions to the Alloy Theoretic Automated Toolkit, Calphad 33 (2) (2009) 266–278.
[35] E. Cockayne, A. van de Walle, Building effective models from sparse but precise data: Application to an alloy cluster expansion model, Physical Review B 81 (2010) 012104.
[36] H. Y. Geng, M. H. F. Sluiter, N. X. Chen, Hybrid cluster expansions for local structural relaxations, Physical Review B 73 (1) (2006) 012202.
[37] C. Wolverton, M. Asta, H. Dreysse, D. D. Fontaine, Effective cluster interactions from cluster-variation formalism. II, Physical Review B 44 (10) (1991) 4914.
[38] M. Asta, C. Wolverton, D. D. Fontaine, H. Dreysse, Effective cluster interactions from cluster-variation formalism. I, Physical Review B 44 (10) (1991) 4907.
[39] J.-S. McEwen, J. M. Bray, C. Wu, W. F. Schneider, How low can you go? Minimum energy pathways for O(2) dissociation on Pt(111)., Physical chemistry chemical physics : PCCP 14 (2012) 16677–16685.
[40] S. Ovesson, B. Lundqvist, W. F. Schneider, A. Bogicevic, NO oxidation properties of Pt(111) revealed by ab initio kinetic simulations, Physical Review B 71 (11) (2005) 115406.
[41] R. B. Getman, W. F. Schneider, DFT-Based coverage-dependent model of Pt-catalyzed NO oxidation, ChemCatChem 2 (11) (2010) 1450–1460.
[42] P. Deshlahra, J. Conway, E. E. Wolf, W. F. Schneider, Influence of dipole-dipole interactions on coverage-dependent adsorption: CO and NO on Pt(111)., Langmuir : the ACS journal of surfaces and colloids 28 (22) (2012) 8408–17.
[43] N. Inoglu, J. R. Kitchin, Simple model explaining and predicting coverage-dependent atomic adsorption energies on transition metal surfaces, Physical Review B 82 (4) (2010) 045414.
[44] F. Calle-Vallejo, N. G. Inoglu, H.-Y. Su, J. I. Martínez, I. C. Man, M. T. M. Koper, J. R. Kitchin, J. Rossmeisl, Number of outer electrons as descriptor for adsorption processes on transition metals and their oxides, Chemical Science 4 (3) (2013) 1245.

[45] Z. Xu, J. Kitchin, Probing the Coverage Dependence of Site and Adsorbate Configurational Correlations on (111) Surfaces of Late Transition Metals, The Journal of Physical Chemistry C (111) (2014) 25597–25601.
[46] J. W. D. Connolly, A. R. Williams, Density-functional theory applied to phase transformation in transition-metal alloys, Physical Review B 27 (8) (1983) 5169–5172.
[47] J. Sanchez, Cluster expansions and the configurational energy of alloys, Physical review B 48 (18) (1993) 13–15.
[48] P. Zhang, Model selection via multifold cross validation, The Annals of Statistics 21 (1) (1993) 299–313.
[49] D. Morgan, G. Ceder, S. Curtarolo, High-throughput and data mining with ab initio methods, Measurement Science and Technology 16 (1) (2005) 296–301.
[50] D. Lerch, O. Wieckhorst, G. L. W. Hart, R. W. Forcade, S. Müller, UNCLE: a code for constructing cluster expansions for arbitrary lattices with minimal user-input, Modelling and Simulation in Materials Science and Engineering 17 (5) (2009) 055003.
[51] A. Seko, Y. Koyama, I. Tanaka, Cluster expansion method for multicomponent systems based on optimal selection of structures for density-functional theory calculations, Physical Review B 80 (16) (2009) 165122.
[52] A. Seko, I. Tanaka, Grouping of structures for cluster expansion of multicomponent systems with controlled accuracy, Physical Review B 83 (22) (2011) 224111.
[53] S. Curtarolo, W. Setyawan, G. L. W. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, D. Morgan, AFLOW: An automatic framework for high-throughput materials discovery, Computational Materials Science 58 (2012) 218–226.
[54] V. Blum, G. L. W. Hart, M. J. Walorski, A. Zunger, Using genetic algorithms to map first-principles results to model Hamiltonians: Application to the generalized Ising model for alloys, Physical Review B 72 (16) (2005) 1–13.
[55] G. L. W. Hart, R. W. Forcade, Algorithm for generating derivative structures, Physical Review B 77 (22) (2008) 224115.
[56] G. L. W. Hart, V. Blum, M. J. Walorski, A. Zunger, Evolutionary approach for determining first-principles hamiltonians., Nature materials 4 (5) (2005) 391–4.
[57] R. Drautz, A. Díaz-Ortiz, Obtaining cluster expansion coefficients in ab initio thermodynamics of multicomponent lattice-gas systems, Physical Review B 73 (22) (2006) 224207.
[58] D. Lerch, O. Wieckhorst, L. Hammer, K. Heinz, S. Müller, Adsorbate cluster expansion for an arbitrary number of inequivalent sites, Physical Review B 78 (12) (2008) 121405.
[59] N. Zarkevich, D. Johnson, Reliable First-Principles Alloy Thermodynamics via Truncated Cluster Expansions, Physical Review Letters 92 (25) (2004) 255702. doi:10.1103/PhysRevLett.92.255702.
[60] E. F. Holby, J. Greeley, D. Morgan, Thermodynamics and hysteresis of oxide formation and removal on platinum (111) surfaces, The Journal of Physical Chemistry C 116 (18) (2012) 9942–9946.
[61] C. Lazo, F. Keil, Phase diagram of oxygen adsorbed on Ni(111) and thermodynamic properties from first-principles, Physical Review B 79 (24) (2009) 245418.
[62] R. B. Getman, Y. Xu, W. F. Schneider, Thermodynamics of environment-dependent oxygen chemisorption on Pt(111), Journal of Physical Chemistry C 112 (26) (2008) 9559–9572.
[63] L. J. Nelson, V. Ozoliņš, C. S. Reese, F. Zhou, G. L. W. Hart, Cluster expansion made easy with bayesian compressive sensing, Phys. Rev. B 88 (2013) 155105.
[64] L. J. Nelson, G. L. W. Hart, F. Zhou, V. Ozoliņš, Compressive sensing as a paradigm for building physics models, Phys. Rev. B 87 (2013) 035125.