Quantifying Errors in Effective Cluster Interactions of Lattice Gas Cluster Expansions

Greg Collinge[a,*], *Alyssa Hensley*[a,b], *and Jean-Sabin McEwen*[a,c,d,e,f*]

[a] The Gene & Linda Voiland School of Chemical Engineering and Bioengineering, Washington State University, Pullman WA 99164, USA

[b] Department of Chemical Engineering and Materials Science, Stevens Institute of Technology, Hoboken, NJ, 07030

[c] Department of Physics and Astronomy, Washington State University, Pullman, WA 99164, USA

[d] Department of Chemistry, Washington State University, Pullman, WA 99164, USA

[e] Institute for Integrated Catalysis, Pacific Northwest National Laboratory, Richland, WA, 99352, USA

[f] Department of Biological Systems Engineering, Washington State University, Pullman, WA 99164, USA

* Corresponding authors:

Jean-Sabin McEwen; 509-335-8580 (phone), 509-335-4806 (fax), js.mcewen@wsu.edu

Greg Collinge; 509-993-6086 (phone), gbcollinge@wsu.edu

Abstract

The promise of lattice gas (LG) cluster expansions (CEs) is that they can describe a given system property to any level of accuracy since the orthogonal "cluster basis functions" span the complete space of available configurations. Such an approach can be constructed to an arbitrarily large surface of a finite number of distinct adsorption sites. Unfortunately, this is only true for the case of an ideal, fixed lattice decorated with components at precise lattice points (the lattice "sites") with no distortions or relaxations subsequently allowed. Since most systems, and surfaces specifically, do not conform to such an ideal set of constraints, errors in the LG CEs must be expected or CE convergence severely hampered. Beyond this, numerical errors in the provided data can complicate the proper construction of a truly predictive and/or physically

significant CE. We show here how reliance on typical statistical tools like confidence intervals cannot be expected to provide an accurate representation of the uncertainty of the effective cluster interactions (ECIs) in the CE due to the nature of the target *ab initio* data and the nature of CEs themselves. We develop a method for estimating these errors that does not rely on statistical assumptions about the model or data. We then use these ECI errors to quantify the fundamental consequences on the uncertainty of ECIs in CEs built from O/Fe(100) data whose surface and adsorbates have been allowed to relax in the typical manner and from O/Fe(100) data whose surface and adsorbates are fixed in ideal lattice positions. We also quantify the effect of using a different density functional theory exchange correlation functional, using these ECI errors to assess the significance in any deviations. In both cases, our method is shown to have incredible utility in the quantification of errors in the ECIs of CEs. While we stick to the lattice gas convention in this work, the method is in principle equally applicable to the Ising convention or, in principle, any linear model of sufficient complexity.

1. Introduction

Fitting cluster expansions (CEs)[1-4] to *ab initio* data—generally from density functional theory (DFT) calculations—is a well-established tool for the characterization and rationalization of alloy and surface behavior. The theory (and to an extent, the validity) of CEs relies on the treatment of systems as a configurational problem: that of decorating an ideal, fixed lattice with substitutional components. Early on in the development of CEs, the effect of geometric relaxations in the underlying structural data used to fit the CE was recognized as an issue[5-12] requiring special attention and having considerable consequence on the energetic behavior of the constituent system including (but not limited to) lowered miscibility-gap temperature,[5] incorrect

predicted ordered structures,[7] and shifted density of states.[8] More recently, it has been suggested that these relaxations are in direct contradiction to the theory of CEs and that some supercell volume/concentration-dependency needs to be included in the effective cluster interactions (ECIs) of CEs or else the CE will fail.[4, 13] The idea behind concentration-dependent ECIs is that in, e.g., substitutional alloys the higher concentration of one component over another will cause the unit cell to shrink or expand to accommodate the extra atoms (depending on the relative covalent radii of the components). If the unit cell of an alloy shrinks or expands, the distance between nearest neighbors shrinks or expands accordingly and the same energetic contribution, or ECIs, should not be expected from clusters made up of these nearest neighbors across unit cells of varying size. Interestingly, volume-dependent ECIs were indeed used long before this in the early 1990s,[14, 15] but they appear to have fallen out of favor since then. Nonetheless, the consequences of geometric relaxations are still of active interest within the community of CE developers/users,[4, 16-22] but a method for directly quantifying their effect on the CEs themselves is still missing.

Here, we are specifically interested in the lattice gas (LG) model[23, 24] and how internal geometric relaxations of surface/adsorbate system calculations affect fitted LG CEs. In such calculations, the supercells used are typically not allowed to relax, meaning the volume is constant and concentration-dependent ECIs, as proposed by Sanchez,[4] will not capture the effect of relaxations in our LG CEs. However, internal relaxations of the supercell's constituent atoms are allowed. Due to the presence of a vacuum layer and often significant surface void space, these internal relaxations can be quite severe—perhaps more so than in substitutional alloy systems. Despite this, concentration-*in*dependent ECIs are also important to the interpretation of ECIs in the LG paradigm, and so we are inclined to assume they will remain a mainstay of LG

CE development. In this case, our aim is to quantify the uncertainty in these ECIs due to relaxations. As the ECIs within LG CEs are typically fit to *ab initio* data, another important factor in the ECI uncertainty is the energy variation induced by the choice of exchange correlation function, so we take the opportunity to also assess the uncertainty of ECIs due to this factor.

In order to quantify the extent to which geometric relaxations (away from ideal lattice positions) and DFT functional affect the quality of CEs, our goal is to assess, quantitatively, the uncertainty in CE's ECIs. Since CEs are generally constructed via a linear regression procedure, it is tempting to appeal to the statistics of linear regression for this purpose since Analysis of Variance (ANOVA) and subsequent confidence intervals (CIs) are often used for precisely this purpose. Unfortunately, the legitimacy of these statistics is based on three basic assumptions[25] that *ab initio* data and CEs do not uphold: (1) normality—that the residuals are normally distributed (or at least distributed in a like manner) with a mean value of zero, (2) homoscedasticity—that the model independent variables (the ECIs) are distributed equivalently (e.g. same variance) about their mean value, which is taken to be the "unbiased estimator" of the true underlying value, and (3) independence—that the independent variables (again, the ECIs) are uncorrelated. Because the *ab initio* data we wish to use in fitting CEs have errors unrelated to the kind of random sampling error seen in experimental data, it is not difficult to see that the first assumption, normality, is very likely violated. More crucially however, the second and third assumptions, homoscedasticity and independence, are violated by the very nature of CEs themselves. This is because higher-body ECIs are expected to have sequentially lesser effect on the final energies or else truncating the CE would not be possible (violation of assumption 2). Also, since higher-body clusters are built up from lower-body clusters (e.g., the removal of any

of the 3 two-body clusters making up a three-body cluster will *always* result in the removal of that three-body cluster) the ECIs will absolutely be correlated (violation of assumption 3). Thus, we should not at all expect CIs (or other ANOVA-derived statistics) to be representative of the CE's true uncertainty.

Here, we will present a method for determining "ECI errors" that avoids statistical assumptions in their entirety. Instead we rely on a direct sampling procedure that takes advantage of our convergent implementation[26] of the leave-multiple-out cross-validation (LMO-CV) score[27]. These ECI errors are then used to allow a direct comparison of four similar systems wherein atoms are either fixed or allowed to relax to their simultaneous energetic minimum; and wherein a standard generalized gradient approximation (GGA) functional and a van der Waals containing functional are used.

2. Theory and Formulation

2.1. Defining ECI errors

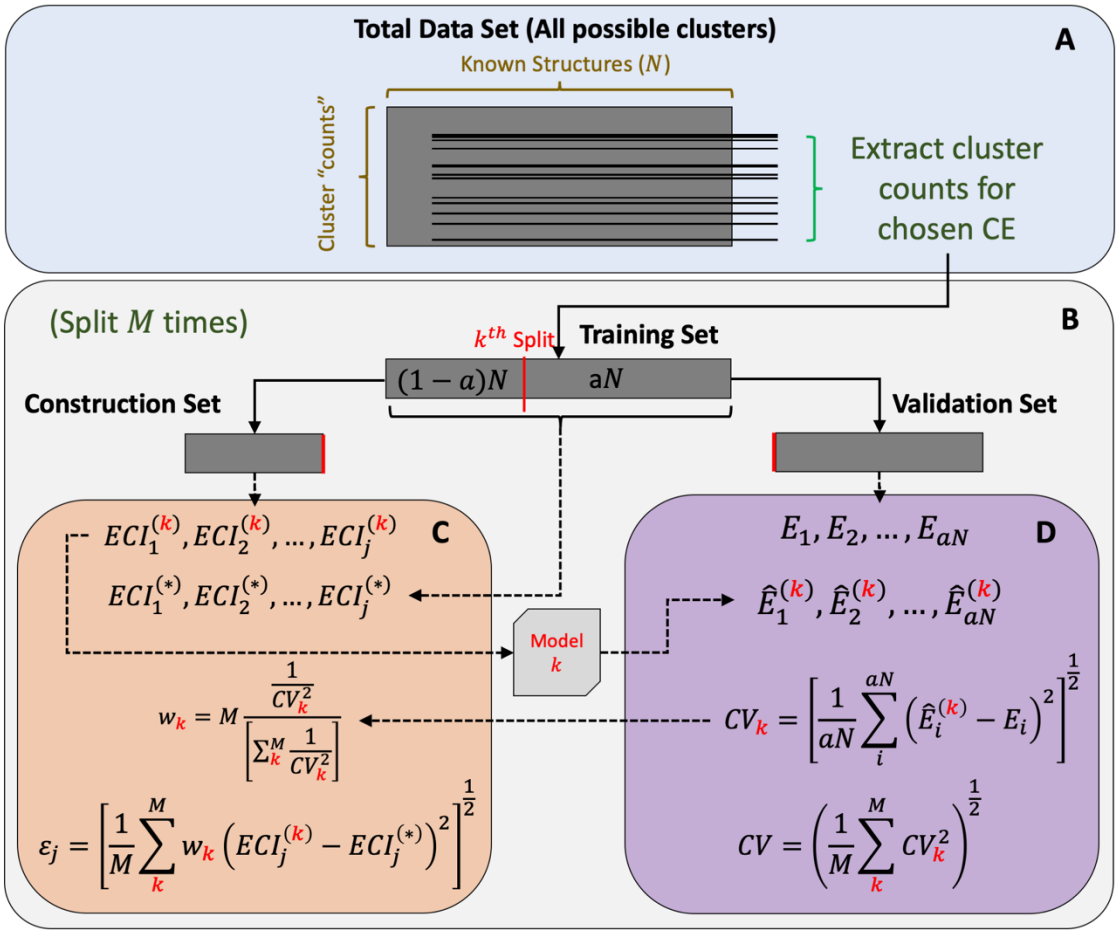


Figure 1. Illustration of the pertinent information extracted during the calculation of the LMO-CV score starting from (A) the complete set of data comprised of $\boldsymbol{N}$ structures where all specified possible cluster 'counts' are included. A subset of clusters corresponding to the chosen CE is extracted from all structures becoming the training set used for (B) evaluation of the LMO-CV score. The data (known structures) are split into various (ideally) random subsets of size $\boldsymbol{(1-a)N}$ (the construction set) and $\boldsymbol{aN}$ (the validation set) where $\boldsymbol{a}$ is the "fraction left out." The data needed to evaluate the ECI errors, ε_j, are sequestered and shown in (C), while the data needed for the evaluation of the LMO-CV score are sequestered and shown in (D). Here, indices i, j, and k represent the i^{th} structure in the validation set, the j^{th} ECI in the CE under consideration, and the k^{th} cut of the training set made to form the validation and construction sets; superscript $(*)$ indicates the ECI value found from the total training set; and M is the number of cuts made. The remaining variables are defined in the main text.

Our procedure for determining ECI errors is shown in Figure 1. Once a complete data set of N *ab initio* structural data (which contains a CE "fingerprint"—i.e., an enumeration of all clusters present—for each structure along with their computed *ab initio* energies) has been collected (Figure 1A), the "training set" can be extracted, which is simply the complete data set with a subset of cluster "counts"—i.e., the integer number of each type of cluster present — selected to form a potential CE (thus implicitly setting all other possible cluster ECIs to zero). As shown in Figure 1B, this training set is then split into a "construction set" (CS) and a "validation set" (VS). The split is ideally made by selecting a random subset of aN structures (where a is the "fraction left out") to be the VS and then using the remaining $(1-a)N$ structures as the CS (we use $a = 0.6$, based on the recommendations of Baumann[27] and considerations described elsewhere[28, 29]). We denote each split and corresponding CS/VS subset of structures with index k and the ECIs are then fit to the CS data to provide a set of ECIs specific to that k^{th} subset of CS structures: $\left\{ECI_j^{(k)}\right\}$, where j demarks the j^{th} cluster in the chosen CE and superscript (k) denotes from which CS subset the ECI was fit. From this newly created CE model, the energies of the structures in the VS (which were not used in the previous fitting) are then predicted ($\hat{E}_i^{(k)}$ for the i^{th} structure; a "hat" denoting prediction) and compared against their actual computed energy values (E_i). A root mean squared prediction error is then assessed for that k^{th} split as

$$CV_k = \left[\frac{1}{aN}\sum_{i}^{aN}\left(\hat{E}_i^{(k)} - E_i\right)^2\right]^{\frac{1}{2}} \quad (1)$$

The LMO-CV score is calculated as the root mean squared CV_k across some M random splits of the training set data (making it the root mean squared error of all predictions):

$$CV = \left(\frac{1}{M} \sum_{k}^{M} CV_k^2 \right)^{\frac{1}{2}} \tag{2}$$

The number of cuts required (M) is chosen so that the LMO-CV score converges to some tolerance (more on this value in Section 3.5). The optimum CE for a data set is that which produces the lowest LMO-CV score and the final model ECIs are assessed by fitting to the *entire* training set, $\left\{ ECI_j^{(*)} \right\}$, where superscript $(*)$ denotes the final model "ultimate" ECI.

To directly assess the "ECI errors," the data shown in Figure 1C is collected alongside the data needed for the evaluation of the LMO-CV score in Figure 1D as described above. As different training set splits are made, a different $ECI_j^{(k)}$ is produced, unique to the subset of structures in the CS. Since the CS has fewer structures than the total training set, we can expect these values to be "better" solutions for those structures and the deviation of these values from the final ECI ($ECI_j^{(*)}$) a good indication of the potential variance in that ECI. However, at the same time, if the model produced with $ECI_j^{(k)}$ gives a poor prediction error in the VS (i.e., a high CV_k or low $1/CV_k$) we lower the significance of that deviation (whether initially large or small). We do this by weighting each deviation with a value proportional to $1/CV_k$. These considerations result in the following equation:

$$\varepsilon_j = \left[\frac{1}{M} \sum_{k}^{M} w_k \left(ECI_j^{(k)} - ECI_j^{(*)} \right)^2 \right]^{\frac{1}{2}} = \left[\frac{1}{M} \sum_{k}^{M} \left\{ \sqrt{w_k} \left(ECI_j^{(k)} - ECI_j^{(*)} \right) \right\}^2 \right]^{\frac{1}{2}} \tag{3}$$

where ε_j is the "ECI error" and w_k is a weighting factor that accounts for the aforementioned poor prediction behavior:

$$w_k = M \frac{\frac{1}{CV_k^2}}{\left[\sum_k^M \frac{1}{CV_k^2}\right]} \quad (4)$$

We use the square of the LMO-CV score in Equation 4 so that the weight, w_k, on the deviation is in fact proportional to $1/CV_k$ once its square root, $\sqrt{w_k}$, is taken. This is normalized such that the sum of all w_k reproduces the number of training sets cuts performed, M. This ensures that if each CV_k were equivalent, the unweighted expression would be recovered. An a priori unweighted version of Equation 3 would have w_k set to unity (i.e., 1).

A technical difficulty arises in the evaluation of the LMO-CV score that deserves mentioning here. Due to the random nature of the training set splits, some structures can become over (or under) represented in the various M number of VS's. This is illustrated in Figure 2A – 2C. The bias can be quantified by defining a "VS residence" vector, $\mathbf{v}$, with N components $\{v_i\}$, one for each structure in the training set, initially starting at 0 (Figure 2A). The bias is then defined as

$$Bias = \max[\mathbf{v}] - \min[\mathbf{v}] \quad (5)$$

As structures are randomly assigned to the VS, their presence is marked by an increment in its VS residence (Figure 2B) and the build-up of bias becomes apparent after a subsequent split of the training set where, due to the random nature of the cuts, some structures are added twice while others are left out twice (Figure 2C). A CV score based on these two training set splits would be biased toward the structures overrepresented in the VS's. To fix this, the underrepresented structures are constantly identified and added to the next validation set, with the remainder being chosen at random. This is shown in Figure 2D. If this is done continuously,

the degree of non-randomness is kept to a minimum and the bias can be kept to below 2, which at large enough M is for all practical purposes completely unbiased.

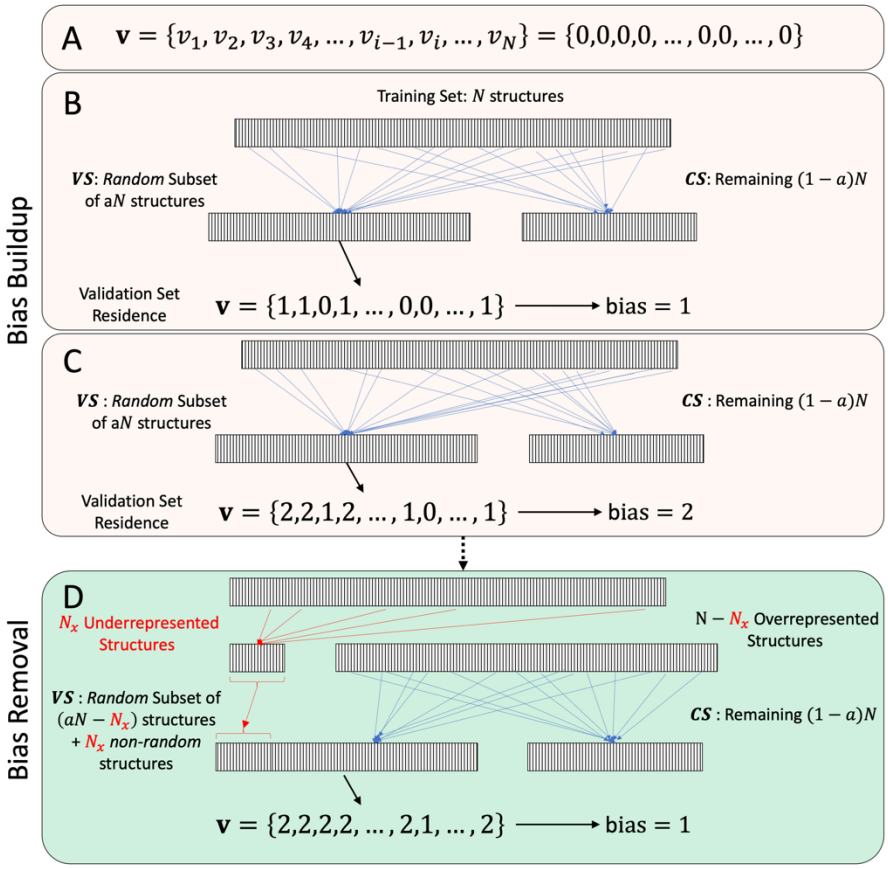


Figure 2. Demonstration of bias build up (orange shaded regions) and bias removal (green shaded region). (A) The initialization of the "VS residence" vector defined in the main text before Equation 5. (B) The training set is split into a validation set (VS) and construction set (VS) choosing $\boldsymbol{aN}$ structures at random in order to form the VS. The VS residence vector indices corresponding to the $\boldsymbol{aN}$ structures chosen are incremented by 1. (C) The same process in (B) is repeated and the formation of a bias in the structures chosen to be in the VS becomes apparent. (D) To remove the bias, those $(\boldsymbol{N_x})$ structures that have been underrepresented are identified (red font and arrows) and added to the VS before the remaining $(\boldsymbol{aN - N_x})$ VS structures are chosen at random to finish forming the VS. In so doing, the bias is reduced.

In order for all these procedures and formulae to be visualized, a LMO-CV score calculation is shown in Figure 3 where the incremental CV score is allowed to converge *without* bias removal in order to demonstrate bias build-up and the resultant behavior in the evaluations of CV_k and w_k before and after bias removal is implemented (as shown in Figure 2). As can be seen in Figure 3A and 3B, the random sampling of the data is largely unaffected by the bias

removal, and importantly, the incremental CV (Figure 3C) score starts converging to a completely different result after the bias is removed, showing in this case that the sampling was biased toward structures that were better predicted than other structures in the data set (since it was converging to a *lower* CV score). The w_k values distribute about a value of 1 in a way consistent with the development of Equation 4. The point at which bias removal begins is apparent in Figure 3D, where a bias of nearly 70 structures had accumulated shortly before 1000 training set splits were made where it is evident that the bias removal algorithm shown in Figure 2 performs very quickly and efficiently. It is important to note that, in practice, the bias removal algorithm is always implemented from the start of the LMO-CV score calculation with negligible effect on the final converged CV score.

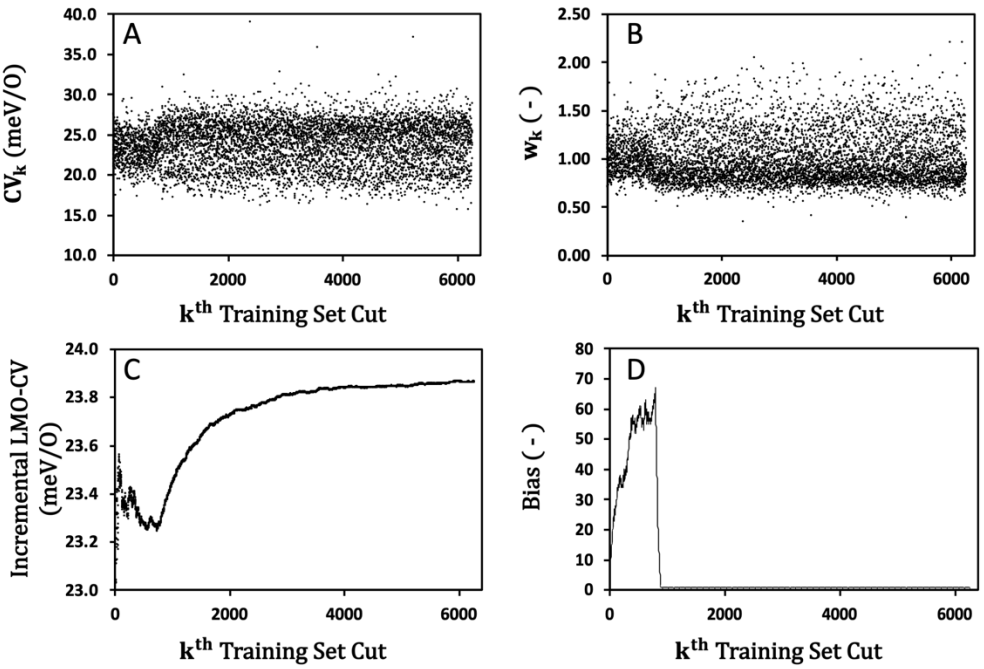


Figure 3. Visualization of various quantities calculated during the evaluation of the LMO-CV score and corresponding ECI errors: (A) $\boldsymbol{CV_k}$ as defined in Equation 1 and Figure 1 for each $\boldsymbol{k^{th}}$ training set cut as also illustrated in Figure 1; (B) the weight, $\boldsymbol{w_k}$, applied to each squared ECI deviation of Equation 3; (C) the incremental LMO-CV score as calculated up to the current $\boldsymbol{k^{th}}$ training set cut (i.e. the root mean squared value of the preceding data in (A)); and (D) the bias as defined in Equation 5.

2.2. Data Sets and Methodology

Our primary motivation for developing the procedure for determining ECI errors is to be able to quantify the uncertainty in CEs brought about by uncertainty or "noise" in the data set, specifically due to geometric relaxations and choice of exchange correlation functional. To do this (and also demonstrate the implementation of the procedure outlined in the previous section), we have developed a LG CE for four different systems: (1) the relaxed-O/Fe(100) RPBE system, using the raw data from Bray et al.,[30] (2) for the fixed-O/Fe(100) RPBE system using data produced for this work, (3) for the fixed-O/Fe(100) optB88-vdW system, using data again produced for this work, and (4) relaxed-H/Fe(100) system using the raw data from Hensley et al.[31]. In the O/Fe(100) systems, O adatoms are confined to the 4-fold hollow sites of the Fe(100) surface (which is the most favorable site[30, 32]), while in the H/Fe(100) system both 4-fold hollow and bridge sites are included due to their similar adsorption energies; a pseudo 3-fold hollow site was also included in the work of Hensley et al.[31] due to massive lattice relaxations during their optimizations (hence why this system is of interest here). In the "relaxed" systems, the top two layers of the Fe(100) model slab and adatoms were allowed to relax completely during geometric optimization, while in the fixed systems, the atoms of the Fe(100) slab and the O adatoms are kept fixed in their "ideal" positions (i.e. the clean surface geometry and the O in the direct center of the 4-fold hollow site at its ideal height for an isolated O adatom on that fixed surface). The data produced for this work, i.e. "fixed" O/Fe(100) systems, were generated in an automated manner using the Alloy Theoretic Automated Toolkit (ATAT)[33-35] while the lattice gas models were found by fitting to the adsorption energies of all data and optimizing using our recently

developed *Ab initio* Mean-field Augmented Lattice Gas Modeling (AMALGM) code[26]. Computational methodology for the *ab initio* calculations is given in the SI. The *ab initio* adsorption energies for all systems along with their final LMO-CV scores and root-mean-squared residuals (RMSRs). The corresponding clusters for these systems are shown in Figure 5. Note that while the clusters shown for the O/Fe(100) systems (Figure 5A – 5C) are exhaustive, only the clusters with ECI magnitudes greater than 10 meV are shown for the H/Fe(100) system (Figure 5D) because this system has more than double the number of clusters than the O/Fe(100) systems.

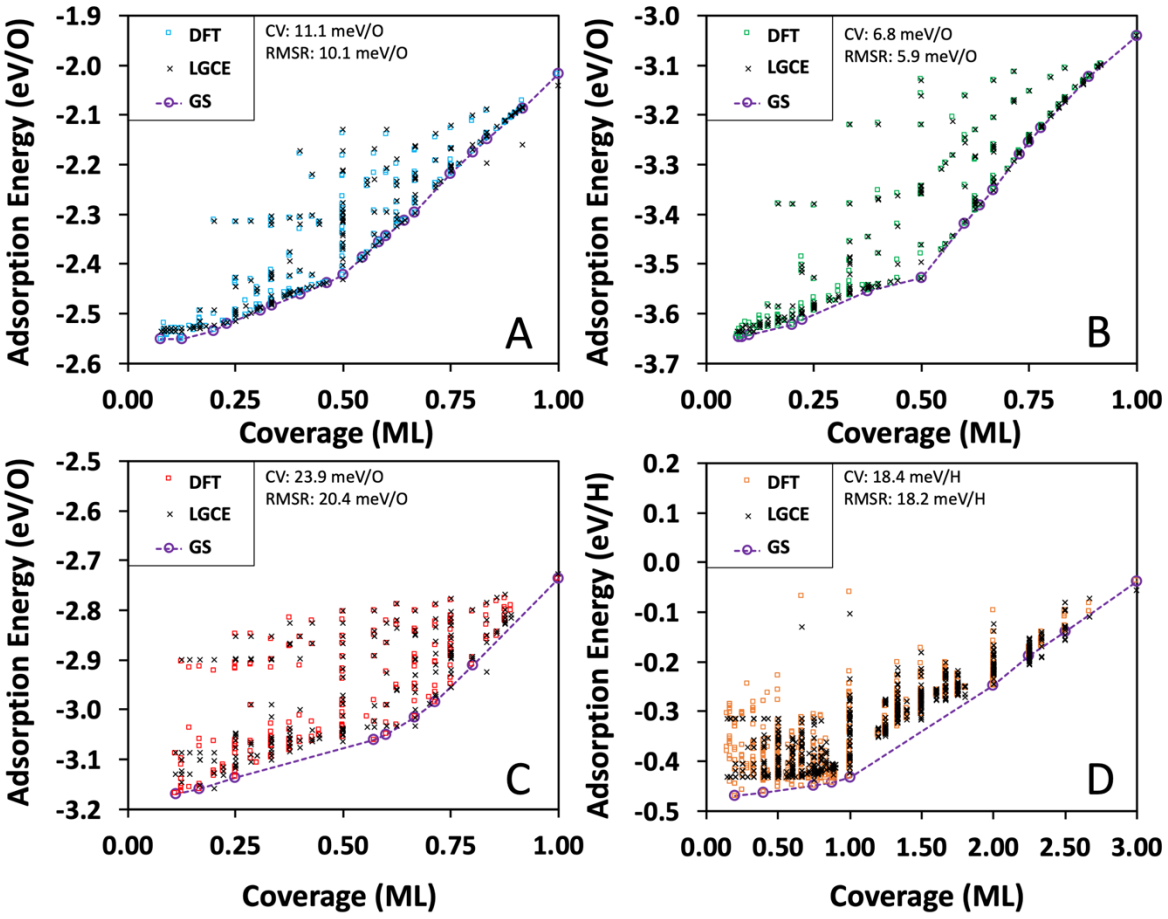


Figure 4. DFT-calculated and LG-predicted adsorption energies for (A-C) the O/Fe(100) system and (D) the H/Fe(100) system. In (A) and (B), the Fe lattice is fixed in its bulk-optimized position and the O is fixed in the ideal position for an isolated O atom on this fixed lattice. In (C) and (D), both adsorbate and top two layers of the lattice are allowed to fully relax. The RPBE functional was used in (A) and (C), while the optB88-vdW functional was used in (B) and (D). The ground states are denoted GS in the

legends. Note that the x-axis in (D) ends at 3.0 ML instead of 1.0 ML due to the greater complexity and configurational space spanned in the H/Fe(100) system (see Ref. [31]).

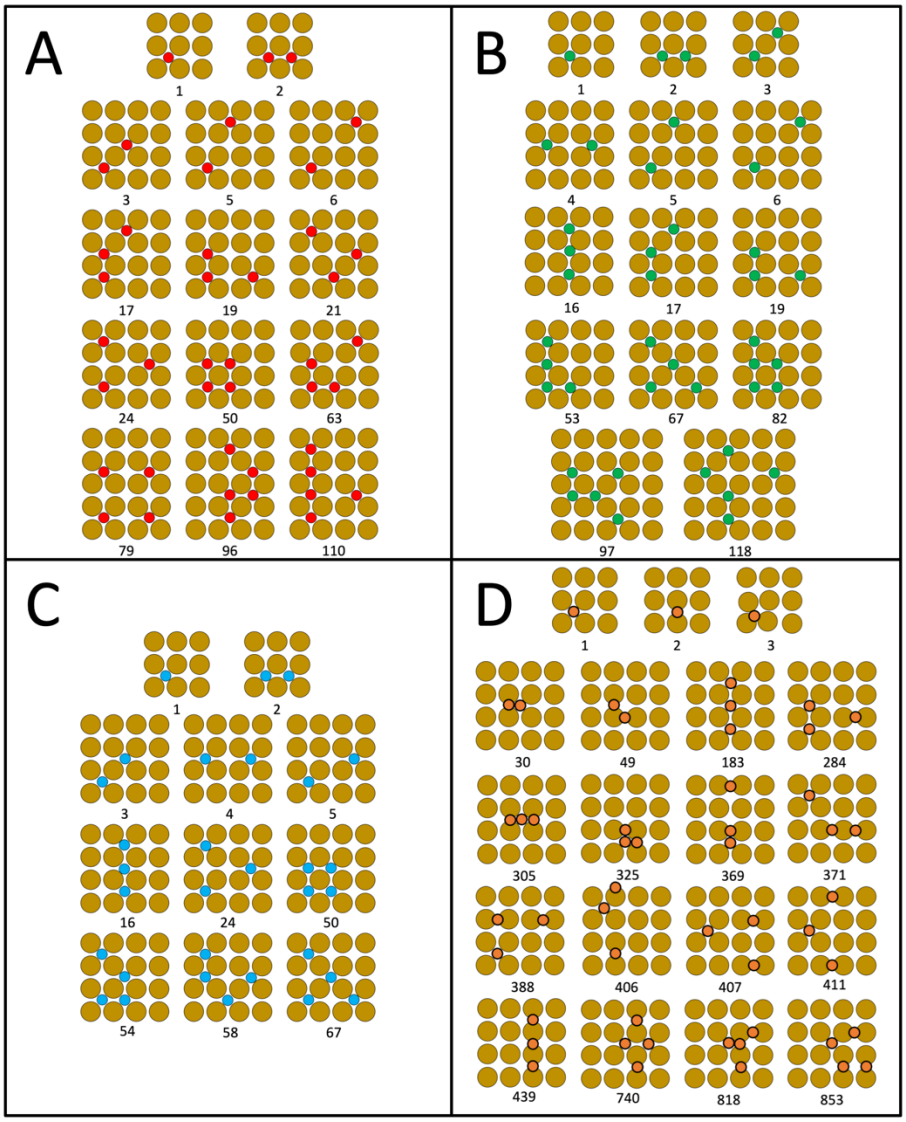


Figure 5. Clusters and their cluster IDs corresponding to the optimum LG CEs found for the systems shown in Figure 4: (A) relaxed-O/Fe(100) RPBE, (B) fixed-O/Fe(100) optB88-vdW, (C) fixed-O/Fe(100) RPBE, and (D) relaxed-H/Fe(100) optB88-vdW. The color of the adatoms corresponds to the color coding established in Figure 4, with red circles, green circles, and blue circles denoting oxygen atoms in their respective system; and orange circles denoting hydrogen atoms. The gold circles denote iron atoms.

3. Results and Discussion

3.1. Visualization of the ECI errors

Before analyzing and comparing the ECI errors found for the optimum LG CEs produced for the four systems shown in Figure 4, we wish to explicitly allow the reader to visualize the direct sampling being made in the evaluation of the ECI errors as defined in Equation 3 as the interpretation of ECI errors is contingent upon the behavior of the ECI deviations from which they are derived. Figure 6 shows the quantity in curly brackets in Equation 3 collected over all training set splits needed to converge the respective LMO-CV scores for the relaxed-O/Fe(100) RPBE (Figure 6A – 6B) and fixed-O/Fe(100) RPBE systems (Figure 6C – 6D) using both the optimum LG CEs for the relaxed system (Figure 6A and 6C) and the fixed system (Figure 6B and 6D). It is immediately apparent when comparing the relaxed-O/Fe(100) to the fixed-O/Fe(100) system that the ECI deviations in the relaxed system have an overall greater variation (~175% greater) than in the fixed system regardless of what CE is used, suggesting greater uncertainty in the relaxed system ECIs that is fundamentally unavoidable. This is apparent because, while the ECI deviations are greater in both systems when using their non-optimum LG CEs (Figure 6B and 6C), the non-optimal CE for the fixed system (Figure 6C) still displays less variation in its ECI deviations than those of the optimum CE for the relaxed system (Figure 6A). Specifically, the relaxed system displays ~70% greater overall variance in this comparison.

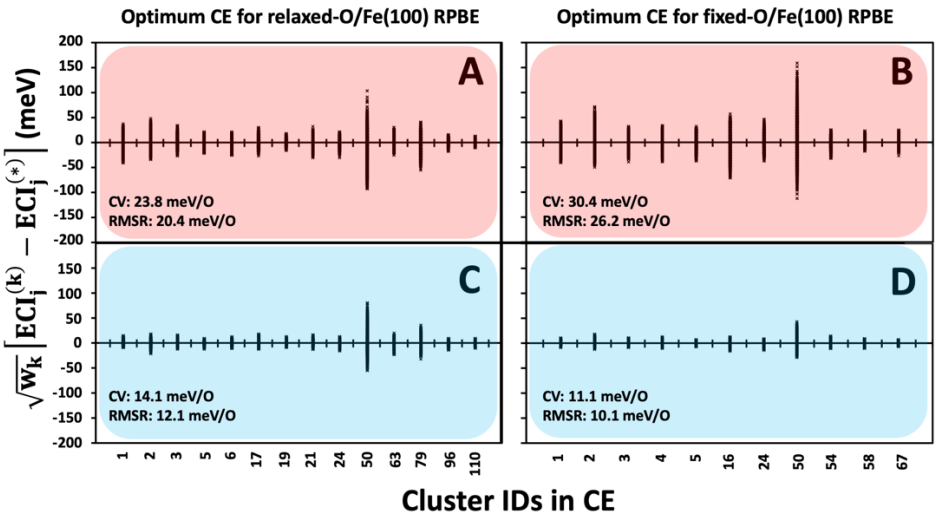


Figure 6. Illustration of how the weighted ECI deviations ultimately distribute about zero for each respective cluster in the displayed CE. Panels (A – B) show these results for the relaxed-O/Fe(100) RPBE system while panels (C – D) show results for the fixed-O/Fe(100) RPBE system. To compare the two on a one-to-one basis, the optimized CE for the relaxed-O/Fe(100) RPBE system is used in A and C while the optimum CE for the fixed-O/Fe(100) RPBE system is used in B and D. The LMO-CV scores and RMSRs for these systems are shown inset for context. Red and blue backgrounds are added to signify which data sets are relaxed or fixed, respectively.

It is important to comment, for the interpretation of the ECI errors we wish to report, that the values in Figure 6 distribute nearly symmetrically about 0 meV for all the clusters. This suggests that their root mean squared value (i.e. ε_j from Equation 3) can indeed be used as a $\pm$ "standard deviation" of the distribution for the final ECIs. A "two sigma" value (i.e. two times the root mean squared value) accurately captures the height of these data and we will therefore use $\pm 2\varepsilon_j$ as the reported ECI error and to determine ECI significance. For comparison, the unweighted version of these distributions of ECI deviations are shown in Figure S2. We should note that, in this work, the ECI error values derived from the unweighted version is negligibly different (<1 meV) from the weighted version.

3.2. Quantifying Uncertainty Due to Relaxations and Choice of DFT Functional

We first compare ECIs and their ECI errors for the O/Fe(100) systems to illustrate the *quantitative* errors induced in CEs by geometric relaxations and choice of exchange correlation functional in the underlying *ab initio* data. Figure 7 shows the ECIs (colored columns) and their associated "two sigma" ECI errors (black error bars) for these systems. The relaxed-O/Fe(100) RPBE system is compared directly to the fixed-O/Fe(100) RPBE system using their respective optimum CEs in Figures 7A and 7B; the ECI errors for these systems are derived from the data shown in Figure 6. Based on the LMO-CV scores for these two systems when using the optimum LG CE for the relaxed-O/Fe(100) RPBE system (Figure 7A), the fixed system (in blue) is actually quite amenable to using a non-optimal CE (its LMO-CV score increases to 14.1 meV/O from its optimum value of 11.1 meV/O—only a 27% increase) and is still more predictive than the optimum value for the relaxed system (23.8 meV/O). We are therefore very confident in the direct comparison between these two systems. The relaxed system LMO-CV score increases by a similar amount (28%) when using the optimum LG CE for the fixed-O/Fe(100) RPBE system (Figure 7B), but the disparity between the two CV scores is great enough that we are less confident in their direct comparison. We therefore turn to Figure 7A for our analysis first.

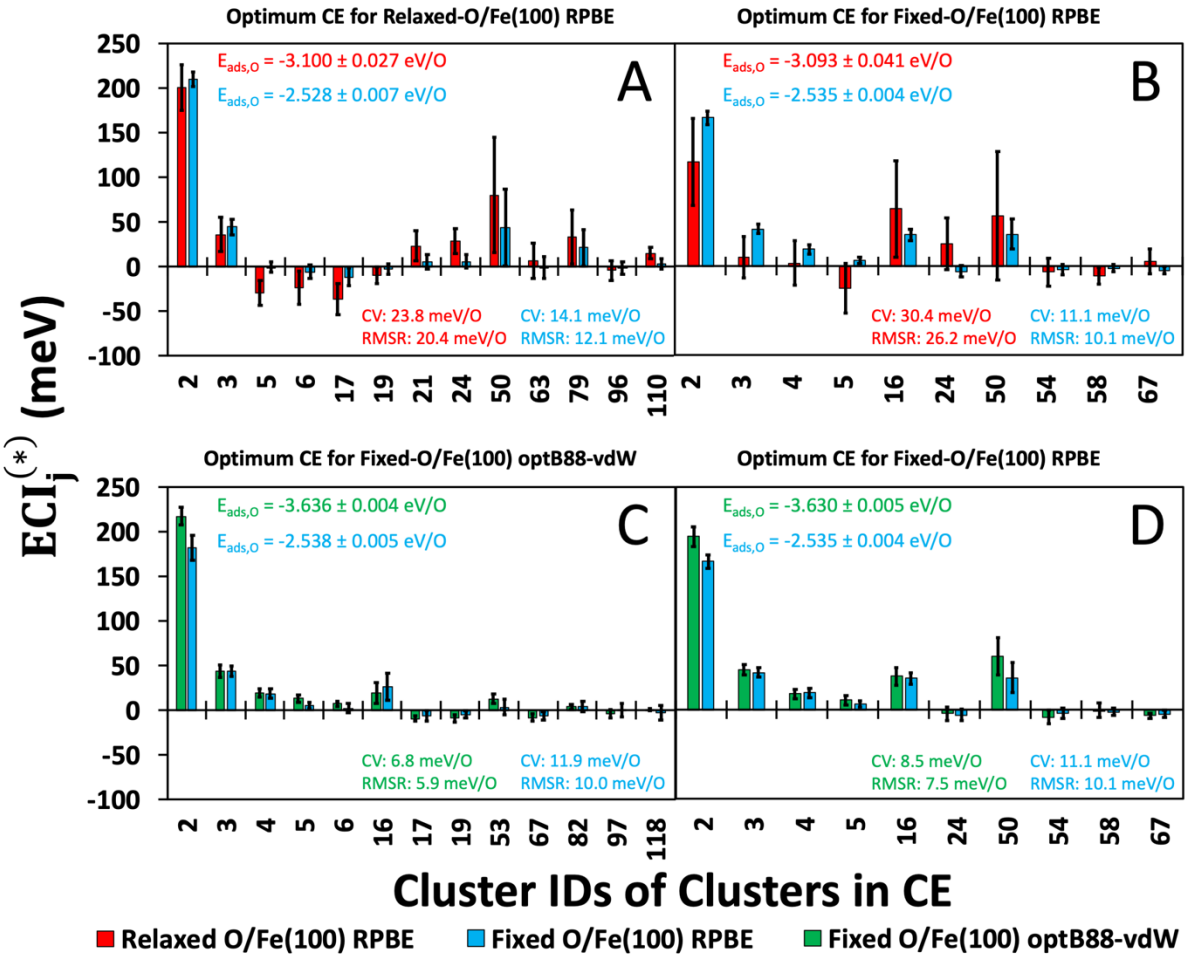


Figure 7. Comparison of the (A – B) relaxed-O/Fe(100) RPBE (red bars and text) vs. fixed-O/Fe(100) RPBE (blue bars and text) systems; and comparison of the (C – D) fixed-O/Fe(100) optB88-vdW (green bars and text) vs. fixed-O/Fe(100) RPBE (blue bars and text) systems. The optimum CE for each data set is used to allow direct comparison of ECIs and ECI errors. The ECI errors are shown as black error bars. Due to magnitude discrepancy, the isolated O/Fe(100) adsorption energy, $E_{ads,O}$ (corresponding to cluster ID #1), is shown in the inset in each panel. The LMO-CV score and RMSR are also shown as insets with the appropriate color coding established in Figure 4.

An interesting feature of the ECIs in Figure 7A themselves is that their character (i.e. repulsive vs. attractive) is essentially the same between the fixed vs. relaxed systems. Only cluster #63 shows contrary character in the two systems, but only barely (-1 meV vs +7 meV in the fixed and relaxed systems respectively). This suggests that relaxations do not necessarily serve to reverse the most basic physics of the underlying system. However, this is not necessarily true if the ECI errors are taken into account, which indicates that for many structures in the data,

ECI values could in fact be repulsive, attractive, or simply 0 meV. It is for this reason that we must deem ECI values with magnitudes less that their ECI errors to be insignificant, and in Figure 7A, this is true of clusters 5, 6, 19, 21, 24, 63, 96, and 110 for the fixed system. For the fixed system in Figure 7A, such an effect is mostly due to the fact that the ECI magnitudes are small to begin with (recall that this is a non-optimal LG CE for this system). In the relaxed system, only clusters 63 and 96 are insignificant—again, primarily because their ECI magnitudes are small. However, comparing between the two systems we can see that for all clusters except 24 and 110, the fixed and relaxed systems' ECI error bars overlap. The relaxed system ECI errors are on average 10 meV greater than then those in the fixed system suggesting that the overlap can be primarily blamed on the increased uncertainty in the relaxed system ECIs. This suggests, rather shockingly, that there are structures within the relaxed data set that could be better predicted using the ECIs derived from the *fixed* data set. Note, however, that there are many more structures within the relaxed data set that cannot be predicted with any level of accuracy using ECIs from the fixed data set. Thus, this finding, while intriguing, should not be used as justification for arbitrarily fixing the underlying *ab initio* structures during data generation.

It is important to reemphasize that this apparent equivalence between relaxed and fixed system ECIs does not mean that the final fixed system LG CE can be used in place of the relaxed system LG CE. Relaxation can affect the magnitude of higher-body interactions and we do not anticipate that the same predictability of the fixed system, as shown in Figure 7A, to be necessarily applicable to other adsorbate systems.[36] On the one hand, there is no way to know *a priori* what this "essentially same" LG CE for the fixed system is because it is not its optimum CE. On the other hand, it was a *set* of ECIs that provided better agreement to the energies, not

just a single perturbation to a single ECI value—meaning, as we've stated previously, that the errors are *correlated*—so we cannot expect to get better fits from just any set of ECIs that fall within the error bars. Nonetheless, this does mean that we are highly *uncertain* that the final ECIs in the relaxed system will accurately predict the energetic consequences in the relaxed structure data that are outside the training set much better than if those data were based on fixed structures. While this does not mean that relaxations cannot be captured by CEs, it does mean that it is incredibly important to be able to quantify the uncertainties inherent in their inclusion, and our method provides exactly that.

The results shown in Figure 7C and 7D, comparing the fixed-O/Fe(100) RPBE system (blue bars) to the fixed-O/Fe(100) optB88-vdW system (green bars), demonstrate the consequences brought about by a change in the DFT functional. Here, the LMO-CV scores are both small and comparable allowing for the direct comparison of both CEs used. Strikingly, the error bars again overlap for virtually every cluster in either Figure 7C or 7D; only the adsorption energy and 1st nearest neighbor (NN) (Cluster #2) are shown to be significantly different. However, this time, there is only a 2 meV discrepancy in average ECI errors, so this is due to the ECIs themselves being practically equivalent for all clusters (again except for the adsorption energy and 1st NN). Thus, in this case, we have shown that the primary consequence of using a different DFT functional is in the adsorption energy and 1st NN, but in practically no other terms. Even discounting the error bars, there is only one cluster (#50 in Figure 7D) that is large enough and different enough between the two data sets to be considered consequentially different (a 24 meV difference). This corresponds to a fairly standard "usual suspect": the most compressed square 4-body interaction (Cluster #50 in Figure 5A); a usual suspect due to it containing four 1st NN subclusters. This means that, in principle, one need not actually collect an entire dataset of

structures using a new functional, but instead simply calculate a handful of interactions corresponding to the "usual suspects" (i.e. a few nearest neighbors and other highly compressed higher-body clusters) and replace these in an already optimized CE. This of course is only potentially true in these fixed systems, but given the uncertainty shown in the relaxed system, this result may well be universal.

3.3. Uncertainty in the H/Fe(100) system

Here, we expand our consideration of ECI error quantification of relaxations in a more complex system, namely H/Fe(100). In this system, both 4-fold hollow and bridge sites were included, but due to massive geometric relaxations of the underlying Fe(100) lattice that are a function of hydrogen coverage, a pseudo 3-fold hollow had to be added to the site definitions in the LG CE.[31] This added complexity along with the underlying severity of relaxation makes for an ideal test case for the analyses already laid out in Section 3.2.

As can be seen in Figure 8, the ECI errors assessed for the H/Fe(100) system are significant. This leads to high uncertainty in the resulting ECIs. Despite this, none of the insignificant ECIs are greater than 10 meV, meaning that the insignificant ECIs would likely have little effect on the chemical behavior for the LG CE. Thus, despite the massive geometric relaxations present in this system, the LG CE has captured the most important physical interactions even with a quantitatively larger degree of uncertainty in their exact values than those for the O/Fe(100) systems. This is also reflected in the reasonably low LMO-CV score of 18.4 meV/H (Hensley et al.[31] fit their LG CE to the data's surface energies, providing a value of 11.5 meV/unit cell, which is difficult to compare to our values fitted to adsorption energies due to different normalization and thus different units). It is also worth noting that the majority of the

most significant ECIs (e.g., Clusters 30, 49, 284, 305, and 325) involve both 4-fold hollow and bridge sites showing that the increased complexity of this system is being captured. Nonetheless, these interactions have non-negligible ECI errors and thus high uncertainty, making it difficult to determine if one source of uncertainty is not the increased complexity itself. Based on the results from Section 3.2, however, we can definitely say that a very major component contributing to these uncertainties is geometric relaxation. Also, it is clear from these results that even weakly bound adatoms like hydrogen can produce quantitatively high ECI uncertainty.

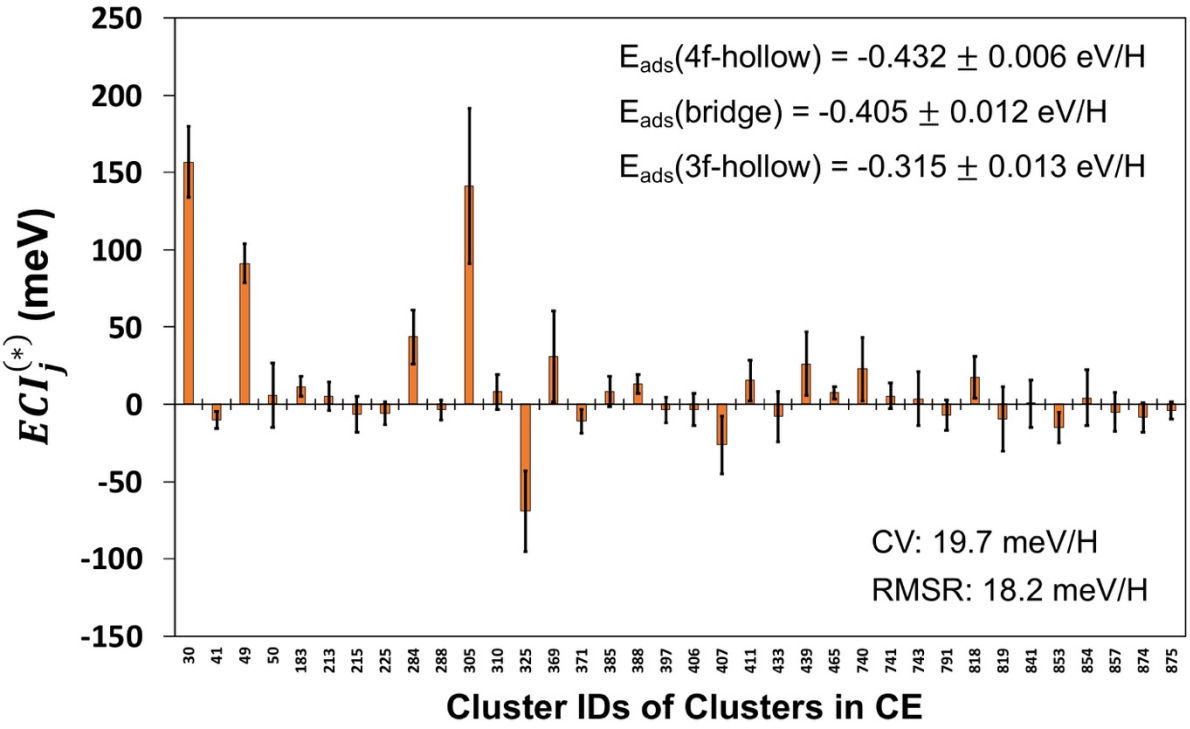


Figure 8. ECIs with their corresponding ECI errors for the optimized CE of the H/Fe(100) optB88-vdW system. Due to magnitude discrepancy, the isolated H/Fe(100) adsorption energies, E_{ads}, (corresponding to clusters #1, #2, and #3) are given in the inset.

3.4. ECI errors vs. Confidence Intervals

To explicitly illustrate the difference between ECI errors and the more customary Student t-test based 95% CIs[37], we present both side-by-side for the relaxed-O/Fe(100) RPBE and fixed-O/Fe(100) RPBE systems in Figure 9. The 95% CI are constructed from Equation 6:

$$95\%\ CI_j = t^{-1}(N - x; 0.95)\frac{RMSR}{\sigma_j} \tag{6}$$

where $t^{-1}(N - x; 0.95)$ is the inverse cumulative Student-t distribution with $N - x$ degrees of freedom and a bound of 0.95, N is the number of structures in the training set, x is the number of clusters specified in the CE, and σ_j is the square root of the variance, σ_j^2, in cluster j which is found from

$$\sigma_j^2 = [(A^T A)^{-1}]_{jj} \tag{7}$$

where A is the $(N \times x)$ design matrix used in regression, $(A^T A)^{-1}$ is the variance-covariance matrix[37], and $[(A^T A)^{-1}]_{jj}$ is its j^{th} diagonal entry.

Figures 9A – 9D show immediately that the ECI errors are not equivalent to their cluster's corresponding 95% CIs and that not even trends are the same between the two. For example, the 95% CIs for the fixed system (Figure 9A – 9B) are indeed smaller than for the relaxed system (Figure 9C – 9D), which is qualitatively similar to the ECI errors, but this is most likely due to the better fit to the data set (i.e., lower RMSR) as will be explained shortly. While the 95% CI appear to capture this basic property, it is important to remember that the form of Equation 6 is based on the (violated) assumptions presented in section 2.1. Thus, we should still not expect these values to be accurate representations of the uncertainty in these systems. This leaves only the question of whether or not our method for determining ECI errors are simply capturing the fixed system's better fit to the CEs as this is the largest factor controlling the CIs.

A comparison of our ECI error estimates to CIs and discussion on the convergence criteria used in the LMO-CV score computation is also included.

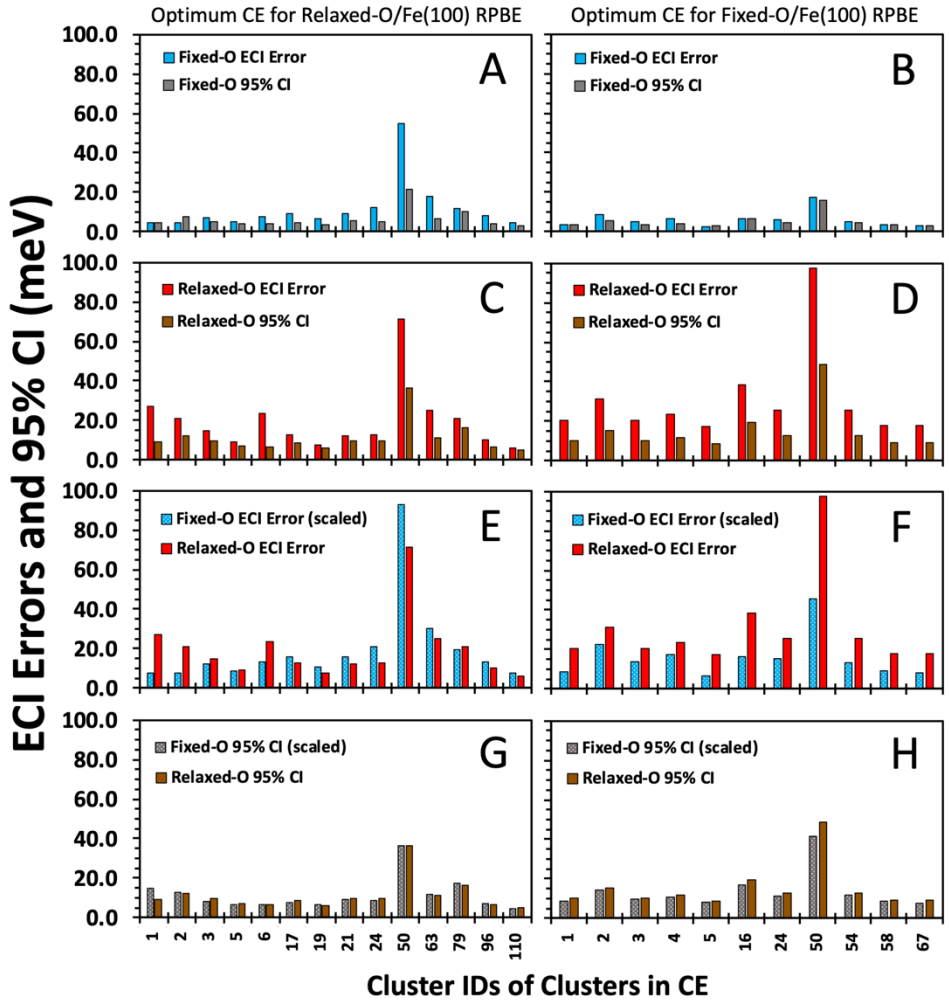


Figure 9. Comparison of ECI errors against 95% CIs for the fixed and relaxed O/Fe(100) RPBE systems using the optimized CE for the relaxed system (left column) and the optimized CE for the fixed system (right column).

As shown in Equation 6, there are three factors that control the value of the CI: the inverse cumulative student-t distribution, the data variance, and the RMSR. With a large enough degree of freedom, $t^{-1}(N-x;0.95)$ approaches the inverse cumulative normal distribution which produces the same output regardless of data set size, and our data set sizes are large

enough that this is the case here. The variance-covariance matrix from which the variance is derived is a function of the structural fingerprints alone, meaning that regardless of the energies of the structures in the dataset, any two data sets with the same structures (say one where the atoms are allowed to relax and one where they are fixed) will output the same variance. The O/Fe(100) data sets here are not exact replicas of each other, but the number of dissimilar structures is quite low. The RMSR, it would therefore seem, is the main controller of the 95% CIs here.

We can then ask if this dependence on the RMSRs is also true of our ECI errors or if they in fact capture something more. To answer this question, we scale up the fixed system ECI errors such that its RMSR matches that of the relaxed system and present the results in Figure 9E – 9F. This is also done to the 95% CIs and presented in Figure 9G – 9H to show the degree of control of the RMSR. As expected, the fixed system 95% CI is quite nearly a perfect match to the relaxed system 95% CI after being scaled up, especially when the relaxed system optimum LG CE is used (Figure 9G). There is clearly more influence from the variance when the optimum LG CE for the fixed system is used (Figure 9H), but the match is still quite significant. This is not true at all for the ECI errors in Figure 9E – 9F. In fact, even despite the RMSR scaling, the nearest neighbor ECI errors (clusters 2 – 6) for the fixed system are still significantly smaller than for the relaxed system, showing that the ECI errors capture the fundamental ideality in the fixed system that is being lost in the relaxed system. The scaled ECI errors for the higher body clusters (clusters 17 – 110) in the fixed system (Figure 9E) are generally significantly larger than the ECI errors in the relaxed system when the fixed system is fit to a non-optimal CE (i.e., the optimum CE for the relaxed system). This suggests that the better fit to the data (i.e., the lower RMSR) of the fixed system is what is primarily being reflected in the overall lower ECI errors in

the higher body clusters seen in the fixed system CEs when a non-optimal CE is used. However, when the optimum CE for the fixed system is used, *every* scaled ECI error is smaller than for the relaxed system. This suggests that there is some connection between the minimizing of the LMO-CV score (leading to the optimum CE) and what property of the system is being primarily captured by the ECI errors. This is perhaps unsurprising since the ECI errors are found from the construction sets used in the determination of the LMO-CV score. Regardless, it is clear that our ECI errors capture the fundamental non-ideality of the relaxed-O/Fe(100) system in a way that is truly different from what CIs might provide.

3.5 Practical Considerations: Effect of Convergence Criteria

In the preceding sections, we have presented data taken from the numerous calculations of LMO-CV scores (and associated ECI errors) using an excessively stringent convergence criteria ("CV tolerance") of 10^{-8} eV to avoid any potential issues of convergence. However, this is a burdensome tolerance criterion resulting in computationally onerous time requirements and choosing a lower tolerance would increase the practicality of our method. A lower tolerance in effect lowers the number of training set splits needed to converge the LMO-CV score ("M" in Figure 1) and thus the amount of time required to arrive at a suitable value. Furthermore, we would expect that the stochastic nature of the training set splits will result in some uncertainty in the final overall value of the CV score and ECI errors but that increased splits will reduce this uncertainty. This deserves explicit testing. To do this, we specify a simple, non-optimal CE containing clusters 1 through 10 for the relaxed-H/Fe(100) optB88-vdW system and then calculate the LMO-CV score and ECI errors 50 times for each of a range of CV tolerances. The results are shown in Figures 10 and 11.

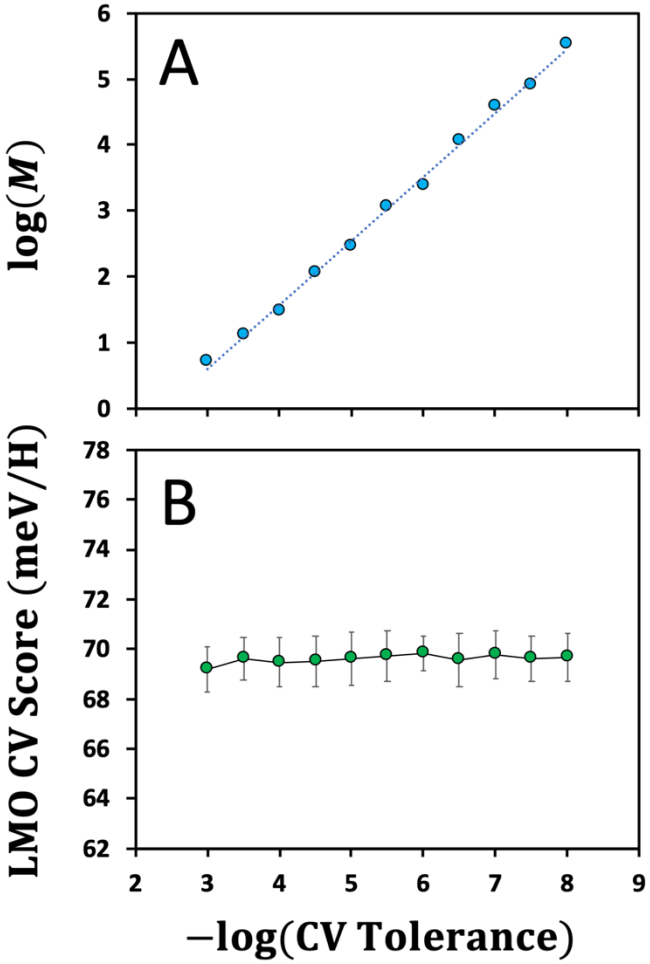


Figure 10. (A) The average number of training set cuts required to converge the LMO-CV score of a set CE for the relaxed-H/Fe(100) optB88-vdW system to the tolerance specified on the x-axis (units in eV) over 50 separate runs of the algorithm presented in Figure 1, and (B) the corresponding average LMO-CV score with two standard deviations of the 50 runs shown as error bars.

Figure 10A shows the relationship between the CV tolerance (in eV) and the average number of splits required to reach that tolerance, M, and Figure 10B shows the corresponding average LMO-CV score with error bars indicating two standard deviations of the 50 LMO-CV scores calculated at each CV tolerance. There is a very linear positive dependence between the logarithm of the (average) number of splits and the negative logarithm of the CV tolerance selected demonstrating that a magnitude decrease in the CV tolerance results in a magnitude increase in the number of splits (and thus time) required. Remarkably, the calculated LMO-CV

scores are largely stable across all convergence criteria and the uncertainty (error bars) remains effectively the same with no appreciable reduction as the CV tolerance is decreased from 10^{-3} eV to 10^{-8} eV (left to right in both Figures 10 and 11). This LMO-CV score uncertainty is approximately ±1 meV/H, which means our method of calculating the LMO-CV score is sufficiently reliable even at the least stringent convergence criteria for this system. However, this is not true for the ECI errors.

In Figure 11, the ECI errors for each cluster in the CE used, clusters 1 – 10, as a function of the set CV tolerance for LMO-CV score convergence. The ECI error values are not as stable across the range of CV tolerances as was the case for the LMO-CV score, and as the CV tolerance is decreased (made more stringent), the maximum uncertainty in their values decreases significantly, appearing to converge to less than ±1 meV by a CV tolerance of $10^{-5.5}$ eV and less than ±0.5 meV by 10^{-7} meV. For the purpose of determining an ECIs significance, a ±1 meV uncertainty is more than sufficient, and since the certainty in the LMO-CV score is ±1 meV/H regardless of the CV tolerance chosen, we would regard a CV tolerance of $10^{-5.5}$ eV a reasonable default value—assuming that the H/Fe(100) system is a representative, worst case scenario for ECI errors. Importantly, at a CV tolerance of $10^{-5.5}$ eV, only ~1000 training set splits are typically required, and in our lab, this takes less than a couple of seconds to achieve. While this result is encouraging, this type of convergence analysis should be performed as a general rule for the individual system of interest.

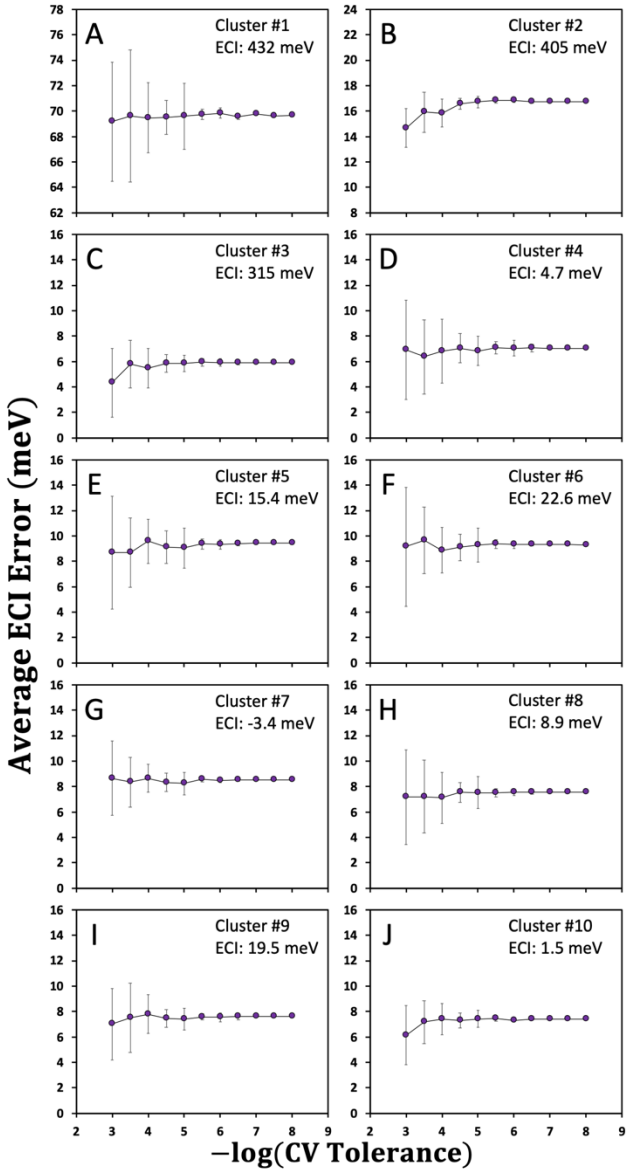


Figure 11. Average ECI errors for clusters within a set CE for the relaxed-H/Fe(100) optB88-vdW system found over 50 separate calculations of the LMO-CV score at the specified tolerance. Two standard deviations of the 50 runs are shown as error bars demonstrating at which tolerance the ECI errors converge. The corresponding ECI values for these clusters are given in the inset. Note that while the bounds change, the y-axes in panels A – J are the same size (16 meV).

3.6 Practical Considerations: Intrinsic Bias in the Training Set

When removing bias from the selection of structures as outlined in Figure 2, we have tacitly assumed that the training set has no *intrinsic* bias. This is a different conceptualization of bias as defined in Equation 5, which we might term an *extrinsic* bias (in that it comes from outside the dataset itself). We have addressed extrinsic bias by ensuring the frequency with

which each structure enters the validation set during the LMO-CV score algorithm is as equal as possible. Therefore, if there is any intrinsic bias in the training dataset, it will be fully reproduced once all extrinsic bias is removed. Formally, this means that the data set should have no intrinsic bias; meaning training data should ideally be made up of a completely random set of structures during structure selection. While a more sophisticated algorithm might be devised to remove any outstanding intrinsic bias, it is outside the scope of this work. However, if this were attempted, the issue of cluster correlation would need to be addressed. For instance, if it is found that a 4-body cluster of some kind were underrepresented in the training data, adding a new structure to the training data that contains this 4-body cluster would necessarily also add the 4-body cluster's constituent 1-, 2-, and 3-body clusters, thereby unintentionally overrepresenting those clusters. This seems to imply that there will always be an intrinsic bias toward lower-body clusters. It is very difficult to tell if such an intrinsic bias is aberrant or not. Since it is a seemingly natural characteristic, perhaps such bias makes the predicted ECI errors more "real," and how would we know the difference? These are open questions deserving of future study. Until then, we must advise that all training data be selected as randomly as possible, and that some amount of training data be set aside to externally cross validate the final CE to ensure its predictiveness has not been compromised by any extant intrinsic bias.

4. Conclusions

We have demonstrated here a method for quantifying the uncertainty in CE ECIs free from major statistical assumptions. We provide the expression and process needed to calculate these "ECI errors" and have demonstrated their use in specifically quantifying the effect of geometric relaxations and functional used in the underlying structural data. As has been described previously, geometric relaxations do indeed cause greater degrees of uncertainty in the

quality of the ECIs, and our method shows quantitively that this corresponds to a general doubling of the ECI errors in CEs built from data where relaxations were allowed as compared to a model where atoms were fixed in ideal lattice positions. This result does not mean that the energetic consequences of relaxations cannot be captured in CEs but does point to the need for being able to quantify the degree of uncertainty that results from them. ECI errors were also able to reveal that the effect of DFT functional is primarily to change the adsorption energy and 1st nearest neighbor ECIs, potentially precluding the need to calculate whole new data sets to build CEs for systems when only the functional is changed. This would allow for the quick screening of functionals when attempting to build CEs.

ECI errors are calculated as part of our recently implemented convergent version of the LMO-CV score, and the worst case uncertainty in the LMO-CV score due to the stochastic nature of the method was shown to be ±1 meV/adsorbate for the H/Fe(100) system while the uncertainty in the ECI errors can be brought to below ±1 meV using suitable convergence criteria. We correspondingly advise that these convergence criteria be tested for any new system of interest. While a LMO-CV score uncertainty of ±1 meV/adsorbate is very sufficient for determining the predictiveness of the vast majority of CEs, it should be noted that we have found that a nonnegligible subset of CEs have LMO-CV scores well within 1 meV/adsorbate of the optimum LMO-CV score found when using AMALGM. ECI errors present an opportunity to fully differentiate this subset of structures by filtering out CEs with insignificant clusters or otherwise give weight to those with the smallest ECI errors. We are actively considering this possibility in our own work.

Overall, the ability to calculate ECI errors as shown here permits researchers to go beyond qualitative descriptions of CE suitability. With this tool, as was done here, direct

comparisons between systems can be performed and conclusions made about their CE's reliability. We posit that such analyses will guide development of CEs with minimal uncertainty and maximal transparency.

Supporting Information

I. Density Functional Theory (DFT) Calculation Parameters for the Fixed-O/Fe(100) Systems. II. Unweighted ECI Error Deviations.

Acknowledgements

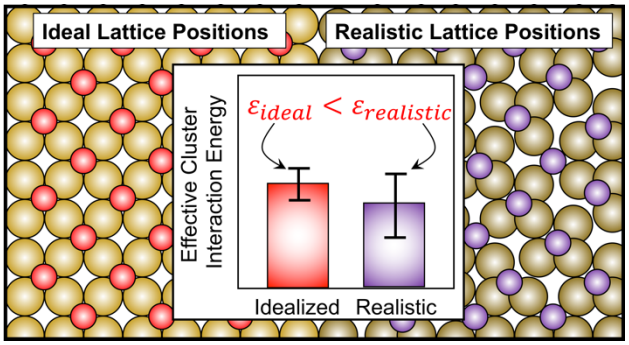
Work at WSU was primarily funded by the U.S. Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Biosciences and Geosciences (DE-SC0014560). It was also funded by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1347973. G.C., Seattle Chapter ARCS Fellow and NSF Graduate Research Fellow, gratefully acknowledges financial support from the Achievement Rewards for College Scientists foundation as well as the National Science Foundation Graduate Research Fellowship Program. This research also used resources from the Center for Institutional Research Computing at Washington State University. PNNL is a multi-program national laboratory operated for the U.S. DOE by Battelle.

REFERENCES

(1) Sanchez, J. M.; Ducastelle, F.; Gratias, D. Generalized Cluster Description of Multicomponent Systems. *Physica A.* **1984**, *128*, 334-350.
(2) Sanchez, J. M. Cluster Expansions and the Configurational Energy of Alloys. *Phys. Rev. B.* **1993**, *48*, 14013-14015.
(3) Sanchez, J. M. Cluster Expansion and the Configurational Theory of Alloys. *Phys. Rev. B.* **2010**, *81*, 224202.
(4) Sanchez, J. M. Foundations and Practical Implementations of the Cluster Expansion. *J. Phase. Equilibria Diffus.* **2017**, *38*, 238-251.
(5) Wei, S. H.; Ferreira, L. G.; Zunger, A. First-Principles Calculation of Temperature-Composition Phase Diagrams of Semiconductor Alloys. *Phys. Rev. B.* **1990**, *41*, 8240-8269.
(6) de Gironcoli, S.; Giannozzi, P.; Baroni, S. Structure and Thermodynamics of Si_xGe_{1-x} Alloys from *ab initio* Monte Carlo Simulations. *Phys. Rev. Lett.* **1991**, *66*, 2116-2119.
(7) Lu, Z. W.; Wei, S. H.; Zunger, A. Comment on `Origins of Compositional Order in NiPt Alloys`. *Phys. Rev. Lett.* **1992**, *68*, 1961-1961.
(8) Lu, Z. W.; Wei, S. H.; Zunger, A.; Frota-Pessoa, S.; Ferreira, L. G. First-Principles Statistical Mechanics of Structural Stability of Intermetallic Compounds. *Phys. Rev. B.* **1991**, *44*, 512-544.
(9) Laks, D. B.; Ferreira, L. G.; Froyen, S.; Zunger, A. Efficient Cluster Expansion for Substitutional Systems. *Phys. Rev. B.* **1992**, *46*, 12587-12605.
(10) Asta, M.; McCormack, R.; de Fontaine, D. Theoretical Study of Alloy Phase Stability in the Cd-Mg System. *Phys. Rev. B.* **1993**, *48*, 748-766.
(11) Silverman, A.; Zunger, A.; Kalish, R.; Adler, J. Effects of Configurational, Positional and Vibrational Degrees of Freedom on an Alloy Phase Diagram: a Monte Carlo Study of $Ga_{1-x}In_xP$. *J. Phys. Condens.* **1995**, *7*, 1167.
(12) Khachaturyan, A. G. *Theory of Structural Transformations in Solids*; Courier Corporation, 2013.
(13) Sanchez, J. M. Reply to "Comment on 'Cluster Expansion and the Configurational Theory of Alloys'". *Phys. Rev. B.* **2017**, *95*, 216202.
(14) Wolverton, C.; de Fontaine, D.; Dreyssé, H. Global Volume Relaxations and Phase Stability in Disordered Pd-Rh Alloys. *Phys. Rev. B.* **1993**, *48*, 5766-5778.
(15) Ozoliņš, V.; Häglund, J. First-Principles Study of Effective Cluster Interactions and Enthalpies of Formation of Substoichiometric VC_{1-x}. *Phys. Rev. B.* **1993**, *48*, 5069-5076.
(16) Nguyen, A. H.; Rosenbrock, C. W.; Reese, C. S.; Hart, G. L. W. Robustness of the Cluster Expansion: Assessing the Roles of Relaxation and Numerical error. *Phys. Rev. B.* **2017**, *96*, 014107.
(17) Einstein, T. L.; Sathiyanarayanan, R. *Multisite Interactions in Lattice-Gas Models*. In *Nanophenomena at Surfaces: Fundamentals of Exotic Condensed Matter Properties*, M. Michailov, Ed. Springer Berlin Heidelberg: Berlin, Heidelberg, 2011; pp 19-37.
(18) Chen, W.; Schmidt, D.; Schneider, W. F.; Wolverton, C. First-Principles Cluster Expansion Study of Missing-Row Reconstructions of FCC (110) Surfaces. *Phys. Rev. B.* **2011**, *83*, 075415.
(19) Li, W.; Huang, L.; Evans, J. W.; Han, Y. Submonolayer Ag films on Fe(100): A First-Principles Analysis of Energetics Controlling Adlayer Thermodynamics and Kinetics. *Phys. Rev. B.* **2016**, *93*, 155416.

(20) Lai, K. C.; Han, Y.; Spurgeon, P.; Huang, W.; Thiel, P. A.; Liu, D.-J.; Evans, J. W. Reshaping, Intermixing, and Coarsening for Metallic Nanocrystals: Nonequilibrium Statistical Mechanical and Coarse-Grained Modeling. *Chemical Reviews* **2019**, *119*, 6670-6768.
(21) Sathiyanarayanan, R.; Stasevich, T. J.; Einstein, T. L. Sensitivity of Short-Range Trio Interactions to Lateral Relaxation of Adatoms: Challenges for Detailed Lattice-Gas Modeling. *Surf. Sci.* **2008**, *602*, 1243-1249.
(22) Han, Y.; Stoldt, C. R.; Thiel, P. A.; Evans, J. W. Ab Initio Thermodynamics and Kinetics for Coalescence of Two-Dimensional Nanoislands and Nanopits on Metal (100) Surfaces. *J. Phys. Chem. C.* **2016**, *120*, 21617-21630.
(23) Binder, K.; Landau, D. P. Square Lattice Gases with Two- and Three-Body Interactions: A Model for the Adsorption of Hydrogen on Pd(100). *Surf. Sci.* **1981**, *108*, 503-525.
(24) Binder, K. Theory of First-Order Phase Transitions. *Rep. Prog. Phys.* **1987**, *50*, 783.
(25) Dobson, A. J.; Barnett, A. *An Introduction to Generalized Linear Models*; Chapman and Hall/CRC: Boca Raton, FL, 2008.
(26) Collinge, G.; Groden, K.; Stampfl, C.; McEwen, J.-S. Formulation of Multicomponent Lattice Gas Model Cluster Expansions Parameterized on Ab Initio Data: An Introduction to the Ab Initio Mean-Field Augmented Lattice Gas Modeling (AMALGM) Code. *J. Phys. Chem. C.* **2019**, *124*, 2923-2938.
(27) Baumann, K. Cross-Validation as the Objective Function for Variable-Selection Techniques. *Trends Analyt. Chem.* **2003**, *22*, 395-406.
(28) Shao, J. Linear Model Selection by Cross-validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486-494.
(29) Zhang, P. Model Selection Via Multifold Cross Validation. *Ann. Statist.* **1993**, *21*, 299-313.
(30) Bray, J.; Collinge, G.; Stampfl, C.; Wang, Y.; McEwen, J.-S. Predicting the Electric Field Effect on the Lateral Interactions Between Adsorbates: O/Fe(100) from First Principles. *Top. Catal.* **2018**, *61*, 763-775.
(31) Hensley, A. J. R.; Collinge, G.; McEwen, J.-S. Coverage Dependent Adsorption of Hydrogen on Fe(100): Determining Catalytically Relevant Surface Structures via Lattice Gas Models. *J. Phys. Chem. C.* **2018**, *124*, 7254-7266.
(32) Bray, J.; Hensley, A. J. R.; Collinge, G.; Che, F.; Wang, Y.; McEwen, J.-S. Modeling the Adsorbate Coverage Distribution over a Multi-Faceted Catalytic Grain in the Presence of an Electric Field: O/Fe from First Principles. *Catal. Today* **2018**, *312*, 92-104.
(33) van de Walle, A.; Asta, M.; Ceder, G. The Alloy Theoretic Automated Toolkit: A User Guide. *CALPHAD* **2002**, *26*, 539-553.
(34) van de Walle, A.; Ceder, G. Automating First-Principles Phase Diagram Calculations. *J. Phase. Equilibria Diffus.* **2002**, *23*, 348-359.
(35) van de Walle, A. Multicomponent Multisublattice Alloys, Nonconfigurational Entropy and Other Additions to the Alloy Theoretic Automated Toolkit. *CALPHAD* **2009**, *33*, 266-278.
(36) Persson, B. N. J. Ordered structures and phase transitions in adsorbed layers. *Surf. Sci. Rep.* **1992**, *15*, 1-135.
(37) Neter, J.; Kutner, M. H.; Nachtsheim, C. J.; Wasserman, W. *Applied Linear Statistical Models*; Irwin Chicago, 1996; Vol. 4.

TOC Graphic


Ideal Lattice Positions
Realistic Lattice Positions
$\varepsilon_{ideal} < \varepsilon_{realistic}$
Effective Cluster Interaction Energy
Idealized
Realistic