Cluster expansion method for multicomponent systems based on optimal selection of structures for density-functional theory calculations

Atsuto Seko,[1,*] Yukinori Koyama,[2] and Isao Tanaka[2,3]

[1]*Pioneering Research Unit for Next Generation, Kyoto University, Kyoto 606-8501, Japan*

[2]*Department of Materials Science and Engineering, Kyoto University, Kyoto 606-8501, Japan*

[3]*Nanostructures Research Laboratory, Japan Fine Ceramics Center, Nagoya 456-8587, Japan*

(Received 8 August 2009; revised manuscript received 30 September 2009; published 20 October 2009)

The cluster expansion (CE) method has been used to evaluate configurational properties in multicomponent systems based on the density-functional theory (DFT) calculations. Appropriate selections of not only clusters but also structures for DFT calculations (DFT structures) are crucial for the accuracy and the efficiency of the CE. In a conventional procedure to construct the CE, the CE error is reduced mainly through an appropriate selection of clusters. In the present paper, we propose an improved procedure that systematically leads to optimal selections of both clusters and DFT structures. DFT structures are chosen to cover as much of the configurational space as possible. During the iterative process, the predictive power of the out-of-sample structures can be increased up to the accuracy that is required to describe alloy thermodynamics. We apply the procedure to configurational behaviors in a simple MgO-ZnO pseudobinary system and in a complex $MgAl_2O_4$ system. The CE error is reduced in both systems, in particular, in the complex system, thereby significantly improving configurational properties at high temperatures compared with the conventional CE procedure.

DOI: 10.1103/PhysRevB.80.165122 PACS number(s): 61.50.Ah, 81.30.Dz

I. INTRODUCTION

The properties of materials are strongly dependent not only on their chemical composition but also on the configurations of solute atoms and/or point defects. Quantitative knowledge of the configuration-dependent properties is, therefore, essential for materials design. The cluster expansion (CE) method[1–3] has been widely used to describe configurational properties. Increases in computational power and advances in numerical techniques enable us to perform a large set of systematic first-principles calculations based on the density-functional theory (DFT) combined with CE calculations. In the CE formalism, the configurational properties are expanded using a basis set of clusters. The expansion coefficients are called effective cluster interactions (ECIs). The configurational energy, E, in a binary system can be expressed using the ECIs, V, and the pseudospin configurational variable, σ_i, for the respective lattice site i as

$$E = V_0 + \sum_i V_i\sigma_i + \sum_{i,j} V_{ij}\sigma_i\sigma_j + \sum_{i,j,k} V_{ijk}\sigma_i\sigma_j\sigma_k + \cdots$$
$$= \sum_\alpha V_\alpha \cdot \varphi_\alpha, \tag{1}$$

where φ_α is called the correlation function of cluster α. In a simplified method, the number of ordered structures for DFT calculations, N, is set to be the same as the number of ECIs, m. Then all ECIs can be determined analytically.[4] An alternative method of CE uses a least-squares technique[5,6] or a linear-programming method[7] to determine a set of ECIs from $N > m$ structures. In the past, the set of clusters was predetermined before constructing the CE. Currently, the set of clusters is usually optimized as the configurational properties are reconstructed using a small number of clusters.[8–10] Since the set of optimal clusters and the set of necessary DFT structures are related, they can be selected simultaneously during an iterative procedure.[11–13] Figure 1 shows a flowchart of an iterative procedure. Initially, a small number of ordered structures are prepared as inputs. An optimal set of clusters that minimizes the cross-validation (CV) score[8,14] is searched for from the pool of candidate clusters using general minimization algorithms.[10] Since the CV score is calculated using the sample inputs, the predictive power for structures far from the inputs is generally lower than that for structures near the inputs. Therefore, the quality of the trial CE should be validated using out-of-sample structures. In the conventional scheme, the predicted ground states and near-ground states are used as additional structures for validation. The procedure is repeated until the predicted ground states and near-ground states converge. The CE error is reduced through the appropriate selection of clusters and the validation of the trial CE using out-of-sample structures.

The conventional iterative procedure has been successfully used to search for the ground and near-ground-state structures. It has also been used to predict properties at finite temperatures. However, it must be noted that the CE error over the whole range of configurations is not necessarily fully reduced by the conventional procedure. In other words, the converged CE obtained by the conventional scheme is not necessarily the optimal CE, in which the CE error is fully reduced. Figure 2 schematically illustrates the situation. The energy of a structure near the input structures can be predicted with a small uncertainty. On the other hand, the energy of a structure far from the input structures cannot be predicted precisely owing to the larger uncertainty of the CE. When input structures with strong correlations are prepared, the predicted energy of a structure far from the input structures has a particularly large uncertainty. To calculate the configuration-dependent properties at finite temperatures, for example, the configurational free energy as a function of temperature, the error of CE over the whole range of configurations, including all excited states, should be carefully examined. It is, therefore, necessary to avoid the localization

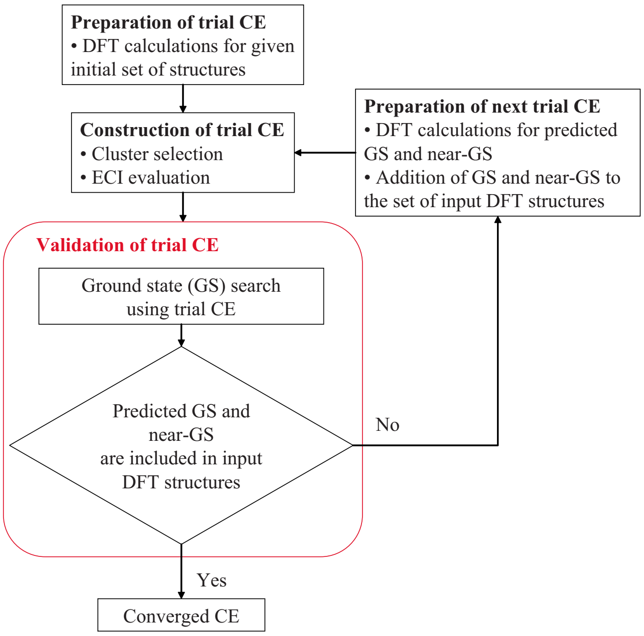
Preparation of trial CE
• DFT calculations for given initial set of structures
Construction of trial CE
• Cluster selection
• ECI evaluation
Preparation of next trial CE
• DFT calculations for predicted GS and near-GS
• Addition of GS and near-GS to the set of input DFT structures
Validation of trial CE
Ground state (GS) search using trial CE
Predicted GS and near-GS are included in input DFT structures
No
Yes
Converged CE

FIG. 1. (Color online) Flowchart of a conventional procedure for constructing the CE.

and strong correlation of input structures in the configuration space.

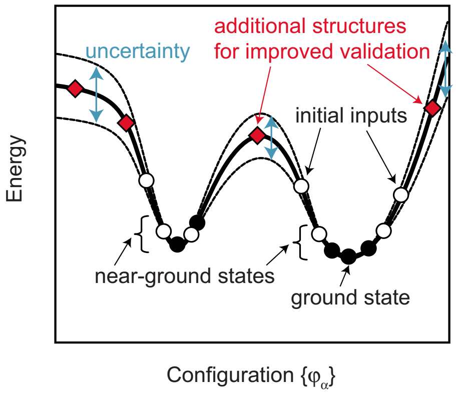

FIG. 2. (Color online) Schematic of CEs constructed from the conventional procedure. Initial input DFT structures are shown by the open circles. Input DFT structures obtained from ground-state searches using trial CEs are shown by the closed circles. Structures far from the input structures have a large uncertainty of predicted energy. To reduce the CE error over the whole range of configurations, additional structures, as shown by the closed diamonds, are required for improved validation of the CE.

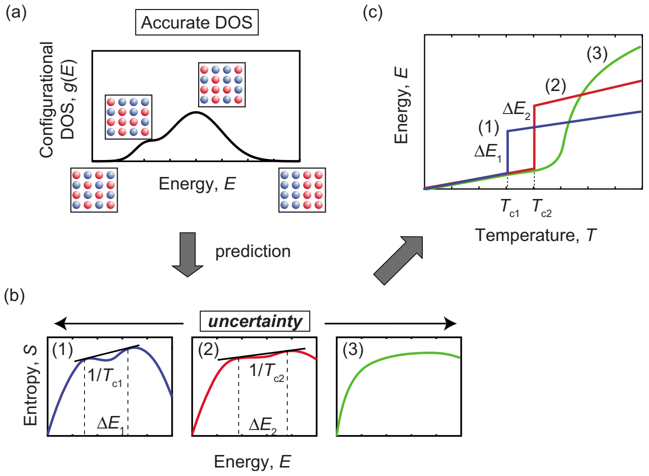

FIG. 3. (Color online) (a) Schematic of the configurational DOS in a binary system. (b) Configurational entropy, which can be evaluated from three types of configurational DOS. The gradient of the common tangent line gives the inverse of the first-order phase-transition temperature. (c) Temperature dependence of the configurational energy.

Order-disorder transition behavior is a typical example of a property that is sensitive to the CE error. The situation is schematically explained in Fig. 3. Assuming that the energies of all different configurations in a finite-size cell are known, the configurational density of states (DOS), $g(E)$, can be made as shown in Fig. 3(a). The CE error can be defined as the difference between the DOS obtained by the DFT and that obtained by the CE. The error is reflected in the uncertainty of the DOS or the configurational entropy, S, given by $S(E)=k_B \ln[g(E)dE]$, where k_B denotes the Boltzmann constant. Since temperature is defined as the reciprocal of the partial derivative of S with respect to the internal energy, $1/T=(\partial S/\partial E)_{V,N}$, the shape of the DOS determines the phase-transition behavior. When a common tangent line is drawn for S as shown in Fig. 3(b)-(1) and (2), two internal energy states can coexist at temperature T_c as shown in Fig. 3(c), thereby indicating a first-order transition. On the other hand, such an abrupt transition disappears when a common tangent line cannot be drawn as shown in Fig. 3(b)-(3). In this case, a second-order transition occurs as shown in Fig. 3(c). When the error of CE over the whole range of configurations is larger, the uncertainty of the predicted phase-transition behavior becomes larger.

In the present study, we propose a modification to the conventional iterative procedure that leads to a reduction in the CE error in a systematic manner. We implement a validation scheme for a trial CE. In the conventional procedure, the validation is performed by examining only the ground and near-ground states. In our validation scheme, additional structures are chosen for the validation of the trial CE so as to cover as much of the configurational space as possible (see Fig. 2) based on the statistics described in Sec. II. The converged CE obtained from our iterative procedure is optimal CE with a fully reduced error.

Most previous CE calculations were limited to fcc or bcc-based binary alloys. Recently more complex structures with nonclosed packed structures have been investigated.[15–20] Our validation scheme proposed in the present study is, in particular, useful for such complex systems where the number of symmetrically independent clusters is much larger than that of fcc or bcc-based binary alloys. In this paper, we discuss two examples that were studied previously by a conventional method.[17,21] The first example is pseudobinary rocksalt-based MgO-25%ZnO, in which the disordering of the fcc-based cation sublattice is examined. The other example is the order-disorder transition of $MgAl_2O_4$ spinel. The temperature dependence of the order parameter in the complex spinel structure is carefully examined.

II. METHODOLOGY

A. Selection of additional DFT structures for validation of a trial CE

Generally, ECIs are evaluated using the least-squares procedure. In elementary multiple linear regression analysis, the variance of the energy of structure i predicted from a CE with m ECIs can be expressed using the input set of structures as[22]

$$\mathrm{Var}[E_{\mathrm{CE}}(i)] = [X_i \cdot (X^{\mathrm{T}}X)^{-1} \cdot X_i^{\mathrm{T}}]\sigma^2, \tag{2}$$

where $E_{\mathrm{CE}}(i)$ is the predicted energy of structure i and σ^2 denotes the variance of the error in the given population. Structure i can be described by the row vector of its correlation functions X_i including the empty cluster. The input set of structures is identified by the $N \times m$ matrix of the correlation functions X. The matrix $(X^{\mathrm{T}}X)^{-1}$ is called the precision matrix. For all structures in the population, the mean variance of the predicted energy is expressed as

$$\begin{aligned}\langle \mathrm{Var}[E_{\mathrm{CE}}(i)] \rangle &= \frac{1}{N_{\mathrm{all}}} \sum_{j=1}^{N_{\mathrm{all}}} [X_j \cdot (X^{\mathrm{T}}X)^{-1} \cdot X_j^{\mathrm{T}}]\sigma^2 \\ &= \{\mathrm{tr}[(X^{\mathrm{T}}X)^{-1}\Sigma] + \mu (X^{\mathrm{T}}X)^{-1} \mu^{\mathrm{T}}\}\sigma^2 \\ &= \Lambda \cdot \sigma^2, \end{aligned} \tag{3}$$

where N_{all} is the total number of structures in the population, Σ is the $m \times m$ covariance matrix of the correlation functions of the structures in the population, and μ is the m-dimensional vector of the mean correlation functions of the structures in the population. The distribution of all structures in the population is characterized by Σ and μ. Approximate values of Σ and μ can be evaluated for many random configurations within a finite number of atoms. Λ can be evaluated using Σ and μ for each set of correlation functions X. It should be emphasized that DFT calculations for many configurations are not required to evaluate Λ. In the present study, we propose a validation method for the trial CE using out-of-sample additional structures that leads to a reduction in the variance of the CE energy. The additional structures are chosen so as to significantly decrease Λ. The CE error can be systematically and efficiently reduced by decreasing the variance of the CE energy. We hereafter call these additional structures "probe structures."

The precision matrix includes the effects of the correlation and variance of the input structures. If the distributions of the correlation functions are narrow or strongly correlated, the elements of the precision matrix become large. This leads to a large variance of the predicted energy. Probe structures can improve the CE by increasing the variances of the distributions and decreasing the correlations between the distributions. Eventually, an input set of nearly independent structures with a large variance is produced, which is used to obtain the optimal CE.

The approximated mean variance of the predicted energy was reported in Ref. 8 as

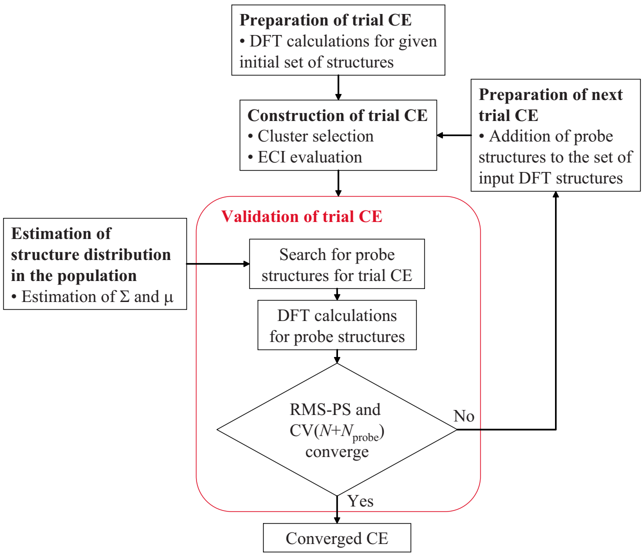
Preparation of trial CE
• DFT calculations for given initial set of structures
Construction of trial CE
• Cluster selection
• ECI evaluation
Preparation of next trial CE
• Addition of probe structures to the set of input DFT structures
Estimation of structure distribution in the population
• Estimation of Σ and μ
Validation of trial CE
Search for probe structures for trial CE
DFT calculations for probe structures
RMS-PS and CV(N+N_probe) converge
No
Yes
Converged CE

FIG. 4. (Color online) Flowchart of our procedure based on the improved validation of trial CEs using probe structures.

$$\langle \mathrm{Var}[E_{\mathrm{CE}}(i)] \rangle \propto \mathrm{tr}(X^{\mathrm{T}}X)^{-1} \cdot \sigma^2. \tag{4}$$

Comparing Eq. (3) with Eq. (4), Σ and μ in Eq. (3) are found to play a role of weighting factors for clusters depending on the distributions of the structures in the population. For example, when the correlation functions of two clusters have different distributions, the cluster with the wider distribution is more significant. Equation (4) is exact only when the structures are distributed spherically and isotropically in the configurational space.

In general, the DFT energies contain the numerical noise, which arises from atomic relaxations, the k-point mesh, and the size of the basis. As pointed out in Ref. 23, the numerical noise in DFT energies greatly influences the accuracy of the CE when many ordered structures are required to extract the ECIs, i.e., when the CE constructed from a small number of DFT structures has a large value of Λ in Eq. (3). When there are no systematic errors from the cluster truncation, the CE error originates only from the numerical noise in DFT energies. The CE constructed from the input set of DFT structures with a large value of Λ has a larger error than the numerical noise. By decreasing Λ by adding the probe structures, the CE error can be decreased to the numerical noise.

B. Iterative procedure

We have proposed the use of probe structures to validate the trial CE. A flowchart of our iterative procedure is shown in Fig. 4. In principle, our procedure is different from the conventional procedure only in the way that the additional structures are selected. In our procedure, the probe structures are selected so as to decrease the variance of the CE energy. The reduction in the CE error over a wide range of configurations is the prime objective of the use of probe structures. The structure that contributes the most to reducing Λ is chosen first. Structures are then selected by order of merit. Our procedure can, therefore, optimize the CE more efficiently than the conventional procedure, in which the ground and near-ground states are used for the validation. Therefore, the ground-state search is omitted in the present procedure in order to show the validity of the present procedure. In a simple system, the ground-state search need not be omitted. The ground-state search can be combined with the present procedure. Confirming the ground-state behavior improves the reliability of the CE. In systems with a large unit cell the exhaustive ground-state search, in which the energies of all configurations are calculated, cannot be performed since the number of configurations increases exponentially with the system size.

Our procedure is divided into five stages. (i) We prepare the initial set of input structures and calculate their DFT energies. (ii) We construct a trial CE by searching for an optimal set of clusters that minimizes the CV score. Here, we minimize the leave-one-out CV score by using the genetic algorithm[10] for the selection. The predictive power of the CE is improved by using the leave-many-out CV scheme. Even if the leave-many-out CV score is used as the prediction score, the predictive power for structures far from the inputs is generally lower than that for structures near the inputs, since the leave-many-out CV score is calculated using the sample inputs. Therefore, the quality of the trial CE, which minimizes the leave-many-out CV score, should be validated using out-of-sample structures. (iii) We search for N_{probe} probe structures that reduce the coefficient Λ in Eq. (3) by the greatest amount. (iv) Then, the DFT energies of the probe structures are calculated. (v) We validate the trial CE using the probe structures. A candidate score used in the validation of the trial CE is simply the root-mean-square (RMS) difference between the DFT and CE energies of the probe structures (RMS-PS). RMS-PS is expressed by

$$(\text{RMS-PS})^2 = \frac{1}{N_{\text{probe}}} \sum_{i}^{N_{\text{probe}}} |E_{\text{CE}}(i) - E_{\text{DFT}}(i)|^2, \qquad (5)$$

where N_{probe} denotes the number of probe structures. $E_{\text{CE}}(i)$ and $E_{\text{DFT}}(i)$ denote the CE and DFT energies of probe structure i, respectively. If RMS-PS is much larger than the CV score, the trial CE fails. However, it is expected to be difficult to rigorously validate the trial CE using RMS-PS because it is more biased than the CV score. A more suitable score for verifying the convergence of the CE is the CV score for $N+N_{\text{probe}}$ structures with clusters selected from N structures, i.e., CV($N+N_{\text{probe}}$). If CV($N+N_{\text{probe}}$) is larger than CV(N) evaluated in stage (ii), the trial CE is considered to fail. The probe structures are then added to the input set and a new iterative step starting from stage (ii) is performed using $N+N_{\text{probe}}$ structures. The iterative steps are repeated until CV($N+N_{\text{probe}}$) converges. In this study, we use both RMS-PS and CV($N+N_{\text{probe}}$) to validate the trial CE.

C. Structure selection in complex systems

Generally speaking, it is more difficult to construct an accurate CE when crystal structures are more complex. Therefore, the use of probe structures is more important in complex systems. There are three reasons why it is difficult to construct an accurate CE in a complex system. (i) There are many candidate clusters. The sampling of important structures to increase the accuracy of the CE for all clusters is a difficult task. In simple systems, it is easy to sample independent structures over a wide range of the configuration space. (ii) Many clusters are required in order to express the configurational thermodynamics. A CE with many clusters has a large uncertainty of the predicted energy easily. Therefore, the CE can only attain the required precision using a larger number of input structures. In simple systems, a smaller number of clusters are needed to describe the configurational thermodynamics. (iii) There are no proper guides for the selection of the number of input structures, N, and the number of clusters, m. In simple systems, because many CE calculations can be found in the literature, appropriate values for N and m can be learned. In the absence of such references, N and m should be determined by trial and error. For these reasons, our procedure involving the use of probe structures is much more useful in complex systems.

III. RESULTS AND DISCUSSION

A. Order-disorder transition in pseudobinary MgO-25%ZnO

To examine our procedure in simple systems, we apply the method to the order-disorder transition in simple pseudobinary MgO-25%ZnO. We start from an initial set of 15 typical prototype structures, including A1 (0%, 100%), Ca_7Ge (12.5%, 87.5%), $L1_2$ and $D0_{22}$ (25%, 75%), Ga_3Pt_5 (37.5%, 62.5%), $L1_0$, $L1_1$, Z2, D4, and "40" (50%) structures. DFT calculations are performed by the projector augmented wave method[24,25] within the local-density approximation[26,27] as implemented in the VASP code.[28,29] The plane-wave cutoff energy is set at 350 eV. The total energies converge to less than 10^{-2} meV. The atomic positions and lattice constants are relaxed until the residual forces become less than 10^{-2} eV/Å. The number of clusters used to describe the configurational energy is fixed at $m=11$. The optimal set of clusters is searched for from the pool of 53 clusters up to quadruplets. To validate the trial CE, $N_{\text{probe}}=3$ probe structures that minimize Λ are searched for using the simulated annealing (SA) procedure with $2\times2\times2$ supercells. Σ and μ, which are required when searching for the probe structures, are estimated from 10,000 randomly selected structures within the $2\times2\times2$ expansion of the conventional unit cell (32 sites). After the construction of the optimal CE, finite-temperature thermodynamic properties are evaluated using canonical Monte Carlo (MC) simulations. Supercells for the MC simulations are constructed by the $20\times20\times20$ expansion of the unit cell. The MC simulations are performed over 8,000 MC steps per cation to calculate the thermodynamic averages after equilibration over 10,000 MC steps per cation. We used the CLUPAN code[17,30] in the series of calculations performed to construct the CE and to evaluate the finite-temperature properties.

We first construct the optimal CE. Figure 5 shows the CV scores of the trial CEs obtained in stage (ii) plotted against the number of input structures. For the validations of the trial CEs, RMS-PS and CV($N+N_{\text{probe}}$) are also shown in Fig. 5. To rigorously estimate the actual error of the trial CEs, we perform DFT calculations for 300 randomly chosen configu-

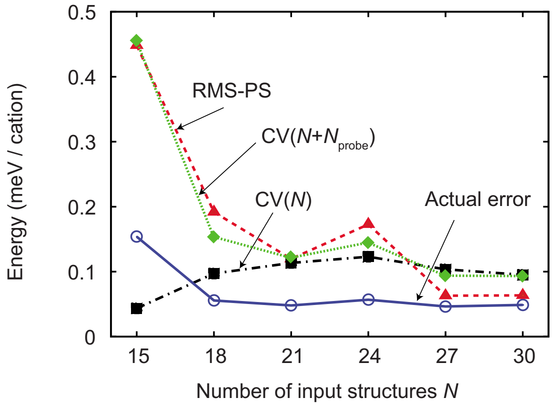

FIG. 5. (Color online) Dependence of the CV score on the number of input structures in the MgO-ZnO system. RMS-PS and CV(N+$N_{\rm probe}$) are also shown. The actual errors of the energies of 300 reference structures predicted from the CEs with N are also shown. The actual error is estimated as $(\Delta E)^2=\frac{1}{N_{\rm ref}}\Sigma_i^{N_{\rm ref}}|E_{\rm CE}(i)-E_{\rm DFT}(i)|^2$, where $N_{\rm ref}$ denotes the number of reference structures used to estimate the error of the CE. $E_{\rm CE}(i)$ and $E_{\rm DFT}(i)$ denote the CE and DFT energies of structure i, respectively. Because the CV score evaluated from N structures without probe structures and that evaluated from N+$N_{\rm probe}$ structures including $N_{\rm probe}=3$ probe structures are calculated using the same set of clusters, CV(N+$N_{\rm probe}$) with N+$N_{\rm probe}$ structures including probe structures are plotted for each value of N on the horizontal axis.

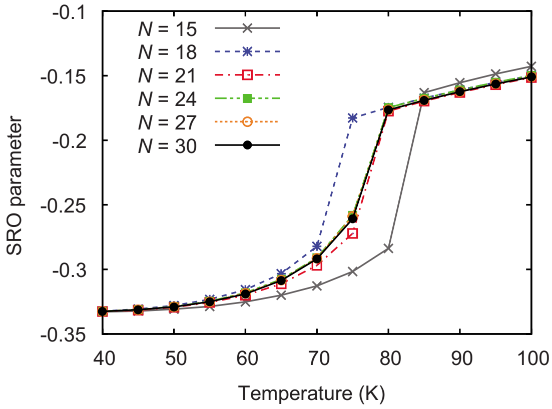

FIG. 6. (Color online) Calculated temperature dependence of the Warren-Cowley SRO of the first NN pair. The ground-state structure is the $D0_{22}$ structure.

rations within the $2\times2\times2$ supercells. The actual error, shown in Fig. 5, is estimated from the RMS difference between the DFT and CE energies of the 300 reference structures. The CV score of the CE constructed from the initial set of 15 structures is evaluated to be 0.04 meV/cation. The CE with $N=15$ predicts the ground state of MgO-25%ZnO to be the $D0_{22}$ structure, which is consistent with the previous result.[21] However, the actual error is 0.15 meV/cation, which is about four times larger than the CV score. Regarding the validation of the CE with $N=15$, RMS-PS and CV (N+$N_{\rm probe}$) are 0.46 and 0.47 meV/cation, respectively, which are also much larger than the CV score. The CV score clearly overestimates the predictive power of the CE with $N=15$. As the number of input structures increases, RMS-PS and CV(N+$N_{\rm probe}$) decrease, and the actual error decreases along with RMS-PS and CV(N+$N_{\rm probe}$). This means that the probe structures are suitable for validating the trial CEs. As can be seen in Fig. 5, CV(N+$N_{\rm probe}$) converges well at the CE with $N=21$. RMS-PS does not converge because it is obtained from $N_{\rm probe}=3$ biased structures, although the values of RMS-PS for $N=21-30$ are smaller than those for N $=15$ and 18. In parallel to the convergence of CV (N+$N_{\rm probe}$), the actual error also converges. However, the CV score has a different value from the actual error when $N=30$ because the former is evaluated from only the biased input DFT energies. The CV score is expected to converge slowly to the actual error upon the addition of more structures.

In all the CEs with $N=15-30$, the ground state is successfully predicted for MgO-25%ZnO to be the $D0_{22}$ structure. The order-disorder transition behaviors predicted from the CEs can be seen in Fig. 6, which shows the calculated temperature dependence of the Warren-Cowley short-range order parameter (SRO) (Ref. 31) of the first-nearest-neighbor (NN) pair for MgO-25%ZnO. The predicted order-disorder transition behavior converges with the CV score and the actual error when $N=21$. Even the CE with 15 input prototype structures can predict the transition temperature with an error of <10 K. The uncertainty of the CE is low in this simple system. This means that the initial set of prototype structures is sufficient to construct an accurate CE with 11 clusters. The procedure based on the addition of probe structures is more useful in complex systems such as $MgAl_2O_4$ than in the system of MgO-ZnO as will be shown in Sec. IV.

B. Order-disorder transition in $MgAl_2O_4$

As an example of a complex system, we investigate the order-disorder transition in $MgAl_2O_4$ spinel oxide between fourfold coordinated tetrahedral and sixfold coordinated octahedral sites in an fcc oxygen sublattice. Spinels with a degree of inversion x that ranges from 0 (normal spinel) to 1 (inverse spinel) are expressed as $(Mg_{1-x}Al_x)[Mg_xAl_{2-x}]O_4$, where the round and square brackets denote the tetrahedral and octahedral sites, respectively. As the temperature increases, spinels have been reported to exhibit disordering by the exchange of tetrahedral and octahedral sites. At the high-temperature limit, the degree of inversion converges to 2/3.

We construct a CE starting from an initial set of 20 highly symmetric (HS) structures, which are ordered by the number of symmetry operations. Because there are no prototype configurations for the spinel reported in the literature, we start with these HS structures in order to cover as much of the configurational space as possible. An initial set of structures with a small number of atoms can also be prepared by using an algorithm proposed recently.[32] The number of clusters used to construct the CE is fixed at 17 including the empty cluster, a point cluster, and five pair clusters. The optimal set of clusters is searched for from the pool of 126 clusters up to quadruplets. $N_{\rm probe}=5$ probe structures are searched for sequentially using the SA within the unit cell of the spinel. Σ and μ, required when searching for the probe structures, are estimated from 10,000 randomly selected configurations

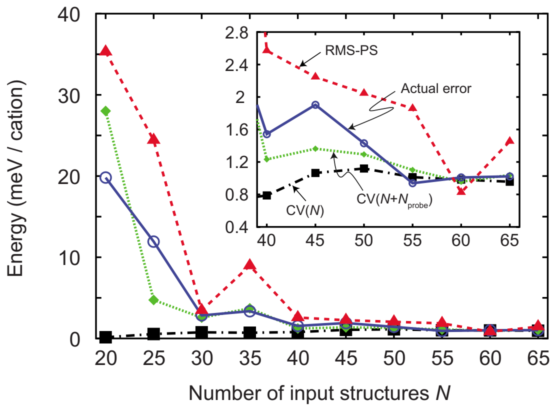

FIG. 7. (Color online) CV score and actual error ΔE along with the number of input structures in $MgAl_2O_4$ spinel. For the validation of trial CEs, RMS-PS and CV(N+N_{probe}) are also shown. The CV(N+N_{probe}) with N+N_{probe} structures including probe structures are plotted for each value of N on the horizontal axis.

within the unit cell. Finite-temperature thermodynamic properties are evaluated using canonical MC simulations with supercells constructed by the 10 × 10 × 10 expansion of the unit cell. The MC simulations are performed over 8,000 MC steps per cation to calculate the thermodynamic averages after equilibration over 10,000 MC steps per cation.

Figure 7 shows the CV scores of the trial CEs. For the validation of the trial CEs, RMS-PS and CV(N+N_{probe}) are also shown in Fig. 7. Figure 7 also shows the actual error estimated from the RMS difference between the DFT and CE energies of the 300 reference structures, which are randomly chosen within the unit cell of the spinel. In a CE with a small number of input structures, both RMS-PS and CV (N+N_{probe}) are much larger than the CV score. The CV score greatly overestimates the predictive power of the trial CEs. This is caused by the strong correlations among the selected clusters and the small variances of the clusters due to the lack of input structures. In the case of the CEs with N =20–35, clusters with strong correlations and with narrow distributions of the correlation functions are selected. Generally speaking, the energies of input structures can be expressed more easily using a set of clusters with strong correlations and narrow distributions than using a set of independent clusters with wide distributions. Therefore, a set of clusters with strong correlations and narrow distributions tends to be selected automatically by the genetic algorithm. As N increases, RMS-PS and CV(N+N_{probe}) decrease. The actual error also decreases along with them. For the CE with N=55, CV(N) and CV(N+N_{probe}) converge. The actual error also converges. Thus, the CV score evaluated from an input set of structures with strong correlations or with a narrow distribution for the selected clusters does not correspond to the predictive power of the CE. On the other hand, the CV score evaluated from an input set of nearly independent structures with a wide distribution for the selected clusters can be a good estimator of the predictive power of the CE.

For the CE with N=40, the ground state can be predicted to be the normal spinel. Structures around the ground state are included in the input set when N=40. In the conventional procedure, once the predicted ground states converge and structures around the ground states have already been included in the input set, the ground-state search cannot improve the CE. Therefore, the CE with N=40 can be regarded as a convergent point in the conventional procedure. However, the actual error of the CE when N=40 is about two times larger than that of the CE when N=55. To improve the CE that is obtained by the conventional procedure, structures other than those around the ground states must be included in the input set.

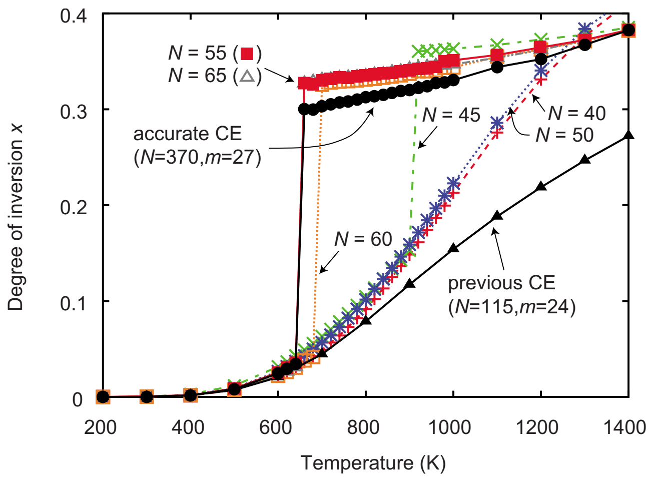

FIG. 8. (Color online) Temperature dependence of the degree of inversion in $MgAl_2O_4$ spinel calculated from the trial CEs with m =17. The temperature dependences predicted from the CE with N =115 and m=24 in our previous paper (Ref. 17) and predicted from an accurate CE with N=370 and m=27 are also shown.

Figure 8 shows the temperature dependences of the degree of inversion predicted from the CEs with N=40–65, in which the ground state is precisely predicted. As the CE and the actual error converge, the temperature dependence of the degree of inversion almost converges. The temperature dependence predicted from the CE in our previous paper,[17] which was constructed by the conventional scheme, is also shown in Fig. 8. In the previous CE, the input structures were composed of random structures in addition to the ground and near-ground-state structures predicted from the trial CEs. Both the present CE with N=40 and the previous CE can be regarded as having converged in the conventional procedure. The temperature dependences at low temperatures below 600 K are independent of the procedure. This implies that structures with low energies are accurately predicted even by the conventional CEs. On the other hand, the temperature dependences at high temperatures (>600 K) are markedly different between the present CE and conventional CEs. To examine the accuracy of the temperature dependences obtained from both the present and conventional CEs, we construct a more accurate CE from 370 structures by minimizing the CV score using 20 multibody clusters. The 370 structures consist of the 70 input and probe structures and the 300 reference structures used to estimate the actual error. The CV score of the accurate CE is 0.38 meV/cation, which is smaller than that of the present CE with N=55 of 1.0 meV/cation.

The accurate CE predicts that the cation-disordering transition of $MgAl_2O_4$ is of the first order with the transition temperature of $T_c = 620$ K. The present CE with $N \geq 55$ is successful in predicting the transition behavior and T_c. On the other hand, the conventional CEs fail to show them. As illustrated in Fig. 3, the difference can be ascribed to the magnitude of the uncertainty in the CE. Structures with high energies are poorly predicted in the conventional CEs. The proposed procedure predicts the high-energy structures accurately despite the use of smaller values of N and m than the previous CE. A small difference can be seen between the present CE with $N \geq 55$ and the accurate CE above the transition temperature, which may be explained by even better accuracy of high-energy structures in the accurate CE. In order to predict the cation-disordering behavior, however, the accuracy by the present CE should be good enough.

IV. CONCLUSION

We have proposed an iterative procedure based on the validation scheme of the trial CE using additional important DFT structures to increase the accuracy of the CE significantly. We applied the procedure to the cation disordering in the MgO-ZnO pseudobinary system and in $MgAl_2O_4$ spinel. Compared with the conventional procedure, the predictive power of the out-of-sample structures is improved. The configurational behavior can be predicted accurately up to high temperatures, in particular, in the complex $MgAl_2O_4$ system. Using the proposed procedure, we can obtain the optimal CE systematically and efficiently with the accuracy that is required to describe alloy thermodynamics.

ACKNOWLEDGMENTS

This study is supported by the Program for Improvement of Research Environment for Young Researchers from Special Coordination Funds for Promoting Science and Technology (SCF) commissioned by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan. This study was also supported by Grant-in-Aid for Young Scientists B.

*seko@cms.mtl.kyoto-u.ac.jp

[1] J. M. Sanchez, F. Ducastelle, and D. Gratias, Physica A **128**, 334 (1984).

[2] D. de Fontaine, *Solid State Physics* (Academic, New York, 1994), Vol. 47.

[3] F. Ducastelle, *Order and Phase Stability in Alloys* (Elsevier, New York, 1994).

[4] J. W. D. Connolly and A. R. Williams, Phys. Rev. B **27**, 5169 (1983).

[5] Z. W. Lu, S.-H. Wei, A. Zunger, S. Frota-Pessoa, and L. G. Ferreira, Phys. Rev. B **44**, 512 (1991).

[6] M. Asta, D. de Fontaine, M. van Schilfgaarde, M. Sluiter, and M. Methfessel, Phys. Rev. B **46**, 5055 (1992).

[7] G. D. Garbulsky and G. Ceder, Phys. Rev. B **51**, 67 (1995).

[8] A. van de Walle, G. Ceder, and J. Phase Equilib. **23**, 348 (2002).

[9] N. A. Zarkevich and D. D. Johnson, Phys. Rev. Lett. **92**, 255702 (2004).

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

[11] S. Müller, J. Phys.: Condens. Matter **15**, R1429 (2003).

[12] V. Blum and A. Zunger, Phys. Rev. B **70**, 155108 (2004).

[13] V. Blum, G. L. W. Hart, M. J. Walorski, and A. Zunger, Phys. Rev. B **72**, 165113 (2005).

[14] M. Stone, J. R. Stat. Soc. Ser. B (Methodol.) **36**, 111 (1974).

[15] A. Van der Ven, M. K. Aydinol, G. Ceder, G. Kresse, and J. Hafner, Phys. Rev. B **58**, 2975 (1998).

[16] A. Van der Ven, C. Marianetti, D. Morgan, and G. Ceder, Solid State Ionics **135**, 21 (2000).

[17] A. Seko, K. Yuge, F. Oba, A. Kuwabara, I. Tanaka, and T. Yamamoto, Phys. Rev. B **73**, 094116 (2006).

[18] A. Seko, A. Togo, F. Oba, and I. Tanaka, Phys. Rev. Lett. **100**, 045702 (2008).

[19] A. van de Walle and D. E. Ellis, Phys. Rev. Lett. **98**, 266101 (2007).

[20] A. Predith, G. Ceder, C. Wolverton, K. Persson, and T. Mueller, Phys. Rev. B **77**, 144104 (2008).

[21] M. Sanati, G. L. W. Hart, and A. Zunger, Phys. Rev. B **68**, 155210 (2003).

[22] C. R. Rao, *Linear Statistical Inference and Its Applications*, 2nd ed. (Wiley, New York, 1973).

[23] A. Díaz-Ortiz, H. Dosch, and R. Drautz, J. Phys.: Condens. Matter **19**, 406206 (2007).

[24] P. E. Blöchl, Phys. Rev. B **50**, 17953 (1994).

[25] G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999).

[26] D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. **45**, 566 (1980).

[27] J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).

[28] G. Kresse and J. Hafner, Phys. Rev. B **47**, 558 (1993).

[29] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).

[30] A. Seko, http://sourceforge.net/projects/clupan (2007).

[31] J. M. Cowley, J. Appl. Phys. **21**, 24 (1950).

[32] G. L. W. Hart and R. W. Forcade, Phys. Rev. B **77**, 224115 (2008).