Grouping of structures for cluster expansion of multicomponent systems with controlled accuracy

Atsuto Seko*
Department of Materials Science and Engineering, Kyoto University, Kyoto 606-8501, Japan

Isao Tanaka
Department of Materials Science and Engineering, Kyoto University, Kyoto 606-8501, Japan and
Nanostructures Research Laboratory, Japan Fine Ceramics Center, Nagoya 456-8587, Japan
(Received 25 January 2011; revised manuscript received 23 May 2011; published 24 June 2011)

Control of errors over the whole range of structures is essential when we combine a large set of density-functional theory calculations and the cluster expansion method for predicting the ground-state structures and configurational thermodynamics of multicomponent systems. Minority structures that are far from a random structure are important for such a prediction. In this paper, we propose a procedure based on the cluster analysis of the structure population, which can adequately take into account the errors of minority structures as well as those of random structures. The usefulness of the procedure is demonstrated by applying it to configurational behaviors of $MgAl_2O_4$ spinel.

DOI: 10.1103/PhysRevB.83.224111 PACS number(s): 61.50.Ah, 81.30.Dz

I. INTRODUCTION

The cluster expansion (CE) method[1–3] is a powerful tool for predicting the ground-state structures, thermodynamics, and configurational properties of multicomponent systems. In order to make reliable thermodynamics with a high accuracy, a combination of the CE method and density-functional theory (DFT) calculations have mostly been adopted. An optimal CE is generally constructed from the DFT results of many ordered structures that are sampled from the total population of structures. In principle, the CE can be made without losing the accuracy of DFT calculations. To perform calculations at such a level, however, it is essential to measure and control the accuracy of the CE properly. Otherwise, the CE may lead to a wrong prediction of the ground states and alloy thermodynamics.

Figure 1 shows a distribution function of the structure population in the configurational space. In the schematic map, group 4 includes the largest number of structures. They correspond to structures near a random structure. On the other hand, group 1 includes the smallest number of structures, which are far from a random structure. They will be hereafter called "minority structures." It should be emphasized that ground-state structures are usually included in minority structures. Indeed, on the basis of the inspection of existing structures in binary compounds, Hart made a hypothesis that the energy shows an extremum (maximum or minimum) in the "least random" structure that has a high relative likelihood index.[4] If this holds true, the accuracy for predicting minority structures should be very important. In this study, we propose a procedure for measuring and controlling the accuracy of a wide range of structures including minority structures on the basis of the distribution of the total population of structures. The use of the cluster analysis of the structure population (CASP) for estimating the accuracy of a wide range of structures will be shown.

II. CLUSTER ANALYSIS OF STRUCTURE POPULATION

A. Accuracy of conventional CE for minority structures

The CE method gives an effective representation of the configurational energy. Within the formalism of CE, the configurational energy E of a binary system is expressed using the pseudospin configurational variable σ_i for the respective lattice site i and the effective cluster interactions (ECIs) V as

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

where φ_α is called the correlation function of the cluster α, which depends only on the atomic configuration. A principal objective is to estimate unknown ECIs from DFT calculations as precisely as the configurational properties can be predicted within the accuracy of DFT calculations.

The cross validation (CV) score[5,6] has been widely accepted as a quantity for controlling the accuracy of the CE. The leave-one-out CV score is obtained from the squared average over N_{DFT} input DFT structures as

$$(\mathrm{CV})^2 = \frac{1}{N_{\mathrm{DFT}}}\sum_{n=1}^{N_{\mathrm{DFT}}}(\hat{E}_{(n)} - E_n)^2, \quad (2)$$

where E_n denotes the DFT energy of structure n, and $\hat{E}_{(n)}$ is the energy of structure n predicted by the CE without using the DFT energy of the structure n. For the accurate evaluation of the CV score, an optimal set of many DFT structures is necessary.[7,8] In such a case, however, the errors of minority structures are only a minor part of the CV score. When the errors of minority structures are much larger than the average error, as shown in Fig. 2, they tend to be underestimated in the CV score.

Such a situation usually occurs when we truncate CE. The truncation error, which is the major source of total CE error, tends to be larger in minority structures in the general use of pseudospin values of $+1$ and -1 since they have larger values of correlation functions than a random structure. To make an accurate CE of such minority structures, an explicit evaluation of the errors of minority structures is essential.

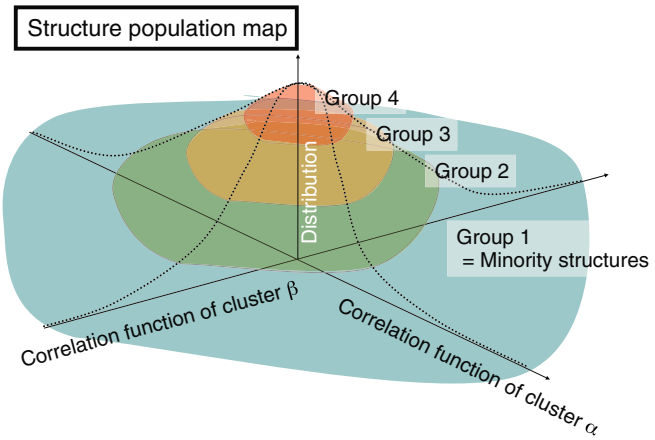

FIG. 1. (Color online) Schematic illustration of distribution function of structure population in space of correlation functions.

B. Procedure of CASP

Cluster analysis generally means the classification of data into meaningful subgroups. CASP enables us to classify structures of similar correlation functions into the same group, as illustrated in Fig. 1. Here CASP is performed by the model-based cluster analysis.[9,10] The likelihood of the correlation functions of all structures in the structure population is modeled by a Gaussian mixture. When the structure population is composed of N structures with the correlation functions $(\boldsymbol{x}_1, \ldots \boldsymbol{x}_N)$, the likelihood L for a model with N_ξ group, given by

$$L(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \tau_1, \ldots, \tau_{N_\xi}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{N_\xi}, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_{N_\xi}) = \prod_{n=1}^{N} \sum_{\xi=1}^{N_\xi} \tau_\xi f_\xi(\boldsymbol{x}_n | \boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi), \quad (3)$$

is maximized, where τ_ξ denotes the probability that a structure belongs to group ξ, and $f_\xi(\boldsymbol{x}_n|\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi)$ is a multivariate Gaussian $(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi)$ for the density of structure n from group ξ, centered at the means $\boldsymbol{\mu}_\xi$. The other geometric characteristics of the Gaussian are determined by $\boldsymbol{\Sigma}_\xi$.

We introduce individual CV scores in each group for estimating the accuracy of a wide range of structures including minority structures. The CV score in group ξ after CASP will be noted by CV-CASP$^{(\xi)}$, which is defined as

$$(\mathrm{CV} - \mathrm{CASP}^{(\xi)})^2 = \frac{1}{N_{\mathrm{DFT}}^{(\xi)}} \sum_{n=1}^{N_{\mathrm{DFT}}^{(\xi)}} (\hat{E}_{(n)} - E_n)^2, \quad (4)$$

where $N_{\mathrm{DFT}}^{(\xi)}$ denotes the number of DFT structures belonging to group ξ. To distinguish CV-CASP from the average CV score, the latter will hereafter be called CV-AVE.

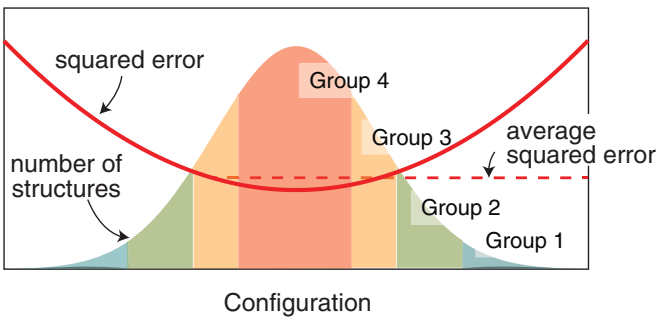

FIG. 2. (Color online) Schematic illustration of typical distribution function of squared nonrandom errors.

III. APPLICATION TO CONFIGURATIONAL BEHAVIOR IN SPINEL OXIDE

A. $MgAl_2O_4$ spinel oxide

We examine the quality of CV-CASP by constructing CEs to determine the configurational properties of cations between fourfold-coordinated tetrahedral and sixfold-coordinated octahedral sites in an fcc oxygen sublattice of $MgAl_2O_4$ spinel oxide. The $MgAl_2O_4$ spinel is a model system with complex interactions.[11] Although the $MgAl_2O_4$ spinel is a typical ionic system, its electrostatic energy changes largely along with the change in the internal coordinates of oxygens. This implies that a large number of many-body clusters is required to express the energetics related to the cation configurations in the $MgAl_2O_4$ spinel. In such a system, errors associated with cluster truncation are anticipated to be large.

B. Computational details

First, CASP is performed in the $MgAl_2O_4$ spinel. Since the number of different structures in the total population is countable infinity, we need to prepare the structure population approximately. To construct an approximated structure population, we search for all symmetrically independent structures within the unit cell of the spinel containing 56 atoms in which 24 are cations. The number of such independent structures is 4222. An approximated structure population can also be prepared by other techniques such as a method for obtaining derivative structures.[12] CASP is then performed by model-based cluster analysis. Here, we consider the correlation functions of 126 clusters up to quadruplets. The likelihood is maximized using the expectation-maximization (EM) algorithm for each of 100 kinds of Gaussian mixture models. We regard the model with the lowest Bayesian information criterion (BIC)[13] among the 100 models as the best one. In the best model, the structure population is divided into four groups, which contain 114, 604, 1071, and 2433 structures, respectively. The ground-state structure, i.e., the normal spinel structure, belongs to group 1 (minority structures).

We then discuss how to find the optimal number of clusters and the optimal set of clusters. Since the accuracy of CE needs to be controlled mainly by four factors, namely, the number of clusters (m), the combination of clusters, the number of DFT structures (N_{DFT}), and the combination of DFT structures, we use all 4222 structures as an optimal set of input DFT structures. First, CEs with up to 40 clusters are constructed using a widely used procedure. The pseudospin values of Mg and Al are assigned to be $+1$ and -1, respectively. The set of clusters that minimizes CV-AVE is searched for from a pool of 126 clusters up to quadruplets using the genetic algorithm (GA)[6] for each number of clusters. The ECIs of the selected

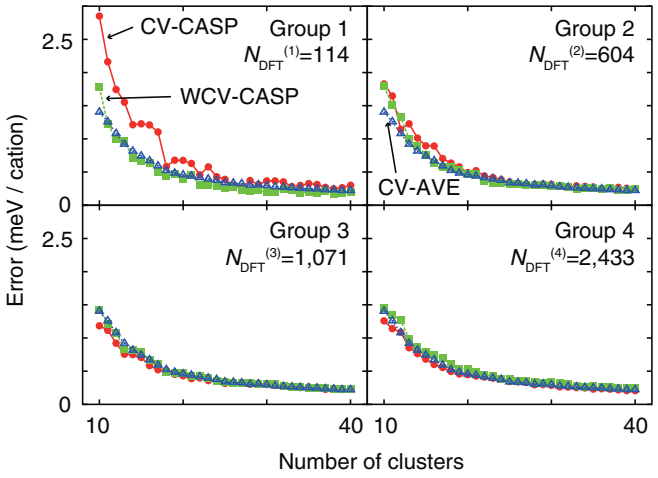

FIG. 3. (Color online) Dependence of the minimized CV-AVE on the number of clusters (open blue triangles). CV-CASPs (closed red circles) and WCV-CASPs (closed green squares) with weights proportional to $1/N_{\mathrm{DFT}}^{(\xi)}$ for four groups are shown together.

clusters are estimated using the least-squares technique. We use the *clupan* code[14,15] to construct the CEs. DFT calculations are performed by the projector augmented-wave (PAW) method[16,17] within the local-density approximation (LDA)[18,19] as implemented in the VASP code.[20,21] The plane-wave cutoff energy is set to 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.

C. Quality of CV-CASP

Figure 3 shows the dependence of the minimized CV-AVE on the number of clusters. When the CV-AVE gradually converges with the number of clusters, a CE with a required precision can be regarded as an optimal CE. For instance, when we require an accuracy of 0.8 meV/cation in Fig. 3, the optimal number of clusters can be found to be $m_{\mathrm{opt}} = 14$. The CV-CASPs for all groups are shown in Fig. 3. The CV-CASP for group 4 is close to the CV-AVE since group 4 contains the largest number of DFT structures in this case. On the other hand, the CV-CASP for group 1, i.e., minority structures, is much larger than the CV-AVE. This indicates that the CV-AVE does not take into account the errors of minority structures. The difference between the CV-AVE and the CV-CASP can be ascribed to the cluster truncation in this case. The averages of the root-mean-square (rms) of the correlation functions for 126 clusters in groups 1 and 4 are 0.303 and 0.190, respectively, which implies that the truncation error is larger in group 1. The use of CASP is essential for obtaining the optimal CE in such a case.

When the errors for minority structures are larger than the average error, the accuracy of the CE for minority structures can be improved by the simultaneous optimization of CE for all the groups. In order to do so, it is natural to minimize the mean square of CV-CASPs as expressed by $[\sum_{\xi}(\mathrm{CV} - \mathrm{CASP}^{(\xi)})^2]/N_{\xi}$. When ECIs are estimated from the weighted least-squares fitting with weights proportional to $1/N_{\mathrm{DFT}}^{(\xi)}$, the RMS of CV-CASPs will be called the weighted CV-AVE (WCV-AVE).

Here, we construct CEs that minimize WCV-AVE with weights proportional to $1/N_{\mathrm{DFT}}^{(\xi)}$. CV-CASP to the weighted fit (WCV-CASP) can also be defined by the same equation as Eq. (4), although the predicted energy is obtained from the weighted fit. The WCV-CASPs for all groups are shown in Fig. 3. Compared with the error with the minimization of CV-AVE, the error for group 1 is reduced. In other words, the minimization of WCV-AVE improves CE particularly for minority structures.

D. Practical procedure for constructing CE based on CASP

Although we have used the approximated structure population as an optimal set of DFT structures so far, a smaller set of DFT structures should be preferred for practical use. In our previous paper,[7] we proposed an iterative procedure for obtaining an optimal set of DFT structures, where CV-AVE minimization and structure selection were repeated. In the procedure, structures that minimize the variance of the predicted energy estimated by linear regression were selected and were called "probe structures." The variance-minimization approach to structure selection was adopted.[5,22] However, since the variance-minimization approach is based on linear regression, assuming that errors are distributed randomly, it is less suitable when a system has a nonrandom error distribution, as illustrated in Fig. 1(b). In such a case, DFT structures should be uniformly sampled in a configurational space. We propose a procedure that combines WCV-AVE minimization and DFT structure selection based on CASP.

The uniform sampling of DFT structures can be achieved by evenly selecting structures from all the groups divided by CASP. Here, CEs are constructed from DFT structures sampled evenly and randomly from all the groups. The number of clusters is fixed at 17. ECIs are estimated by the weighted fit with weights proportional to $1/N_{\mathrm{DFT}}^{(\xi)}$. The set of clusters is optimized as WCV-AVE is minimized. Since DFT structures are picked up evenly from all the groups, all the DFT structures have the unit weight and WCV-AVE corresponds to CV-AVE.

To examine the quality of the CEs constructed by the proposed procedure, CEs are constructed from two kinds of DFT structures prepared by different sampling procedures. One is composed of high-symmetry structures (HSs), which have multiple symmetry operations. The other is composed of randomly selected structures (RAs). The CE error is approximately estimated from the RMS difference between the DFT and CE energies for all the structures in the structure population. Since the accuracy of the CEs made from the CASP and RA samplings is dependent on the selected structures, we perform CE ten times for each N_{DFT} and estimate CE error by averaging the errors of ten CEs. Figure 4(a) shows the dependence of CE error on the number of DFT structures. As can be seen in Fig. 4(a), the CEs with the CASP and RA samplings are better than the CE with the HS sampling. In the CEs with the CASP and RA samplings, the errors are almost the same and converge at $N_{\mathrm{DFT}} = 120$.

In order to examine the accuracy for a wide range of structures from a different viewpoint, the dependence of CE error on the DFT energy at $N_{\mathrm{DFT}} = 120$ is shown in Fig. 4(b). In the CE with the CASP sampling, structures with a wide range of energies can be most precisely predicted among three sampling

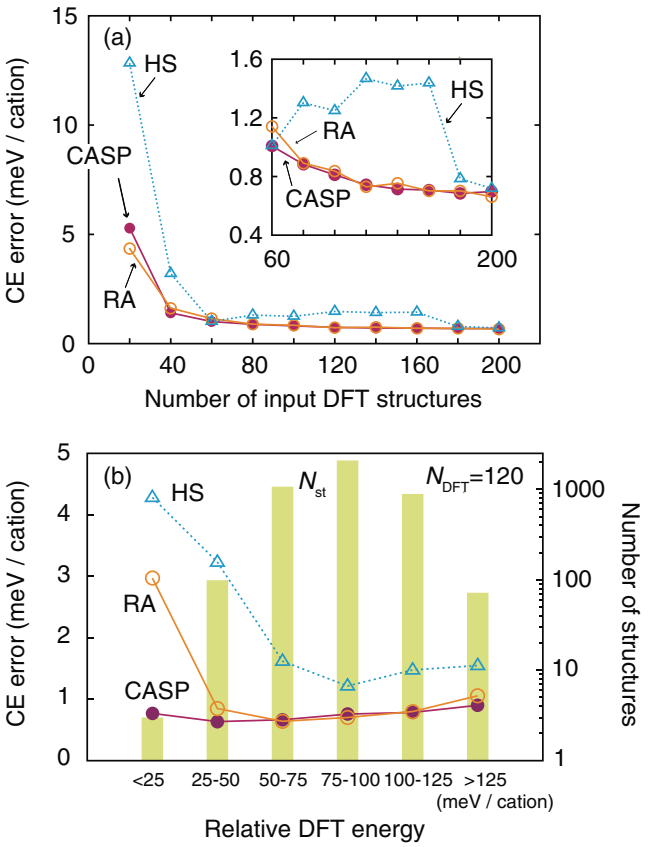

FIG. 4. (Color online) (a) Errors of CEs constructed using three types of DFT structure sampling. (b) DFT energy dependencies of CE errors made from three types of 120 DFT structure. The number of structures belonging to a group classified on the basis of the DFT energy is also shown. The relative energy is measured from the energy of the normal spinel.

approaches. On the other hand, the CE with the RA sampling has a low accuracy for structures with a low energy. It can reconstruct structures only in the energy range 50–125 meV.

When constructing an optimal CE on the basis of WCV-AVE minimization without using the CE error estimated from the DFT energies for the structure population, the convergence of WCV-AVE should be examined using the probe structures in the same way as reported in our previous paper.[7] Structures selected evenly from each group classified by CASP can be adopted as the probe structures. When the WCV-AVE evaluated with the probe structures of a trial CE converges, an optimal CE with m clusters is obtained. Finally, by performing the CE for various m values, the number of clusters can be determined using WCV-CASP.

IV. SUMMARY

We have proposed a CE technique that is reliable for predicting a wide range of structures including minority structures. The procedure is particularly useful when the CE errors of minority structures are larger than the average error. Such a situation occurs when we truncate CE in Eq. (1). To find an optimal CE in such a system, adequate estimation of the errors of minority structures is essential. In the proposed procedure, CASP enables us to distinguish structures and estimate the errors in each group. The accuracy of CE over a wide range of structures can be improved by optimizing CE in all the groups simultaneously by WCV-AVE minimization. This would lead to a more accurate prediction of the ground-state structures, thermodynamics, and configurational properties of multicomponent systems.

ACKNOWLEDGMENTS

This study was supported by a Grant-in-Aid for Young Scientists (A) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. I.T. acknowledges support in the form of both a Grant-in-Aid for Scientific Research (A) and a Grant-in-Aid for Scientific Research on Priority Areas "Nano Materials Science for Atomic Scale Modification 474" from MEXT, Japan.

*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 Phys. **47**, 33 (1994).
[3] F. Ducastelle, *Order and Phase Stability in Alloys* (North-Holland, Amsterdam, 1994).
[4] G. L. W. Hart, Nat. Mater. **6**, 941 (2007).
[5] A. van de Walle and G. Ceder, J. Phase Equilib. **23**, 348 (2002).
[6] G. L. W. Hart, V. Blum, M. J. Walorski, and A. Zunger, Nat. Mater. **4**, 391 (2005).
[7] A. Seko, Y. Koyama, and I. Tanaka, Phys. Rev. B **80**, 165122 (2009).
[8] B. Arnold, A. Díaz Ortiz, G. L. W. Hart, and H. Dosch, Phys. Rev. B **81**, 094116 (2010).
[9] C. Fraley and A. E. Raftery, J. Am. Stat. Assoc. **97**, 611 (2002).
[10] C. Fraley and A. E. Raftery, *MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering*, Technical Report 504 (University of Washington, Department of Statistics, 2006) (revised 2009).
[11] A. Seko, K. Yuge, F. Oba, A. Kuwabara, I. Tanaka, and T. Yamamoto, Phys. Rev. B **73**, 094116 (2006).
[12] G. L. W. Hart and R. W. Forcade, Phys. Rev. B **80**, 014120 (2009).
[13] G. Schwarz, Ann. Statist. **6**, 461 (1978).
[14] A. Seko, J. Am. Ceram. Soc. **93**, 1201 (2010).
[15] A. Seko, *clupan* [http://sourceforge.net/projects/clupan].
[16] P. E. Blöchl, Phys. Rev. B **50**, 17953 (1994).
[17] G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999).
[18] D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. **45**, 566 (1980).
[19] J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).
[20] G. Kresse and J. Hafner, Phys. Rev. B **47**, 558 (1993).
[21] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).
[22] T. Mueller and G. Ceder, Phys. Rev. B **82**, 184107 (2010).