Cluster expansions in multicomponent systems: precise expansions from noisy databases

2007 J. Phys.: Condens. Matter 19 406206

(http://iopscience.iop.org/0953-8984/19/40/406206)

IOP PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **19** (2007) 406206 (19pp) doi:10.1088/0953-8984/19/40/406206

Cluster expansions in multicomponent systems: precise expansions from noisy databases

Alejandro Díaz-Ortiz[1], **Helmut Dosch**[1] **and Ralf Drautz**[2]

[1] Max-Planck-Institut für Metallforschung, Heisenbergstraße 3, D-70569 Stuttgart, Germany
[2] Department of Materials, University of Oxford, Parks Road, Oxford OX1 3PH, UK

Received 25 June 2007
Published 11 September 2007
Online at stacks.iop.org/JPhysCM/19/406206

Abstract

We have performed a systematic analysis of the numerical errors contained in the databases used in cluster expansions of multicomponent alloys. Our results underscore the importance of numerical noise in the determination of the effective cluster interactions and in the expansion determination. The relevance of the size of and information contained in the input database is highlighted. It is shown that cross-validatory approaches by themselves can produce unphysical expansions characterized by non-negligible, long-ranged coefficients. A selection criterion that combines both forecasting ability and a physical limiting behavior for the expansion is proposed. Expansions performed under this criterion exhibit the remarkable property of noise filtering. A discussion of the impact of this unforeseen characteristic of the cluster expansion method on the modeling of multicomponent alloy systems is presented.

1. Introduction: numerical noise and convergence of cluster expansions

Numerical noise in first-principles input data arises from the finite convergency of several computational quantities, such as the k-point mesh, the size of the basis (e.g. energy cut-off in plane-wave-based methods) and zeroing in the forces of atomic positions not fixed by symmetry in the unit cell, among others. This calculation uncertainty can be bound by well known approaches, i.e. using special sets of k-points, smearing mechanisms for the Brillouin zone integration and, of course, the systematic increase of all the relevant parameters until the physical quantity of interest does not vary within certain limits.

Naturally, increasing the accuracy of the first-principles data comes at the price of an increased computational time. For simple systems, characterized (in the sense of a cluster expansion) by few ordered structures with small unit cells, this might be of little importance. However, relevant materials, either from the fundamental or the applied point of view, are usually multicomponent with large unit cells and characterized by complex interactions, i.e.,

many ordered structures are needed to extract the effective cluster interactions (ECIs). For these materials, arbitrarily increasing the accuracy of the first-principles data is an important issue.

The pertinent questions are, in any case, if a higher accuracy in the first-principles database directly translates into more precise cluster expansions and, if so, how this can be gauged. Answering these questions is important in order to find a positive trade-off between the computational expense and the physics embedded in the ECIs. This is one of the aims of the present contribution.

Closely related, on the other hand, is the issue of how to determine a converged expansion. For the enthalpy of formation several routes have been proposed over the years [1–4]. In essence, all these strategies rely on the following *aufbau* principle: a cluster expansion is proposed in terms of an initial database. This cluster expansion is then used for a ground-state search. If new ground states are predicted, then their structures are included in the input database, and a new cluster expansion is constructed. The new expansion usually encompasses a new set of cluster figures and/or different values of the ECIs. This process is repeated until no new ground states are found [1–3]. However, for physical quantities others than the enthalpy of formation, this scheme cannot be fully applied. Consider, for example, the magnetization in binary alloys or the bandgap in semiconductor compounds: in general, there is not a direct (simple) correlation between extrema in such quantities and the ground states in the system.

In consequence, much effort has been put into estimating the residual of a given cluster expansion and several schemes have emerged as a result. Such methods can be roughly divided into two categories. On one hand are the approaches that assess a cluster expansion by its predictive ability, that is, by how accurately a given expansion predicts data not included in the fitting set [1, 4]. On the other hand are the methods that ascertain the quality of an expansion by judging limiting cases, for example the fully random alloy [4]. Clearly, the latter is a physical, perhaps more difficult, approach, whereas the former faces the problem via statistical analysis and, therefore, it is more amenable for automated calculations. This explains the recent popularity of expansion-selection schemes based on the minimization of the cross-validation estimate of the prediction error (see the appendix). In a sense, the prediction error contains a measure of the expansion, i.e. how much of a given cluster expansion is missing. Expansions with small prediction and fitting errors are believed to be 'more complete' and therefore to provide a 'better' (more accurate) description of the underlying physics.

In what follows, we shall see that cluster expansions with very low prediction errors do not necessarily imply a better physical description of the system. Moreover, we will show that in order to produce meaningful expansions it is necessary to account for the numerical noise of the input data—a factor altogether neglected in previous analyses of the cluster expansion [5]. As a result, we will propose a selection criterion for cluster expansions that encompasses the physical limiting behavior together with forecasting ability.

Our strategy relies on the study of archetypical systems for which the underlying interactions (the ECIs) are known by construction, that is, they are defined *a priori*. In all cases, we calculate a physical quantity of interest, e.g. the enthalpy of formation, for a set of input structures. In order to develop systematics, we manipulate the prototype database by adding different levels of Gaussian noise. The resulting databases—characterized by the variance σ^2 of the Gaussian noise—are then analyzed using the variational approach to the cluster expansion (VCX) and a large pool of cluster figures, from which the relevant interactions are selected. Undeniably, this approach provides an advantageous test bed for (cluster expansion) method developing, since the obtained expansions can be cross-checked against the 'control system', that is, the exact ECIs and cluster figures.

The rest of the paper is organized as follows: section 2 is devoted to introducing the main theoretical tools used here, i.e. the cluster expansion method and its variational approach

(the VCX), together with statistical concepts such as model and subset selection. The prototype systems are presented and discussed in section 3. We close this paper with the conclusions in section 5.

2. Theory and methods

The interaction of the cluster expansion community with statisticians has resulted in such a positive thrust for the first-principles thermodynamics of multicomponent systems [1–3, 6, 7]. In this section, we aim to review, albeit briefly but always within the context of cluster expansions, statistical concepts such as model and subset selection, model combining, etc. These concepts are extremely useful in placing the cluster expansion method for multicomponent systems as a problem of subset selection in linear models. More importantly, as we shall see, such concepts are instrumental in determining meaningful (physical) expansions.

2.1. Cluster expansion method as a variable model selection problem

2.1.1. Cluster expansion method. Consider a crystal of N sites characterized by the configuration vector of all occupation lattice sites $\boldsymbol{s}_\ell = \{s_1, s_2, \ldots, s_N\}$. The occupation variable s_i is $+1$ if lattice site i is occupied by an atom A or $s_i = -1$ if its occupied by an atom B. Many physical properties of materials depend on the (atomic) configuration degrees of freedom $\boldsymbol{s}_\ell$. This is the case of the formation enthalpy in metallic alloys or the band-gap in semiconducting compounds.

The seminal idea of the cluster expansion method (CE) is to propose a linear relationship between a physical property F and some function of $\boldsymbol{s}_\ell$ [8–10],

$$F = \sum_\alpha f_\alpha \Phi_\alpha(\boldsymbol{s}_\ell), \tag{1}$$

where the expansion coefficients, f_α, are given by the scalar product between F and the expansion (cluster) functions Φ_α,

$$f_\alpha = \langle F, \Phi_\alpha \rangle. \tag{2}$$

When the cluster functions Φ_α are expressed in terms of orthogonal discrete Chebyshev polynomials, it can be shown that, first, their configuration averages are the well known (and widely used) multisite correlation functions [11], and, second, they constitute a *complete* and *orthogonal* basis in the configurational space [12, 13].

Accounting explicitly for the point-group symmetry of each cluster figure α allows us to rewrite equation (1) as

$$F(\boldsymbol{s}_\ell) = J_0 + \sum_\alpha D_\alpha J_\alpha \Phi_\alpha(\boldsymbol{s}_\ell), \tag{3}$$

where J_α are the effective cluster interactions associated with cluster figures α. The numbers of symmetry-equivalent clusters having identical ECIs are represented by D_α. We have explicitly separated the configuration invariant term J_0 so that the sum in (3) runs over all non-empty clusters.

Some remarks are in order. (a) The orthogonality is, of course, a matter of convenience but the completeness of the basis functions is fundamental to describe *any function* F of the configuration. (b) Once the corresponding expansion coefficients J_α are determined, we can easily compute F for *any configuration* $\boldsymbol{s}$ of the system, including all ordered and disordered states. (c) In a rigorous manner, the determination of an infinite number of expansion

coefficients (i.e. the ECIs) requires a database containing an infinite set of configurations (observations). The viability of equation (3) resides in the notion that configurational degrees of freedom and the crystal structures in alloys are strongly correlated and thus amenable to be described by small number of parameters [7]. Remarks (a) and (b) represent the strength of the method. Its importance, however, is embedded in (c): the expansion can be very well represented by selecting the few most relevant terms.

Note that we favor the term 'selecting' over 'truncating' since the latter gives the (wrong) impression that a convergency radius can be defined for expansion (3), an assumption that is unwarranted from the theory. Certainly, early truncation can lead to gross errors and several examples have been documented in the past [14–16].

Alternatively, hierarchical approaches, where cluster figures are included in the expansion together with all their subclusters, have been proposed recently [17, 18]. Most of these hierarchical approaches have their roots in concepts from the cluster variational method (CVM) [19–21], a technique to construct a hierarchy of consistent approximations to the configuration entropy of lattice systems [22, 23]. So far, however, it is not clear if these approaches *à la* CVM lead to better expansions or convergence—see, for example, the clear account of Blum and co-workers [3].

Modern approaches to the cluster expansion method in multicomponent systems are based on the definition of a pool of cluster figures that is used in the evaluation of equation (3) [2, 3, 6, 24]. A distinctive characteristic of this pool is the lack of a design principle. In other words, the cluster pool contains as many as possible pair and many-body cluster figures up to a given number of vertices and a maximum average bond length (or vertex distance). The elements of the pool can be identified by an index p that runs from 1 to the total number of cluster figures N_c. A pool of cluster figures with these characteristics offers an *unbiased* approximation to equation (3):

$$F(\boldsymbol{s}_\ell) \approx J_0 + \sum_{p=1}^{N_c} D_p J_p \Phi_p(\boldsymbol{s}_\ell). \tag{4}$$

2.1.2. Variable model selection. Posed in this way, it is clear that the cluster expansion represents a special case of the model selection problem, where the standard form is

$$F = \Phi V + \varepsilon, \tag{5}$$

with Φ an $n \times N_c$ full-rank matrix of known constants (predictors), V a N_c vector of unknown parameters, and the errors ε. Issues of model selection arise in many physical and statistical problems that deal with a large number of observed variables [25]. Some of the most commonly encountered problems involve the use of multiple-regression techniques, where it is often desired to find a relatively simple function that models the collected data [26, 27].

On the other hand, it is often entertained that some of the components of V are zero. In this case, model (5) is called the 'full model' and the process of selecting a subset model from the $2^{N_c} - 1$ subset fits is called 'variable model selection' [26, 28–31]. The characteristic of variable model selection is the huge number of candidate models to be considered. Even restricting ourselves to small models, exhaustively evaluating the $2^{N_c} - 1$ possible subsets is prohibitively expensive (for $N_c = 30$ there are $\sim 10^9$ different subsets). Selection criteria based on predictive error estimates obtained by intensive computing methods such as cross-validation are very popular. Cross-validatory methods rely on splitting the data into two parts: one part is used to determine the model (by some goodness-of-fit criterion) and the other is reserved for validation. The rest of the paper will make intense use of the cross-validation scheme which is further discussed in the appendix.

In the cluster expansion method, many of the ECIs are expected to be zero. Therefore, we could apply some of the statistical developments surveyed early to handle equation (4), with the advantage of having a deterministic set of predictors (the cluster functions, Φ_p or Φ_α) that embody the physics of the problem.

2.2. Unbiased cluster expansions

It is clear that aiming for cluster pools with large N_c increases the probability to have a good (controlled) approximation to the correct model—i.e., the exact expansion is recovered when $N_c \to \infty$. However, setting cluster pools with large N_c confronts us with the predicament of having to scan $2^{N_c} - 1$ possible subset models. Even for moderate (typical) values of $N_c = 30$–40, there are between 10^9 and 10^{12} different subset models.

Within this framework, Hart and co-workers [2, 3] have recently proposed an ingenious method to select the relevant (leading) ECIs in expansion (4) by applying a genetic algorithm (GA) to optimize the prediction error in a cross-validatory scheme. By selecting a small number of cluster figures ($N_{GA} \sim 5$) out of a pool of several decades ($N_c \sim 50$), Hart *et al* have cluster expanded first-principles data for several alloy systems with very good prediction errors. Their final selection contained cluster figures that, in general, do not comply with any of the popular hierarchy or compactness criteria. So far, the application of the GA approach has been limited to many-body cluster figures (pairs were fitted separately with an ad hoc decay rule in [2, 3]). However, the method is certainly applicable to all types of cluster figures, including pairs [32].

On the other hand, restricting the number of relevant terms in the expansion to N_{GA} limits the sampling of possible expansions (models) to

$$\begin{pmatrix} N_c \\ N_{GA} \end{pmatrix} = \frac{N_c!}{N_{GA}!(N_c - N_{GA})!},$$

that is, a small fraction of the space spanned by the N_c terms in the cluster pool. This translates as an additional (undesired) factor of uncertainty in the cluster expansion (equation (4)). In other words, confining the expansion to N_{GA} terms naturally rules out the possibility that an optimal expansion with more than N_{GA} terms can be found. Of course, a systematic increase of $N_{GA} \to N_c$ certainly would overcome this limitation, but a numerical overhead associated with enlargement of the search space needs to be considered.

We have recently proposed a variational approach to the cluster expansion (VCX) that is capable of handling the entire pool of cluster figures in a very efficient numerical way. The VCX combines both model selection and model averaging, where the subset selection (cluster expansion) is driven only by the nature of the observation database [6, 24].

Both the GA and VCX methods provide *unbiased* cluster expansions. We refer the interested reader on the GA to the excellent accounts of Hart and co-workers [2, 3], whereas we shall discuss the VCX in detail next.

2.3. Variational approach to the cluster expansion method: subset selection and model combining

Keywords for current methodologies to the cluster expansion in multicomponent systems are *selection* and *forecast*. All of the current methods fall in the category of variable model (subset) selection, namely, a single expansion is chosen and then used to make predictions in the configurational space or to calculate the finite-temperature properties of the system [14, 15]. This is true whether the selection is performed on biased (e.g. ad hoc assumptions on the compactness of the expansion) or unbiased (e.g. the GA approach) grounds. The predictive ability of the expansion, on the other hand, is now considered one of the most

important attributes of an expansion and, in consequence, cross-validatory approaches have been introduced into the selection scheme as a standard ingredient [1, 2, 6, 17, 18].

2.3.1. Model combining. Subset selection methods, however, do not guarantee a best model (expansion) but instead they provide a useful ordering of the potential candidates. A practical recommendation is to use a primary criterion to reduce the number of possibilities to a manageable small size (best candidates) and then employ a second criterion (e.g. cross-validation) to discriminate the best one for the final selection [33]. The recent development, in the context of cluster expansions of multicomponent systems, based on extensive sampling by GAs (to downsize the possibilities space) and the application of cross-validatory selection (for the final selection), is a good example of this approach [2, 3].

Model averaging or model combining offers a different route based on averaging a variety of plausible competing models which are considered with appropriate posterior probabilities instead of having to choose a single best model [33, 34]. In this setting, as an example, we can entertain two models of type (5):

$$F = a_1 + b_1 x + \varepsilon_1 \qquad \text{(model I)}, \tag{6a}$$

$$F = a_2 + \varepsilon_2 \qquad \text{(model II)}, \tag{6b}$$

where a_1, a_2, and b_1 are constants and ε_i are errors ($i = 1, 2$). The posterior probabilities are evaluated from the data as p_1 and p_2. There are now three possibilities. (i) We choose a single model with the highest posterior probability (best prediction capabilities) and use it to make predictions. (ii) We make two forecasts, one for each model in (6), and assign them the corresponding probabilities. (iii) We combine the two predictions in (ii) into a *single weighted prediction*. This latter forecast is effectively the outcome of the model averaging approach and it has a lower mean square prediction error in the long run than either of the individual forecasts [33]. Selecting (iii) implicitly suggests that there is a combined model for which

$$F = p_1 a_1 + p_2 a_2 + p_1 b_1 x + \varepsilon_3 \qquad \text{(model III)}. \tag{7}$$

Model-combining methods that generalize the above ideas have been proposed in recent years to deal with the uncertainty in model selection. The common characteristic of such methods is that they avoid selecting one model by averaging or combining the (best) candidate models [34]. The drawback in this type of approach is that the probabilities can be selected in a more or less arbitrary way, i.e., parts of the model space may be over-represented or under-represented by the analyst's preferences. In section 2.3.2, we shall introduce a variational approach that averages over all possible subset models, assigning them predictive weights based solely on the nature of the database. As a consequence, no user-bias is introduced in the process of model (expansion) combining and/or selection.

2.3.2. Variational approach to the cluster expansion method. The variational approach to the cluster expansion (VCX) is constructed upon the all-important ideas of prediction ability and subset selection but with the additional element of model combination. Before introducing the mathematical description of the method, we would like to review the conceptual framework of the VCX.

Consider the cluster expansion of equation (4) and let $\mathcal{M}$ be the set of all $2^{N_c} - 1$ possible models (expansions). It is clear that for a given cluster pool there are only two options, either the correct model $\mathcal{M}_0$ is contained in $\mathcal{M}$ or it is not. In both cases, a weighted average of all the possible models (in the sense of section 2.3.1) will produce a final model with a better (combined) forecast ability than any of the individual models [33, 34]. In the former case

(when $\mathcal{M}_0 \in \mathcal{M}$), for example, the weighting process would simply render model $\mathcal{M}_0$ with unit probability; that is, the correct model (expansion) will be selected.

In order to make this approach practical, we still have to overcome the problem of the enormous number of possible models, i.e. $2^{N_c} - 1$, for which weights have to be determined based on the individual prediction ability of each model. We have sorted out this problem in two steps. First, we cast the ECIs as functions of the weights and we lifted the normalization restriction on the weights. Second, we optimized the distribution of weights by calculating its variations with respect to the prediction error using cross-validatory techniques.

This approach has the additional advantage that backward elimination can be easily implemented [26], i.e., a cluster pool of size N_c is decimated into a sub-pool of size $N_c - 1$ by removing a cluster figure such that the prediction error for the $N_c - 1$ increases the least, and so until the remains of the pool reach a prescribed size. Because of its variational nature, the backward elimination (decimation) process either maintains or increases the prediction error. In different words, for a given database, the combined expansion associated with N_c produces better or equal forecasts than the combined model associated with $N_c - 1$.

Mathematically, our first step of making the ECIs dependent on the weights can be accomplished by minimizing the *penalized* fitting error

$$\Delta^2_{\mathrm{VCX}} = \frac{1}{n}\Big[\sum_{\ell}\Big(F_\ell - \sum_p D_p J_p^k \Phi_p(\boldsymbol{s}_\ell)\Big)^2 + \sum_p (w_p D_p J_p)^2\Big], \tag{8}$$

with respect to the J for a given set of weights w_p ($p = 1, \ldots, N_c$). Here n represents the total number of observations, e.g., the number of input structures calculated by first principles.

Since the ECIs are now (implicit) functions of the weights, i.e. $J_p = J_p(\boldsymbol{w})$ where $\boldsymbol{w}$ is the N_c-component vector associated with the cluster pool, we can determine the distribution of weights that minimizes the prediction error

$$\frac{\partial \Delta_{\mathrm{pred}}}{\partial w_p} = 0. \tag{9}$$

Both equations (8) and (9) are quite general conditions independent of the method chosen to estimate the forecasting proficiency of the expansion. In the rest of this paper, we shall use a cross-validatory leave-one-out scheme to estimate the prediction error. In this case, Δ_{pred} takes the following form:

$$\Big(\Delta^{\mathrm{CV}}_{\mathrm{pred}}\Big)^2 = \frac{1}{n}\sum_k \Big(F_k - \sum_p D_p J_p^k(\boldsymbol{w}) \Phi_p(\boldsymbol{s}_k)\Big)^2 \tag{10}$$

where the $J_p(\boldsymbol{w})$ are determined without taking into account structure F_k (see the appendix for more details).

Regarding equation (8), it is important to emphasize its generality in the sense that it does not impose any restrictions on the range, compactness or decay behavior of the cluster figures and interactions [16]. The particular form of the penalty term, however, is not unique and one can entertain different functional forms, e.g. $\exp(w_p J_p)^2$. Since the prediction error Δ_{pred} is the objective function to be minimized, the specific form of J weight dependence is, in principle, not critical. In practice, nevertheless, choosing one functional form over other might benefit the numerical minimization process of Δ_{pred}. In other words, depending on the specifics of the system and on the physical quantity to be expanded, a particular form of the penalty term can produce a better defined prediction-error landscape. We have learned that this is indeed the case for the spin cluster expansion [35], where the configurational variables are the continuous projections of the local spin [36].

In a practical implementation of the VCX method, one starts from a random choice of weights w_p, that are used to determine the corresponding set of J_p (minimizing equation (8)).

This set of interactions are in turn used to evaluate the prediction error. Varying the weights and evaluating the associated effective interactions, one can determine the global minimum of the prediction error (equation (9)). This last step deserves special care and robust minimization techniques have to be used. Once the minimization of Δ_{pred} is achieved, the VCX method renders large (small) values for the weights associated with non-relevant (relevant) cluster figures. If the true expansion is contained in the pool of cluster figures, then it would be characterized by a set of weights with small values. Using backward-reduction techniques, we can decimate the initial cluster set, i.e. cleaning the expansion from the irrelevant cluster figures until an optimal expansion is found [6].

3. Prototype systems

Assessing the reliability and robustness of a method for determining the effective cluster interactions in a real system is very difficult for the simple reason that we do not know the correct answer *a priori*. This very fact makes the comparison of different expansions, obtained either by different methods or by the same one, also very difficult; i.e., the predictive power of each expansion can be compared but, as we shall see later on, this is not enough to judge the goodness of a cluster expansion.

In prototype systems, on the other hand, the defining interactions are known by construction (i.e., they are defined *a priori*). Therefore, systematic analyses can be performed and the robustness of the method can be gauged. In the case of cluster expansions, where the predictors are deterministic (i.e. the cluster functions Φ_p), the level of complexity is defined by the number and strength of the expansion coefficients. Simple systems are therefore characterized by few cluster functions with large coefficients, whereas complex systems have many cluster functions with large and small coefficients. In this classification issues like frustration (among the interactions) can be easily accommodated as a secondary level within the two main categories.

3.1. Simple systems: few predictors with large coefficients

Consider a bcc-based system defined by the first three pair interactions and the more compact three- and four-body effective cluster interactions as depicted in figure 1(a). We have named the cluster figures as follows: the number of vertices in the figure is written first, then its rank in that group according to the average bond length, e.g., 2P3 stands for the third pair while 4P1 for the first four-body cluster. The associated values of the ECIs are given below each cluster figure; e.g., the nearest-neighbor pair (2P1) interaction is 1.0. Since this is a prototype system, units are arbitrary in the sense that they have, naturally, energy units but they are not associated with any particular system.

Using these ECIs, we can compute the enthalpy of formation, i.e. the energy of a given configuration $\boldsymbol{s}$ referred to that of the concentration-weighted average of the pure components (A and B),

$$\Delta E(\boldsymbol{s}) = E(\boldsymbol{s}) - xE(A) - (1-x)E(B), \tag{11}$$

as a function of the atomic concentration x. We have plotted this quantity in figure 1(b) for 80 bcc-based ordered structures, including several special quasi-random structures [38–40]. Many of the input structures have been used before in the analysis of Fe-based alloys [24, 37] and some other bcc-based binary systems of Mo, Nb, Ta, and W [41]. Negative values of ΔE indicate a trend to form ordered structures, whereas positive values mark unstable structures against phase separation.

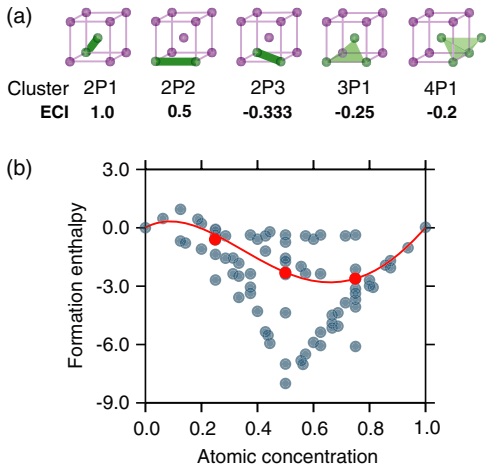

Figure 1. (a) Cluster figures and effective cluster interactions used to generate the enthalpy of formation modeling a simple prototype system—few predictors (cluster functions) with large expansion coefficients. A cluster figure and its associated ECI are named as follows: the number of vertices in the figure is written first, then its rank in that group according to the average bond length; e.g., 2P3 stands for the third pair while 4P1 for the first four-body cluster. (b) Enthalpy of formation obtained from the interactions defined in (a) when applied to a set of 80 bcc-based ordered structures (see [24] for a description). The formation enthalpy for the random alloy is the solid (red) line. Special quasi-random structures (with 16 atoms) are shown in solid circles.

It is not the objective of this paper to exhaustively search for all ground-state configurations (as, for example, in [24, 37]). The random-alloy limit can be easily computed within the cluster-expansion method since the average product of the occupation variables can be replaced by the product average of such variables. The configurational average of the cluster functions Φ_α appearing in equation (3) can be written as

$$\langle \Phi_\alpha \rangle_{\text{random}} = (2x - 1)^{|\alpha|} \tag{12}$$

with $|\alpha|$ the number of sites (vertices) in cluster figure α. At low atomic concentration the system separates into phases even when fully disordered—as seen from figure 1(b), where the formation enthalpy of the random alloy is plotted (red line) together with the obtained values for 16-atom special quasi-random structures (SQSs) [38–40].

We subjected the enthalpies of formation obtained from the ECIs described in figure 1 to the VCX methodology using a pool of 31 cluster figures encompassing pairs, triplets, and quadruplets[3]. When the cluster pool is large enough, so that all relevant cluster figures are included, the VCX method finds an expansion that reproduces accurately enough the input database, i.e. selecting the relevant terms from the cluster pool by assigning them very small weights whereas non-relevant cluster figures would have very large weights. Figure 2 shows the evolution of the prediction error as we decimate (backward-reduction) the original cluster pool. There are several points worth stressing here.

First, *the VCX variationally samples the full expansion space associated with a given cluster pool*. For a 31-cluster pool, the number of possible expansions is $2^{31} \sim 10^9$. The VCX method selects the best expansion by optimizing the weight distribution in a variational

[3] In all our tests, the outcome from the VCX did not depend on the cluster pool size as long as it contained all the relevant terms. We show this particular set merely for convenience in presenting the data. Pools containing as many as 60 clusters (spanning a space of $2^{60} \sim 10^{18}$ possible expansions) were tested without finding any difference from smaller cluster sets.

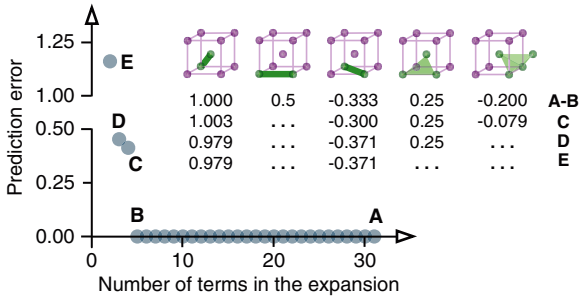

Figure 2. Prediction error as a function of the number of terms in the expansion. The correct (exact) ECIs are recovered in the first step when $N_c = 31$ (point A) and remain the same down to point B. Further decimation (backward reduction) of the pool increases the errors as shown in C–E. The inset shows the actual numerical values of the ECIs from A to E as produced by the VCX.

way for a given cluster pool. Removing a cluster figure from the original set implies again the entire optimization of the expansion space (now of size 2^{30}). In the next step, a pool of 29 cluster figures is then optimized (with an associated space of size 2^{29}), and so on. In general, this optimization process is carried out for *every* cluster pool during the decimation process.

Second, *the VCX finds the correct expansion in the first step*. Both fitting and prediction errors are zero for the entire pool ($N_c = 31$) and down to a decimated set of five cluster figures—that is, all the optimal expansions between points A and B in figure 2. This implies that either a single (correct) solution was found in the first step when the cluster pool had $N_c = 31$, or that 27 different expansions with zero fitting and prediction errors were determined by the VCX during the backward reduction. (We shall see below that the former case is indeed the correct one.)

Third, *the VCX and the decimation process produce robust expansions*. The correct solution is found in the first step when the original cluster figure pool is considered and such expansion is maintained as long as the pool contains all the relevant terms. Moreover, decimating the cluster pool beyond the five relevant cluster figures has little influence on the actual values of the remaining ECIs as compared with the known (exact) ones. This is shown in the lower panel of figure 2, where the numerical values of the ECIs are shown as we decimate the cluster beyond the optimal expansion (i.e. from point C to E).

In figure 3 we have plotted the ECIs obtained when the entire (original) cluster pool is optimized together with the associated weight distribution (lower panel). *Relevant* cluster figures (and associated ECIs) are signaled by the shadowed bars. Notice that relevant terms have finite ECIs and zero weights. Very large weights (zero ECIs), on the other hand, are assigned to *marginal* cluster figures that otherwise would compromise the predictive power of the expansion, whereas *irrelevant* cluster figures have naturally both ECIs and weights equal to zero. It is important to emphasize that the correct (exact) expansion is obtained at the first step even without any decimation procedure. This characteristic is present even in more complex situations, as we will see below, thus confirming the VCX as a very efficient and reliable method—the optimization of a pool of 31 clusters and 80 input structures, that is, the determination of the correct cluster expansion, was virtually effortless.

3.2. Complex systems: many predictors with small and large coefficients

The complexity of a system, and therefore of its cluster expansion, depends on the number of relevant terms (predictors or cluster functions) *and* on the value of the associated ECIs. The

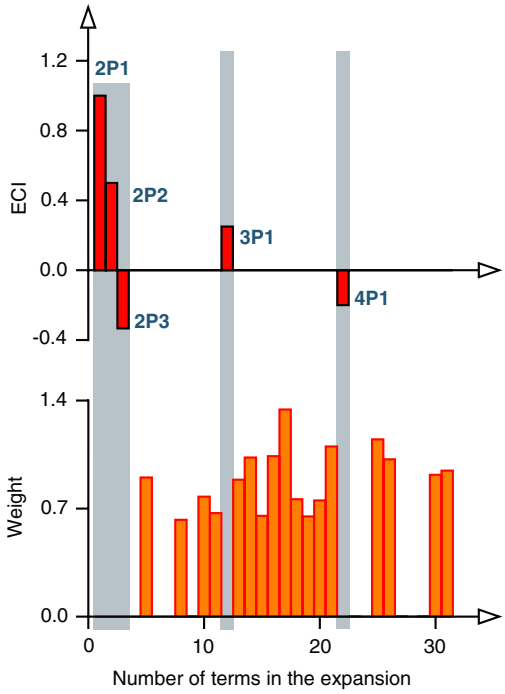

Figure 3. Effective cluster interactions as obtained from the VCX for a cluster pool with $N_c = 31$. In the lower panel, we show the value of the weights associated with each ECI (and cluster figure). Arbitrary units have been used for the values of the weights. Shadowed areas indicate nonzero ECIs and their corresponding weights.

former point is easily and intuitively understood, i.e., the space of possible models (expansions) grows exponentially with the number of terms in the expansion. The latter point, however, it is a little bit more subtle and deserves some additional words.

The determination of an expansion characterized by terms with mixed (small and large) values for the ECIs represents a difficult task if we aim to find *all* the relevant terms. Most fitting strategies can pinpoint the cluster figures associated with large ECIs, whereas terms with small ECIs can be easily misrepresented in the expansion. This means that, in general, a cluster figure with a small ECI contributes little to the physical quantity being cluster expanded. It also means that the contribution of such expansion terms with small ECIs can be either mimicked by a combination of other cluster figures not included in the expansion or that the contribution of such terms can be assimilated by the values of the ECIs associated with the cluster figures already present in the expansion.

The robustness of a cluster expansion can be then defined on such ideas: a scheme capable of retrieving all the relevant terms is therefore more robust than one that misses the terms with small ECIs without mimicking their effect via the combination of additional cluster figures. The least robust scheme is one that easily misses small ECIs and compensates the goodness of fit by adding a number of unphysical (artificial) terms.

We have modeled a complex prototype bcc-based system with the cluster figures and ECIs shown in figure 4(a). We have selected many-body cluster figures that are very likely to be contained in moderate-sized cluster pools and, more importantly, that have shown up in several real-alloy investigations [2, 3, 24, 37, 41]. The cluster figures, however, do not follow any

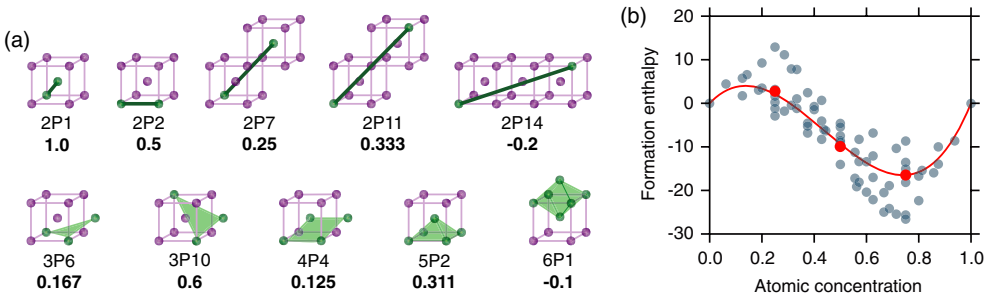

Figure 4. (a) Cluster figures and effective cluster interactions of a complex bcc-based prototype system, i.e. with many mixed (small and large) terms. The nomenclature is the same as in figure 1(a). (b) Enthalpy of formation obtained from the interactions defined in (a) when applied to a set of 80 bcc-based ordered structures (see [24] for a description). The formation enthalpy for the random alloy is represented by the solid (red) line. Special quasi-random structures (with 16 atoms) are shown in solid circles.

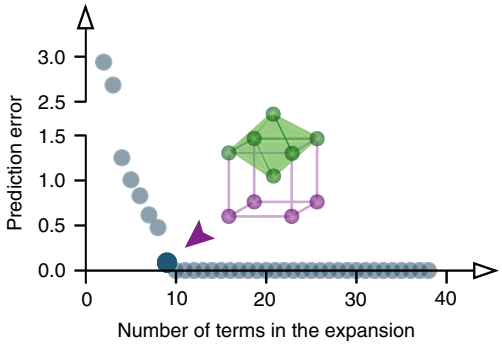

Figure 5. Prediction error as a function of the number of terms in the expansion. The correct (exact) ECIs are recovered in the first step when $N_c = 38$ (original cluster pool) and they stay the same down to $N_c = 10$, where only the relevant cluster figures remain in the pool. Removing the 6P1 cluster figure (at $N_c = 9$) slightly increases the error.

compactness [1, 17] or invariance [18] criteria nor do their associated ECIs decay in any specific way [2, 3, 16]. This is indeed a complex system for which the correct determination of ECIs and cluster figures is a challenging task, especially for schemes based on the above-mentioned criteria.

An input database was generated using such ECIs and cluster figures on the same set of 80 bcc-based ordered structures as we used in section 3.1 and [24]. The outcome can be seen in figure 4(b) where the enthalpy of formation for the ordered structures (open circles) have been plotted as a function of the atomic concentration. The (exact) formation enthalpy of the random alloy is denoted by the solid (red) line and 16-atom SQSs are shown in solid circles for comparison. As expected from the input interactions, the phase diagram exhibits rather interesting and complex features: it is highly asymmetric and shows phase separation at one concentration extreme whereas strong ordering tendencies are seen at the other concentration end.

Subjecting the formation enthalpies of figure 4(b) to the VCX resulted in figure 5, where the fitting and prediction errors are plotted as functions of the number of cluster figures in the pool. The initial pool contained 38 elements encompassing cluster figures from pairs to sextuplets (see footnote 3). The main characteristics observed in this figure were already

present in the simple case discussed in section 3.1, i.e., the correct (exact) expansion was found at the first step when the original cluster pool of 38 elements (spanning $2^{38} \sim 10^{11}$ possible expansions) is considered. In other words, the correct ECIs were retrieved after optimizing the original 38-cluster pool. This set of relevant cluster figures (and ECIs) was naturally maintained during the decimation procedure until all the irrelevant and marginal clusters were sorted out (at $N_c = 10$).

Further decimation of the pool resulted in an increase (from zero) of the fitting and prediction errors. Among the ten possible clusters, the method selected the 6P1 cluster figure as the term to be removed next, because in such a case the errors increased the least. The remaining nine-term expansion showed virtually no change in the expansion coefficients. Even when the expansion is further decimated down to five relevant terms (with errors of about unity), the coefficients are still very close to the original values, i.e., 2P1 = 0.947, 2P7 = 0.269, 3P6 = 0.196, 3P10 = 0.572, and 4P4 = 0.132 (cf figure 4(a)).

4. Noisy databases

In this section, we shall address the effect of the numerical noise on the cluster expansions. That is, we consider that F can be expanded as

$$F(\boldsymbol{s}_\ell) \approx J_0 + \sum_{p=1}^{N_c} D_p J_p \Phi_p(\boldsymbol{s}_\ell) + \varepsilon(\boldsymbol{s}_\ell) \tag{13}$$

where ε represents an additive Gaussian noise characterized by a mean zero and standard deviation σ. In the case of cluster expansions of multicomponent systems, imperfection arises from only two distinct sources: an incomplete set of predictors (cluster functions) or numerical noise in the input database. Although both cases are interesting on their own, we will focus on the latter for the rest of the paper.

There are many investigations on the systematic errors and limitations of first-principles calculations of materials, e.g. the over-binding of the local-density approximation or the over-estimation of the magnetic energy by the generalized gradient approximation to the exchange and correlation energy [42]. However, we are not aware of any study aiming to characterize the distribution of numerical errors as they occur in actual situations and therefore our choice for a Gaussian distribution.

4.1. Cluster expansions under Gaussian noise

A popular selection criterion is to choose an expansion with the least prediction error (as estimated by cross-validatory techniques). This might be a good criterion when the input database is noise free, i.e. the optimal expansion is the one with zero fitting and prediction errors and with the least number of terms. However, when the database contains finite numerical errors, the cross-validation estimate of the prediction error is biased for the subset models (i.e. possible expansions contained in a finite cluster pool). This means that, in some instances, the cross-validation score can be smaller than the actual noise in the database [30].

We added numerical (Gaussian) noise to the exact database of figure 1(b). The amount of numerical noise is characterized by the standard deviation σ of the associated distribution. Figure 6(a) shows the prediction and fitting errors as functions of the number of cluster figures when $\sigma = 0.43$ and a 31-cluster pool is used in conjunction with the VCX method. The first thing that strikes us from this figure is that, in contrast to its exact counterpart (figure 2), both errors behave in a softer way, thus blurring the separation between relevant and non-relevant (marginal and irrelevant) cluster figures. A second feature apparent from figure 6(a) is that

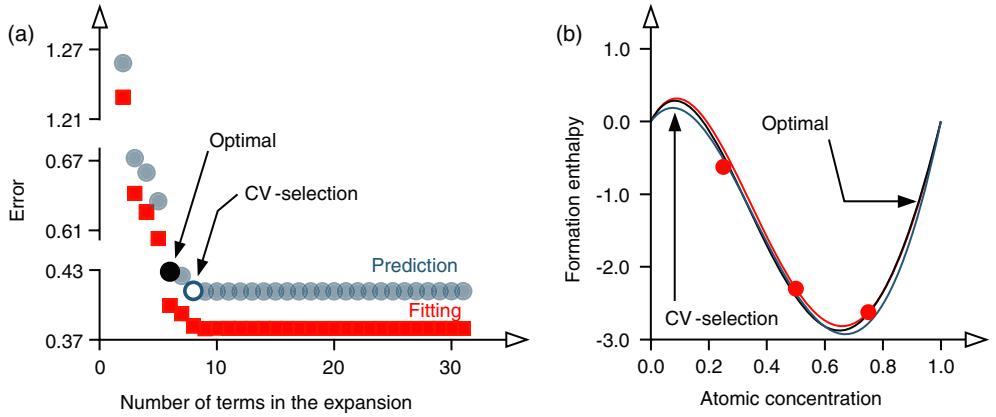

Figure 6. (a) Fitting (circles) and prediction (squares) errors as functions of the number of terms in the expansion. Numerical (Gaussian) noise, characterized by $\sigma = 0.43$, has been added to the input database before subjecting it to the VCX. In the inset two possible expansions are marked, i.e. the optimal expansion and the cross-validation selection. (b) Enthalpy of formation for the random alloy as evaluated from the exact data (red), optimal (black) and cross-validation (green) selections. Solid circles denote the 16-atom SQS.

both fitting and prediction errors can be far below the standard deviation of the input data ($\sigma = 0.43$). Incidentally, this second characteristic exemplifies the efficiency and reliability of the VCX: for a large enough cluster pool the method processes the input data so well that even the numerical noise is well accounted for (in terms of both fitting and prediction).

Clearly, an expansion-selection scheme based only on estimates for the prediction error, e.g. using cross-validatory techniques, can yield expansions with unphysical terms. In the inset of figure 6(a) we have indicated what would be the cross-validation selection together with the optimal (physical) expansion. The latter has been chosen as the smallest expansion satisfying a secondary selection criterion of the form

$$\min\left|\sum_k \Big(F_k - \sum_p D_p J_p \Phi_p(\boldsymbol{s}_\ell)\Big)^2 - n\sigma\right|. \tag{14}$$

An enthusiastic reader might point out that criterion (14) certainly produces *consistent* expansions but that the physical nature of the expansion has to be judged by its compliance with a certain physical limit. We fully agree with this view and, accordingly, we propose the random alloy as the natural physical limit for a truly consistent cluster expansion. A fully disordered alloy can be thought of as an 'ordered' structure with an infinite number of atoms, so that every atomic position is randomly occupied and no long-range order is hence possible.

In figure 6(b) we compare the formation enthalpy for the random alloy as obtained from the exact data (red line) with the two possible selections. Notice that the optimal selection—notwithstanding that it contains *fewer* terms than the cross-validation selection—represents a *better* approximation to the real (exact) one.

Admittedly, the difference between the two possible cluster expansions is rather small for low levels of noise. Criterion (14) becomes pertinent when the database contains higher levels of noise. The frontier separating relevant from non-relevant expansions widens with noise and, without a secondary selection criterion, the selection process of a meaningful expansion become arduous. This can be seen clearly in figure 7, where the input database has been subjected to a noise of $\sigma = 0.79$, i.e. almost twice that in figure 6. The cross-validation choice is now separated by almost a decade of possible expansions from the optimal expansion

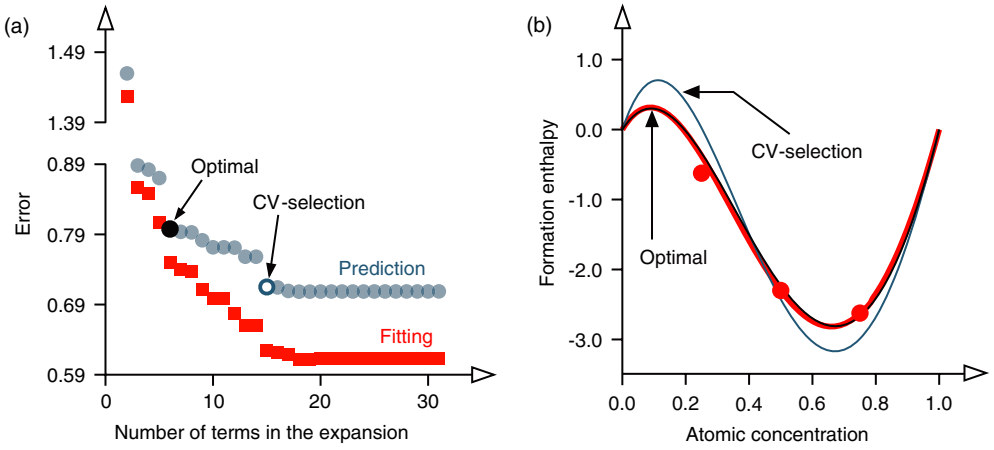

Figure 7. (a) Fitting (circles) and prediction (squares) errors as function of the number of terms in the expansion. Numerical (Gaussian) noise, characterized by $\sigma = 0.79$, was been added to the input database before subjecting it to the VCX. In the inset two possible expansions are marked, i.e. the optimal expansion and the cross-validation selection. (b) Enthalpy of formation for the random alloy as evaluated from the exact data (red), optimal (black) and cross-validation (green) selections. Solid circles denote the 16-atom SQS.

as selected by the VCX in conjunction with equation (14). In figure 7(b) we can see that the random-alloy limit is very well reproduced by the optimal selection whereas the cross-validation choice strongly overestimates both the ordering and phase-separation tendencies at high and low concentrations, respectively.

Therefore, the predictive power of an expansion, as estimated by cross-validatory techniques, is not by itself a good physical criterion. Cluster expanding physical data using the prediction error as the only guidance may result in expansions with unphysical (artificial) terms. Meaningful cluster expansions can only be selected considering the numerical uncertainty in the input database, for example, using equation (14).

4.2. The importance of the database

The concept of a cluster expansion of the configurational degrees of freedom of a physical quantity is an important concept because of its viability: a converged expansion can be obtained by retaining a finite number of relevant terms. Since the expansion coefficients are obtained through a linear-inversion method (e.g. a least-squares fit), it is natural to think that only a few input structures (of the order of the number of relevant cluster functions) should be necessary. In general, however, this is not the case.

The idea of selecting the most appropriate set of input structures is, nevertheless, an appealing one and some methods based on variance reduction have been proposed [1]. However, this is a very difficult task since the suitableness of such an input-set strictly depends on the particular expansion (subset model) being considered. Although it may be possible that a given structure-set is appropriate for more than one expansion, the enormous number of possible expansions (see our early discussion in section 2.1.1) makes this 'fine tuning of input structures' approach difficult to implement in practice. The size of the database plays, in the end, an important role in determining a meaningful expansion.

Here, both simple and complex input databases will be analyzed in two variants: the original database with 80 structures and a subset thereof containing 45 entries with

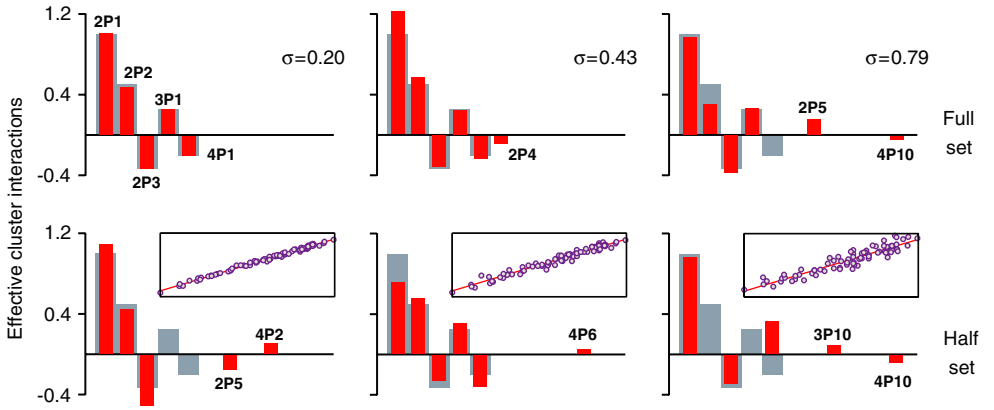

Figure 8. ECIs for a simple prototype system under Gaussian noise. The actual interactions, for both the full and half sets, are displayed with reference to the exact set (shadowed bars). The height and the position of the bars denote the value of the ECI and its type; i.e., everything that falls outside the shadowed bars points to ECIs associated with other cluster figures different from the true ones. For $\sigma < 0.20$ the half set still contains enough information so that the exact ECIs are always recovered (not shown). The inset in the lower panels depicts the noise-treated versus the exact data for each noise level.

concentrations $x \geqslant 0.5$. We then treated both databases with different levels of Gaussian noise. Clearly, a variety of other different sorting-down schemes can be entertained, e.g. random selection or considering only structures close to the ground-state line. However, our pick (hereafter the 'half set') has the advantage of being rather simple to define with the added complexity of sampling only half of the concentration range. In any case, the conclusions drawn from this section will be independent of the particular partitioning of the data.

Figure 8 shows the evolution of the ECIs as the noise level is increased. In all cases, the methodology described so far was applied to select the best possible expansion. For low levels of noise, that is, from $\sigma = 0$ to 0.12, the VCX recovered the exact cluster figures and ECIs for both the full and half sets (not depicted in the figure). Differences between the two sets are first found for moderated Gaussian noise of $\sigma = 0.20$: the 'best' expansion obtained for the half-set (lower panels) is quite different from the real one, whereas for the full set (upper panels), the VCX still yields the right answer. Eventually, when the noise level is high both input-data sets fail to provide enough information and the true expansion cannot be recovered anymore.

This simple yet illustrative exercise underscores two all-important concepts: first, the significance of having enough information (e.g. a large database) that a meaningful (physical) cluster expansion can be performed. The efficiency of a selection scheme cannot replace the need for enough information in the database. Second, the amount of information in a database is not an absolute concept but a relative one, that depends on the level of noise, as it does on the sampling of the configurational space, that is, on the input structures. As a consequence, in order to produce meaningful (i.e. consistent and physical) expansions it is more important to have a good database with moderate or even modest precision than to have scarce yet highly accurate data.

5. Conclusions: the noise-filtering properties of cluster expansions

Traditionally, cluster expansions of first-principles data have been considered to be, in a best-case scenario, as precise as the accuracy of the first-principles data they represent.

Table 1. Optimal effective cluster interactions determined from the entire database (80 observations) as functions of the noise level. For the noise-free database, the true (exact) cluster figures and their associated ECIs are recovered (cf figures 4(a)–5). For a finite level of noise up to $\sigma = 0.1$, our methodology produces expansions that are virtually the exact ones. The robustness of the VCX is exemplified for $\sigma = 0.2$, where the optimal expansion does not include the 6P1 term yet it does not contain any further artificial terms. Notice that very noisy databases contain less information, resulting in poor expansions. See the discussion in the text for further information.

Figure	$\sigma = 0$	$\sigma = 0.05$	$\sigma = 0.10$	$\sigma = 0.20$	$\sigma = 0.42$	$\sigma = 0.79$
2P1	1.000	1.002	1.010	0.995	0.887	0.479
2P2	0.500	0.500	0.489	0.450	···	···
2P7	0.250	0.249	0.245	0.256	0.251	···
2P11	0.333	0.334	0.333	0.348	0.281	···
2P14	−0.200	−0.200	−0.199	−0.222	···	···
3P6	0.167	0.168	0.167	0.161	0.171	0.198
3P10	0.600	0.599	0.600	0.605	0.593	0.575
4P4	0.125	0.126	0.123	0.121	0.128	···
5P2	0.311	0.304	0.316	0.325	0.299	···
6P1	−0.100	−0.102	−0.080	···	···	···
2P6	···	···	···	···	···	0.929
2P9	···	···	···	···	−0.148	−0.299
2P10′	···	···	···	···	···	0.321
2P13	···	···	···	···	···	−0.499
2P13′	···	···	···	···	0.104	0.243
2P15	···	···	···	···	···	0.118
4P1	···	···	···	···	···	0.233
4P3	···	···	···	···	···	0.189
6P2	···	···	···	···	···	−0.086

In consequence, many popular schemes gauge the goodness of an expansion based on its capabilities to reproduce some features of the original database, e.g. the ground states [1–3, 17]. Regardless of this being a good criterion or not, this example underscores the current emphasis on cluster expansions, that is, in obtaining expansions with good fitting and prediction capabilities of the 'raw' database, i.e. the physical data plus the numerical errors.

Inevitably, cluster expansions performed in such way lead to the configurational description of numerical noise contained in the database. This manifests, typically, in expansions with non-negligible and long-ranged ECIs associated with quite open cluster figures (as seen in figure 8 and table 1). For simple systems, such artificial ECIs can be easily ruled out and the appropriate description can be recovered. Nevertheless, even unsophisticated binary alloys can embody a fair degree of complexity. For such complex materials, cluster expanding the raw data would make it difficult to separate the effects of the numerical noise from the physical effects, at the risk of attributing some physical meaning to the former.

The results presented in this paper have shown the importance of broadening our view of cluster expansions to account explicitly for the effects of the numerical noise in the database. Our investigation has revealed that cluster expansions act as noise filters. This unexpected and remarkable feature has a deep impact on the way we approach the modeling of materials properties and how we construct future databases. The emphasis should be now on the amount of information contained in the database and on the consistency of the expansion. This implies the analysis and characterization of the numerical noise in the *ab initio* database.

Finally, in this paper we have used the VCX method as our working tool. However, qualitative and even some quantitative aspects must be present in any unbiased approach to the

cluster expansion. Notably, we expect that the genetic algorithm of Hart and co-workers [2, 3] displays the same noise-filtering behavior as reported here.

Acknowledgments

The authors gratefully acknowledge useful discussions with Daniel Steiauf and Reinhard Singer on the variational approach to the spin-cluster expansion. This work is supported by the Alexander von Humboldt Foundation (AD-O).

Appendix. Prediction error and cross-validation

A model (an expansion) is judged by its ability to reproduce the data on which it is based and, more importantly, by its forecasting proficiency. The predictive ability of the associated fitted model can be characterized using the moments of $\widehat{V}$ (a p-vector estimator of V in equation (5)).

Suppose that Φ_f is a known p-vector and that F_f is an observation (measurement) consistent with model (5) and independent of $\widehat{V}$. Under these circumstances, F_f can be considered as a future observation (measurement). The predicted value of F_f based on the selected model is then $\widehat{F}_f = \Phi_f \widehat{V}$ and the error in prediction is $F_f - \widehat{F}_f$. Therefore, the predictive power of the model is reflected by the statistical properties of $F_f - \widehat{F}_f$ for different choices of Φ_f. Ideally, one would like to obtain the distribution function of $F_f - \widehat{F}_f$, but this is unrealistic in most cases. Instead, the mean-squared error is usually used as a summary.

Performing additional observations for the phenomenon in question is always advisable, but in general this is not always possible either because the observation conditions may have changed or because it represents unbearable costs. In such cases, it is therefore quite natural to consider partitioning the available data into 'construction' and 'validation' sets. The former is used to select a model that, in turn, is assessed using its predictions against the latter set. A possible objection to using the splitting of the database is the loss of information. However, in moderate and large data sets, where the splitting is more practical, this cost is typically negligible [28] (see section 3).

The cross-validation scheme removes the arbitrariness in the division of the data (of size n) by considering a construction sample of size $n - 1$ and a validation sample of size 1 in all the n possible ways [43–45]. Certainly, different partitions of the data can be entertained, e.g. leaving r terms out for the validation subsample and building the model with a construction subsample of size $n - r$ [29, 46]. For the purposes of this paper, the very popular leave-one-out cross-validation is satisfactory enough and it will be used throughout the rest of the paper. Different data-splitting strategies have been reviewed in [29, 46] and some of them are tested in [3] in the context of cluster expansions.

Cross-validation is a method for model selection in terms of the predictive ability of the models. In terms of model (5), the squared prediction error in the leave-one-out cross-validation can be written as

$$\Delta_\gamma^{\mathrm{CV}} = \frac{1}{n}\sum_{i=1}^{n}\left(F_i - \Phi_\gamma V_\gamma^i\right)^2, \tag{A.1}$$

where Φ_γ is the predictor associated with a (sub-) model of size q_γ and V_γ^i is the corresponding least-squares estimator of equation (5) when the observation i is not in the construction subsample. An optimal model can be selected on terms of the prediction error and the number of predictors. This is a very popular criterion in the context of cluster expansions.

References

[1] van de Walle A and Ceder G 2002 *J. Phase Equilib.* **23** 348
[2] Hart G L W, Blum V, Walorski M J and Zunger A 2005 *Nat. Mater.* **4** 391
[3] Blum V, Hart G L W, Walorski M J and Zunger A 2005 *Phys. Rev.* B **72** 165113
[4] Ferreira L G, Wei S-H and Zunger A 1989 *Phys. Rev.* B **40** 3197
[5] Díaz-Ortiz A and Dosch H 2007 *Phys. Rev.* B **76** 012202
[6] Drautz R and Díaz-Ortiz A 2006 *Phys. Rev.* B **73** 224207
[7] Curtarolo S, Morgan D, Persson K, Rodgers J and Ceder G 2003 *Phys. Rev. Lett.* **91** 135503
[8] Sanchez J M and de Fontaine D 1979 *Modulated Structures-1979* (*AIP Conf. Proc.* vol 53) ed J M Cowley, J B Cohen, M B Salamon and B J Wuensch (New York: AIP) p 133
[9] Sanchez J M, Kikuchi R, Yamauchi H and de Fontaine D 1980 *Theory of Alloy Phase Formation* ed L H Bennett (Warrandale, PA: The Metallurgical Society of AIME) p 289
[10] Sanchez J M and de Fontaine D 1981 *Structure and Bonding in Crystals* vol 2, ed M O'Keeffe and A Navrotsky (New York: Academic) p 117
[11] Kawasaki K 1973 *Phase Transitions and Critical Phenomena* vol 2, ed C Domb and M S Green (New York: Academic) p 465
[12] Sanchez J M, Ducastelle F and Gratias D 1984 *Physica* A **128** 334
[13] Sanchez J M 1993 *Phys. Rev.* B **48** 14013
[14] de Fontaine D 1994 *Solid State Phys.* **47** 33
[15] Zunger A 1994 *Statics and Dynamics of Alloy Phase Transformations* ed P E A Turchi and A Gonis (New York: Plenum) p 361
[16] Laks D B, Ferreira L G, Froyen S and Zunger A 1992 *Phys. Rev.* B **46** 12587
[17] Zarkevich N and Johnson D D 2004 *Phys. Rev. Lett.* **92** 255701
[18] Sluiter M H F and Kawazoe Y 2005 *Phys. Rev.* B **71** 212201
[19] Kikuchi R 1951 *Phys. Rev.* **81** 988
[20] Barker J A 1953 *Proc. R. Soc.* A **216** 45
[21] Morita T 1953 *J. Phys. Soc. Japan* **12** 753
[22] Sanchez J M and de Fontaine D 1978 *Phys. Rev.* B **17** 2926
[23] Vul D A and de Fontaine D 1993 *Mater. Res. Soc. Symp. Proc.* **291** 401
[24] Díaz-Ortiz A, Drautz R, Fähnle M, Dosch H and Sanchez J M 2006 *Phys. Rev.* B **73** 224208
[25] Linhart H and Zucchini W 1986 *Model Selection* (New York: Wiley)
[26] Miller A J 1990 *Subset Selection in Regression* (London: Chapman and Hall)
[27] McQuarrie A D R and Tsai C-L 1998 *Regression and Time Series Model Selection* (Singapore: World Scientific)
[28] Picard R R and Cook R D 1984 *J. Am. Stat. Assoc.* **79** 575
[29] Shao J 1993 *J. Am. Stat. Assoc.* **88** 486
[30] George E I 2000 *J. Am. Stat. Assoc.* **95** 1304
[31] Rao C R and Wu Y 2001 *Institute of Mathematical Statistics Lectures Notes* (*Monograph Series* vol 38) p 1
[32] Hart Gus L W 2006 private communication
[33] Chatfield C 1995 *J. R. Stat. Soc.* A **158** 419
[34] Yuan Z and Yang Y 2005 *J. Am. Stat. Assoc.* **100** 1202
[35] Drautz R and Fähnle M 2004 *Phys. Rev.* B **69** 104404
[36] Singer R and Fähnle M 2007 private communication and unpublished
[37] Drautz R, Díaz-Ortiz A, Fähnle M and Dosch H 2004 *Phys. Rev. Lett.* **93** 067202
[38] Zunger A, Wei S-H, Ferreira L G and Bernard J E 1990 *Phys. Rev. Lett.* **65** 353
[39] Wei S-H, Ferreira L G, Bernard J E and Zunger A 1990 *Phys. Rev.* B **42** 9622
[40] Jiang C, Wolverton C, Sofo J, Chen L-Q and Liu Z-K 2004 *Phys. Rev.* B **69** 214202
[41] Blum V and Zunger A 2005 *Phys. Rev.* B **72** 020104(R)
[42] Perdew J P, Ruzsinszky A, Tao J, Staroverov V N, Scuseria G E and Csonka G I 2005 *J. Chem. Phys.* **123** 062201
[43] Stone M 1974 *J. R. Stat. Soc.* B **36** 111
[44] Geisser S 1974 *Biometrika* **61** 101
[45] Geisser S 1975 *J. Am. Stat. Assoc.* **70** 320
[46] Burman P 1989 *Biometrika* **76** 503