Structure-property maps and optimal inversion in configurational thermodynamics

Björn Arnold,[1,*] Alejandro Díaz Ortiz,[1,†] Gus L. W. Hart,[2] and Helmut Dosch[3]
[1]*Max Planck Institute for Metals Research, Heisenbergstraße 3, D-70569 Stuttgart, Germany, EU*
[2]*Department of Physics and Astronomy, Brigham Young University, Provo, Utah 84602, USA*
[3]*DESY, Notkestraße 85, D-22607 Hamburg, Germany, EU*
(Received 20 January 2010; revised manuscript received 9 March 2010; published 25 March 2010)

Cluster expansions of first-principles density-functional databases in multicomponent systems are now used as a routine tool for the prediction of zero- and finite-temperature physical properties. The ability of producing large databases of various degrees of accuracy, i.e., high-throughput calculations, makes pertinent the analysis of error propagation during the inversion process. This is a very demanding task as both data and numerical noise have to be treated on equal footing. We have addressed this problem by using an analysis that combines the variational and evolutionary approaches to cluster expansions. Simulated databases were constructed *ex professo* to sample the configurational space in two different and complementary ways. These databases were in turn treated with different levels of both systematic and random numerical noise. The effects of the cross-validation level, size of the database, type of numerical imprecisions on the forecasting power of the expansions were extensively analyzed. We found that the size of the database is the most important parameter. Upon this analysis, we have determined criteria for selecting the optimal expansions, i.e., transferable expansions with constant forecasting power in the configurational space (a structure-property map). As a by-product, our study provides a detailed comparison between the variational cluster expansion and the genetic-algorithm approaches.

DOI: 10.1103/PhysRevB.81.094116 PACS number(s): 61.50.Ah, 61.66.Dk, 81.05.Bx

I. RATIONAL DESIGN, STRUCTURE-PROPERTY MAPS, AND TARGETING PHYSICAL PROPERTIES

Rational design of molecular systems and solid-state materials relies on the knowledge of the effective potentials or interactions to tailor motifs with favorable properties. In a combinatorial high-throughput approach the task is, in principle, simple: To solve the Schrödinger equation for all viable conformations and combinations of a list of candidate components. In practice, however, this is unfeasible due the astronomical size of the chemical space (i.e., the set of all spatial and chemical conformations available to the system) that must be scanned to optimize a target property.[1]

Constructing maps that relate structure to physical properties is at the core of rational design strategies. This approach looks for correlations between a set of measurements (experimental observations and/or quantum-mechanical calculations) of an observable F and a potential-energy surface V. In a practical fashion, the functional map replaces the true V dependence of $F[V]$ with a functional $f(\boldsymbol{v})$ through a suitable transformation from V to a set of key variables $\boldsymbol{v}=(v_1,\ldots,v_N)$. Members of the potential energy surface are thus characterized by different $\boldsymbol{v}$'s. The choice for the particular form and nature of $\boldsymbol{v}$ depends, of course, on the problem at hand. Data-centered methods[2,3] and basis-functions expansions[4,5] are among the most popular choices in materials science although neural-network approaches for inverting intermolecular potentials also have been reported in the literature.[6]

Cluster expansion (CE) (Ref. 7–9) is the method of choice to map the configurational dependence of many physical (scalar) properties in crystalline systems, including formation enthalpies,[10,11] Curie temperatures,[12] and magnetic moments.[13] Recently, the CE method has been extended to account for anisotropic properties such as the piezoelastic tensor in semiconducting materials.[14] In crystalline compounds, the atoms form a periodic lattice where the potential V does not depend on the spatial coordinates but only on the chemical identity of the atoms sitting at every site $i(=1,\ldots,N)$ of the crystal. The system can be then characterized by the configuration vector $\boldsymbol{s}=(s_1,s_2,\ldots,s_N)$, where s_i indicates the chemical identity of an atom at site i (e.g., $s_i=\pm 1$ in a binary system).

The CE method assumes a linear map between F and some functions of $\boldsymbol{s}$ (Ref. 7)

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

The expansion coefficients or effective cluster interactions (ECIs) f_{α} are obtained by inverting Eq. (1)

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

that is, they are defined as the scalar product between the observable F and the expansion functions Φ_{α}. The Φ_{α} are the so-called cluster functions. They are associated with a subset of sites denoted by the index α. In a compact way, Eq. (1) states the intuitive idea of decomposing a physical quantity F into its point, pair, triplet, etc., contributions.[15] The cluster-expansion method formalizes this idea, expressing the cluster functions Φ_{α} in terms of orthogonal discrete Chebyshev polynomials.[8] The cluster functions constitute a complete and orthogonal basis in the configurational space and their averages are the well-known (and widely used) multisite correlation functions.[16] The orthogonality is, of course, a matter of convenience but the completeness of the basis functions is fundamental in describing any function F of the configuration.

Generating structure-property maps involves an inversion problem where the Schrödinger equation is solved for a set of training cases to determine the f_α's. Usually this is the most expensive step in the construction of the map since, as we shall see below, the size of training set depends on the numerical uncertainty of F. Once constructed, a structure-property map offers an accurate and straightforward description of a physical property F, without the full computational overhead of solving the Schrödinger equation.

The applications of such a map are diverse, from evaluating F in a very complex (computational unaffordable) configuration to calculating phase diagrams in the temperature composition space.[10,11] An appealing application of a structure-property map is that it can be used to optimize physical properties,[17] that is, finding the configuration that targets F directly from Eq. (1) without the evaluation of the Schrödinger equation a prohibitive number of times. The search procedure, however, has many technical intricacies—the complexity of the configuration space makes it difficult to handle the large number of local possible solutions that grow exponentially with the size of the sampling unit cell.[18]

Interesting and useful as they are, all these applications rely on a robust inversion method that is resilient to numerical noise in the data. The inversion method must provide structure-property maps that truly represent the configurational dependence of an observable F, or equivalently, an inversion approach defining effective cluster interactions (f_α) *even* in the face of numerical inaccuracies in the training database. A realistic analysis of the error propagation in the inversion process is very demanding because both data and noise have to be treated on equal footing. Inverting a noisy database provides, in principle, a family of maps or expansions that are statistically consistent with the error distribution of the database. Previous investigations have shown that the combined use of cross-validatory[19,20] and unbiased approaches to the CE method provide expansions that effectively filter moderate levels of random noise out of the training database, and a selection criterion has been developed based on the *a priori* knowledge of the noise level.[21,22]

This paper has the threefold purpose of (i) addressing the issue of quantifying how numerical inaccuracies in the training set propagate through the inversion process, (ii) to propose an approach to determine optimal inversions that lifts the requirement of knowing in advance the noise level in a given database, and (iii) to compare independent inversion methods. To those aims, we have considered simulated numerical imprecisions in the form of both systematic (rounding and saturation) and random (Gaussian) errors in training databases. The structure-property maps were constructed using two different unbiased approaches, one of stochastic nature (genetic algorithm)[23,24] and the other following a variational principle (variational cluster expansion).[22] Our results show that cross-validatory techniques can be used to recursively and self-consistently determine the noise level in the database. A tradeoff between the effort of filtering the noise out of the database or enlarging the information contained thereof is also analyzed. We close the paper with a summary and the conclusions.

II. SIMULATED DATABASES AND TYPES OF NUMERICAL INACCURACIES

Simulated databases were constructed *ex professo* for a binary alloy with the following effective Hamiltonian

$$F = 4.0\bar{\Phi}_2^1 + \frac{3}{2}\bar{\Phi}_2^2 - 2.0\bar{\Phi}_2^3 + 3.0\bar{\Phi}_3^1 - \frac{6}{5}\bar{\Phi}_4^1, \tag{3}$$

where $\bar{\Phi}_m^n$ is the nth m-body correlation function, i.e., the configurational average of the cluster function Φ_m^n (defined as the m product of occupation variables).[8,16] Thus, the first three terms in the right-hand side of Eq. (3) correspond to pair interactions while the last two are associated to most compact three and four-body interactions in a bcc-based alloy. Our choice for the underlying lattice simply reflects the facts that (1) bcc-based alloys have been less investigated than fcc-based systems and (2) a variety of alloys with technological applications crystallize in a bcc environment for a wide range of compositions, e.g., the Ti-based gum metals,[25,26] the refractory alloys from the Mo, Ta, W, Nb quartet,[27] or the magnetically relevant Fe-Co alloys.[13,28]

We have built several databases by evaluating F (here the formation enthalpy in mRy/atom) for different sets of ordered configurations, i.e., different sets of $\bar{\Phi}_m^n$. Two distinct approaches were followed: the first and more traditional one, was to collect the naturally appearing bcc-based structures with unit cells of moderate size and then complement this set with superlattices in the (001), (110), and (111) directions. This approach yielded a set of 80 ordered bcc-based structures with unit-cell sizes ranging between 2 and 20 atoms. All the elements of the 80 database have been used previously in first-principles cluster expansions of binary alloys.[13,27,28]

A second set of ordered structures was produced by generating all 629 irreducible derivative superstructures with eight or less atoms in the bcc-based unit cell.[29,30] The 629 database was then partitioned into the 160 and 320 databases containing the first 160 and 320 (sorted in increasing number of atoms in the unit cell) structures of the 629 database. These short-ranged structure sets allow for a systematic investigation of the impact of the configurational information contained in a given set (see below).

The breadth of a database, in terms of the concentration span and the number of structures, offers a heuristic estimation of the information contained therein. It is clear that the 629 database contains far more configurational information than the 80 database [cf. Fig. 1(a)]. However, the density of ordered structures [Fig. 1(b)] shows that the 80-database set offers a comparable sampling (in terms of sparsity although obviously definitely far more limited in absolute terms) to the 629 database. In other words, the 80 database offers a good compromise between size and information and as an alternative when building up a large database is impractical, e.g., when the calculations are too complex or the experimental data are limited.

Since these training sets have been constructed from the *a priori* knowledge of the ECIs, we can use such data sets as testing grounds for inversion algorithms. An optimal inversion procedure should retrieve the true Hamiltonian (3) even

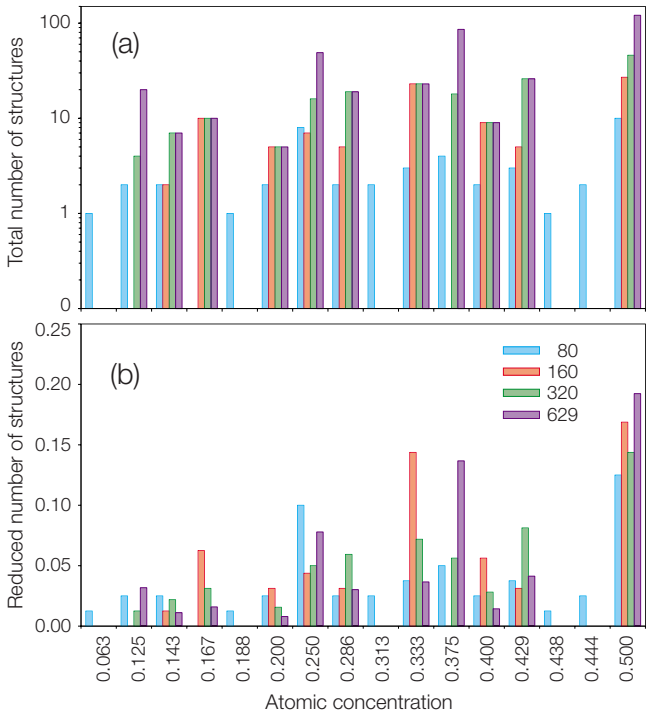

FIG. 1. (Color online) (a) Total number of ordered structures as a function of the atomic concentration for several databases. The 80 database was constructed in the traditional way, i.e., collecting naturally occurring ordered structures of moderate size whereas the 629 database contains all ordered structures up to eight atoms per unit cell. The 160 and 320 databases contain the smallest 160 and 320 ordered structures of 629 database. (b) Reduced number of structures (normalized to the total number in each set) as a function of the atomic concentration. The number of structures in a database is symmetric around equiatomic concentration.

in the case of significant imprecisions in the database. Imprecisions or numerical noise in a database can arise from a variety of situations and sources, from convergence of the first-principles calculations to rounding errors in handling the databases to instrument saturation effects in collecting experimental data. Characterizing all possible types of numerical errors falls out the scope of this paper. Instead, we are interested in investigating how commonly appearing *random* and *systematic* errors travel through the inversion process. To that aim, we have chosen rounding imprecisions as an example of latter

$$R(F,\alpha) = \alpha \ \text{floor}\left(\frac{F}{\alpha} + \frac{1}{2}\right), \tag{4}$$

where floor (x) returns the integer part of x.[31,32] The net effect of $R(F,\alpha)$ is to collapse the data into a discrete number of values as $\alpha \to 1$ as seen in Fig. 2.

A more severe type of systematic imprecisions are saturation errors. Typical saturation functions are of the type

$$S(F,\alpha) = \frac{F_{\max}}{2}\left(1 + \frac{\tanh \alpha\tilde{F}}{\tanh \alpha}\right) + \frac{F_{\min}}{2}\left(1 - \frac{\tanh \alpha\tilde{F}}{\tanh \alpha}\right), \tag{5a}$$

$$\tilde{F} = 2\left(\frac{F - F_{\min}}{F_{\max} - F_{\min}}\right) - 1, \tag{5b}$$

where $\alpha(>0)$ and $F_{\max}(F_{\min})$ is the maximum (minimum) value of F in the database. Figure 3 displays the database after being transformed according to Eq. (5) for different values of α. Notice that large values of α erase the configurational dependence by collapsing the data in two bands corresponding to their extremal values (i.e., −8.0 and 1.0 mRy/atom).

On the other hand, the more general type of random imprecision is an additive Gaussian-distributed noise

$$G(F,\sigma) = F + X(\sigma), \tag{6}$$

where X is a random (real) variable, normally distributed with zero mean and standard deviation σ, i.e., with a probability density function

$$P(X) = \frac{\exp(-X^2/2\sigma^2)}{\sigma\sqrt{2\pi}}. \tag{7}$$

Figure 4 shows the Gaussian treated vs the exact database for several values of σ. As we shall see in the next sections, the lack of hard bounds and the random nature for this type of errors make them difficult to handle in the inversion process.

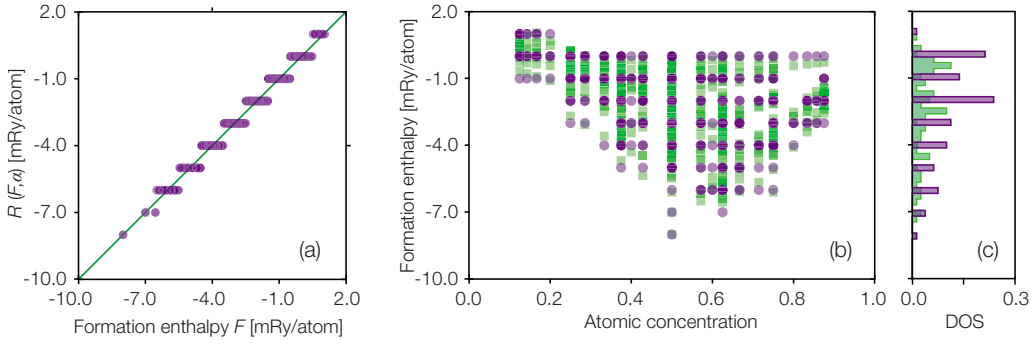

FIG. 2. (Color online) (a) $R(F,\alpha)$ vs F for the 629 database for $\alpha=1.0$. (b) Exact (squares) and rounded (circles) values for the formation enthalpy as a function of the atomic concentration. (c) Density of states (DOS), i.e., normalized number of structures per enthalpy of formation, for both the exact and noised data.

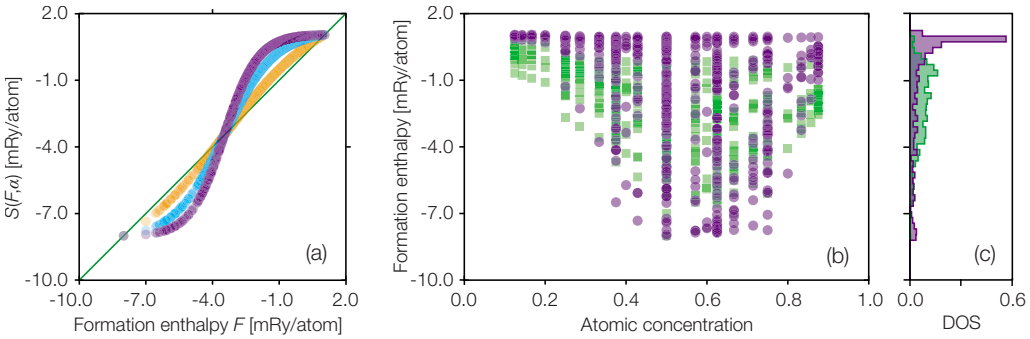

FIG. 3. (Color online) (a) $S(F,\alpha)$ vs F for the 629 database for $\alpha=1.0$, 2.0, and 3.0. (b) Exact (squares) and saturated (circles) values for the formation enthalpy as a function of the atomic concentration. The saturated data corresponds to the extreme case of $\alpha=3$. (c) DOS, i.e., normalized number of structures per enthalpy of formation, for both the exact and noised data ($\alpha=3$).

III. CONSTRUCTING THE MAP

In its most popular incarnation, the cluster expansion renders a map where ECIs are independent of the concentration [cf. the f_α in Eq. (3)]. This is a very attractive characteristic for many applications such as phase diagram calculations or exhaustive configurational sampling to optimize a given physical property. An expansion in terms of concentration-independent ECIs has, however, the disadvantage of lacking a convergence radius, i.e., the relevant ECIs associated to given cluster figures do not follow any preordained compactness or decay criteria. This is a major issue of the method and several approaches to determine the relevant terms in Eq. (1) have been proposed over the years. Early truncation of the expansion can lead to flagrant errors and several examples of this have been documented in the past.[10,11,33] Hierarchical routes,[34,35] mostly inspired by the success of the cluster variational method (CVM) in lattice systems,[36–40] have also been advanced, although it is not clear whether these cluster expansions *à la* CVM lead to converged expansions.[24]

Contemporary approaches to the cluster-expansion method rely on statistically approximating Eq. (1) in the *least unbiased* way. This is achieved by selecting the expansion terms from an *undesigned* set of cluster figures. Usually this designerless cluster set 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).[23,24] Variational or evolutionary approaches (see below) can be used to construct the configurational map by selecting the terms appearing in the expansion and minimizing

$$\Delta^2_{\mathrm{Fit}} = \frac{1}{N_s}\sum_{\ell=1}^{N_s}\left[T(F_\ell) - \sum_\alpha f_\alpha \Phi_\alpha\right]^2 \tag{8}$$

for a database with N_s entries (e.g., ordered structures) and where the cluster figures α are selected from an undesigned pool. Notice that Eq. (8) explicitly considers the imprecisions in F via the transformation function $T=G$, S, or R.

Equation (8) defines the goodness-of-fit, thus guaranteeing that the effective Hamiltonian reproduces all the information in the training set ($\ell=1,\ldots,N_s$). The forecasting abilities, however, are not warranted by Eq. (8) but by the *optimized* N_v-out cross-validatory estimation of the prediction error[20]

$$\Delta^2_{\mathrm{Pred}} = \frac{1}{L_v N_v}\sum_{i=1}^{L_v}\sum_{\ell}\left[T(F_\ell) - \sum_\alpha f^i_\alpha \Phi_\alpha\right]^2, \tag{9}$$

where the inversion is performed using a construction set of size $N_c=N_s-N_v$. The prediction power of the expansion is then tested against the remaining structures (not used in the fit) for all L_v possible validation sets of size N_v.

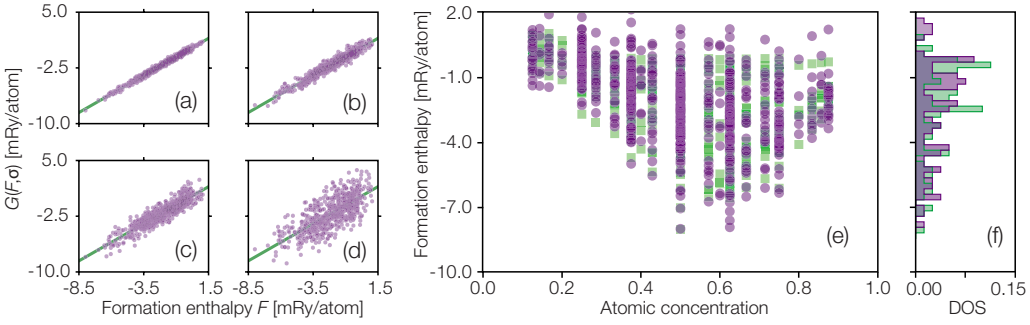

FIG. 4. (Color online) [(a)–(d)] $G(F,\sigma)$ vs F for the 629 database with $\sigma=0.2$, 0.43, 0.79, and 1.6 mRy/atom, respectively. (e) Enthalpy of formation vs atomic concentration for the exact and noised ($\sigma=0.79$ mRy/atom) databases. (f) DOS, i.e., normalized number of structures per enthalpy of formation, for both the exact and noised data ($\sigma=0.79$ mRy/atom).

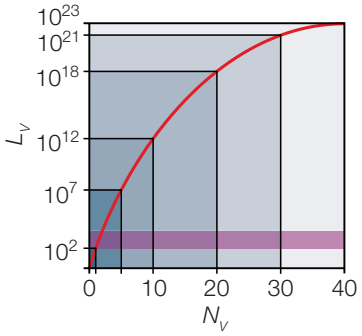

FIG. 5. (Color online) Number of validations sets L_v as a function of the size of the set N_v for a database with $N_s=80$ entries (solid line). Using smaller subset of L_v (solid band) provides statistically comparable estimations of the prediction error.

Direct optimization of the prediction error for moderate databases ($N_s \sim 80$) and cluster-pool sizes (e.g., $N_{\mathrm{CP}}=50$) in a conservative leave-many-out regime ($N_v=10$) renders search spaces of the order of Avogadro's number. This is an astronomical number that calls for smart optimization methods and practical approximations. Before discussing a couple of the former, we would like to draw the attention to the number of validation sets $L_v=N_s!/N_v!N_c!$ depicted in Fig. 5 for $N_s=80$ as a function of N_v (the size of the validation set). For $N_v=10$ and $N_s=80$, a typical case discussed in the rest of the paper, $L_v \sim 10^{12}$.

Notice, however, that what mainly improves the estimation of the prediction error in Eq. (9) is N_v and not L_v. The size of the validation set accounts for the amplitude of the fluctuations in the different fittings due to inaccuracies in the database. Including all L_v possibilities only guarantees that all fluctuations are accounted. Therefore, it is reasonable to assume that a much-smaller-sized subset of L_v could render statistically comparable estimations of the prediction error. Our extensive and systematic inversions have shown that this is indeed the case and reliable estimations for the leave-many-out cross validation attained with small number of validation sets ($\sim 10^3-10^4$).[41]

On the other hand, selecting the relevant terms in Eq. (1) that optimize both the fitting and prediction errors is a complicated task. First, the expansion terms are highly correlated, that is, unless all the relevant terms actually appear in the expansion, a least-square fit cannot distinguish between real and fictitious contributions. Second, many terms in Eq. (1) have zero ECIs, turning the CE into a subset-model-selection problem,[42,43] i.e., the optimal CE have to be selected from the $2^{N_{\mathrm{CP}}}-1$ possible expansions compatible with a cluster pool of size N_{CP}.[22] In this paper, we have used two methods that solve this problem by optimizing the entire cluster pool at once in a variational way (the variational cluster expansion) or by selecting the relevant terms following an evolutionary approach (genetic algorithm).

A. Variational approach

The variational cluster expansion (VCX) (Refs. 22 and 44) optimizes both the fitting and the prediction error in a variational way for the entire cluster pool. This is accomplished by turning the discrete ECIs into functions of a continuous variable $\boldsymbol{w}=\{w_1,w_2,\ldots,w_{N_c}\}$ via the penalized goodness-of-fit

$$\Delta^2_{\mathrm{VCX}}=\frac{1}{N-N_v}\left\{\sum_{\ell}\left[T(F_\ell)-\sum_{\alpha}f_\alpha\Phi_\alpha\right]^2+\sum_{\alpha}(w_\alpha f_\alpha)^2\right\}. \tag{10}$$

The last term in the right-hand side links the fitting with the prediction error in Eq. (9) since now $f_\alpha=f_\alpha(\boldsymbol{w})$. The form of the penalty term is not unique and other functional relationships can be entertained as long as they are continuous and differentiable functions with a minimum in the weights $\boldsymbol{w}$ domain.

The global minimum of the prediction error Δ^2_{Pred}, now a functional of the weights, is formally achieved as

$$\frac{\partial\Delta^2_{\mathrm{Pred}}}{\partial\boldsymbol{w}}=0. \tag{11}$$

In the praxis, one starts with an arbitrary choice of weights w_α's that are used to determine the corresponding set of f_α's through Eq. (10). This set of ECIs is then used in the evaluation of the prediction error Eq. (9). Since the ECIs, Δ^2_{VCX}, and Δ^2_{Pred} are continuous functions of the weights, fast numerical routines can be employed. Equation (11) renders large (small) values for weights associated with nonrelevant (relevant) cluster figures. An optimal expansion can be defined as the one containing only relevant terms (i.e., with $\boldsymbol{w}\sim\mathbf{0}$), and therefore, satisfying the following condition:

$$\min|\Delta^2_{\mathrm{VCX}}-\Delta^2_{\mathrm{Fit}}|. \tag{12}$$

This is an interesting property useful in dealing with databases containing significant numerical imprecisions.

On the other hand, the irrelevant terms of an expansion can be removed using backward-reduction techniques,[42] i.e., the initial cluster pool of size N_{CP} is decimated into a subpool of size $N_{\mathrm{CP}}-1$ by removing a cluster figure (term) from expansion such that the prediction error for the $N_{\mathrm{CP}}-1$ expansion increases the least, iteratively until the remainder of the pool reaches a prescribed size. Because of its variational nature, a VCX expansion associated with N_{CP} produces better or equivalent forecasts than one with $N_{\mathrm{CP}}-1$.

B. Evolutionary approach

For practical reasons, the cluster expansion must always be truncated. The VCX above is one approach to truncation. Other systematic truncations, based on a variational approach, have been advocated (van de Walle[45,46] and Zarkevich and Johnson[34]) but they converge rather slowly—often the important terms come late in the hierarchy. The choice of whether or not to include a particular cluster in the expansion is a "yes-no" question so the truncation problem has a discrete solution space. The natural correlation of the problem, the discrete nature of the solution space, and the astronomic size of the solution space make the problem intractable by gradient-based methods and ill suited to simulated annealing.

The evolutionary approach is a favorable solution because it is well suited to highly correlated problems and can find near optimal solutions while exploring only a tiny fraction of the solution space. The evolutionary approach seeks the n most important clusters, selecting from a relatively large pool of clusters, m, perhaps several hundred. The size of the search space it explores then is $\binom{m}{n}$. For a typical pool of several hundred clusters and an expansion with a few dozen terms, the search space is bigger than Avagadro's number—far too large for a direct search (see Fig. 5).

The evolutionary approach searches for the optimal cluster expansion via a genetic algorithm. Candidate solutions (individuals) are generated randomly at first and evaluated for their predictive power (fitness score). The best solutions are combined (mating) in subsequent iterations (generations) to yield improved solutions (offspring). For our tests with noise-added data in this paper, the evolutionary approach used a cluster pool of the smallest 27 pair clusters, 58 triplets, and 15 four-vertex clusters (100 clusters total). The population size was 54 individuals and ran for 100 generations. In each generation, 40 children solutions were created, the top four replacing the four least-best parent solutions.

IV. SIMULATED INVERSION

We have used the databases generated in Sec. II aiming to develop systematics that can be applied later on in real-alloy scenarios where the level or type of numerical imprecision is unknown. Along this line, the first question to be addressed is the one of the inversion performance using databases free from numerical errors. In all cases, as long as the true cluster figures were contained in the cluster pool, the true configurational Hamiltonian [i.e., Eq. (3)] was always retrieved, irrespective of the (i) database size, (ii) approach followed to build the map (VCX or GA), or (iii) the cross-validatory estimation for the prediction error, i.e., leave-one-out cross validations turned to be good enough. The issue of multiple solutions was not found in our inversions with the exact (noise-free) data.

A. Systematic errors

1. Rounding

Rounding off is, perhaps, the simplest and most frequently encountered numerical inaccuracy in databases. Increasing the α value in the error transformation $R(F,\alpha)$ [Eq. (4)] erases the configurational dependence in the database by collapsing neighboring data points to a common value. For example, $\alpha=0.1$ will fold data up to the first decimal point while $\alpha=1.0$ will bring all points to the nearest integer value [see the formation enthalpy of Fig. 2(b) or the density of states in Fig. 2(c)]. Figure 6 shows the leave-one-out prediction error as a function of the number of terms in the expansion for the 80 database and $\alpha=1$. The optimal expansion [see Eq. (12)] satisfies the condition

$$\Delta^2_{\text{Pred}} > \sigma_R(\alpha)^2, \tag{13}$$

where $\sigma_R(\alpha)$ is the standard deviation of $R(x,\alpha)$ with x uniformly distributed in $[-\alpha/2,\alpha/2]$, that is,

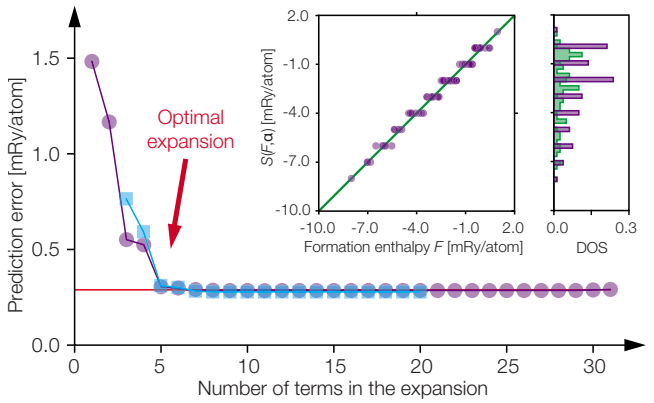

FIG. 6. (Color online) Prediction error as a function of the number of terms in the expansion as calculated by the VCX (circles) and GA (squares). The 80 database was rounded off to the integer part, i.e., $R(F,\alpha=1)$ (see insets). The optimal expansion has the least number of terms (five) and it is closer to the limiting error $(2\sqrt{3})^{-1}\approx 0.289$ (horizontal line).

$$\sigma_R(\alpha) = \frac{\alpha}{2\sqrt{3}}. \tag{14}$$

A good way to think about σ_R is as the extent of the noise introduced in the database through rounding. Criterion in Eq. (13) states that an optimal expansion has a prediction power not better than the noise in the database. In other words, an expansion that can discriminate features in the database that are smaller than the noise level, is certainly "fitting" the noise as a configurational degree of freedom.

Applying criterion in Eq. (13) to the data in Fig. 6, i.e., $R(F,\alpha=1)$, retrieved an expansion with the five correct cluster figures and a cross-validation score of 0.31 mRy/atom for both the VCX and GA. The retrieved ECIs have a mean-squared error (MSE) of 0.01 mRy/atom and 0.05 mRy/atom for the VCX and GA, respectively. The retrieved ECIs can be brought down to virtually the exact ones by increasing the size of the database N_s. For instance, using the VCX on the 629 database together with a leave-one-out cross validation renders an ECI-MSE of 0.004 mRy/atom. On the other hand, improving the estimation of the prediction error by increasing the size of the validation set N_v, seemed to have almost no effect on the quality ECI-MSE. Inverting the rounded $(\alpha=1)$ 160 database using the VCX rendered ECI-MSE's equal to 0.0123, 0.0123, 0.0123, 0.0123, and 0.0125 for $N_v=1$, 5, 10, 20, and 40, respectively. The ECI-MSE was also independent of the number of validation sets $L_v(\geq 5000)$ for both the VCX and GA. For databases with rounding imprecisions characterized by $\alpha<1$, the retrieved ECIs and cluster figures were virtually indistinguishable from the exact ones irrespective of the method used (VCX or GA).

This behavior can be understood as follows: rounding imprecisions erase the configurational fine structure by closing together neighboring F into a common value $R(F,\alpha)$, for example, the integer part when $\alpha=1$. This type of noise is "sharp" (very well bounded) and therefore a validation set of size $N_v=1$ will capture the same fluctuations of the ECIs as, say, $N_v=5$ or 10. In other words, asking more questions on a

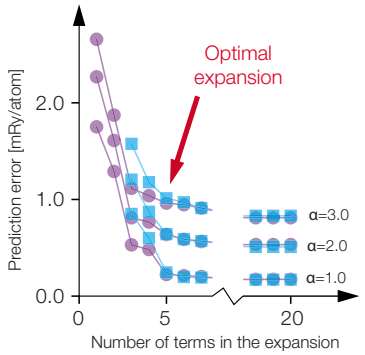

FIG. 7. (Color online) Prediction error (leave-one-out cross validation) as a function of the number of terms in the expansion as calculated by the VCX (circles) and GA (squares). The 80 database was saturated using transformation $S(F,\alpha)$ for $\alpha=1.0$, 2.0, and 3.0.

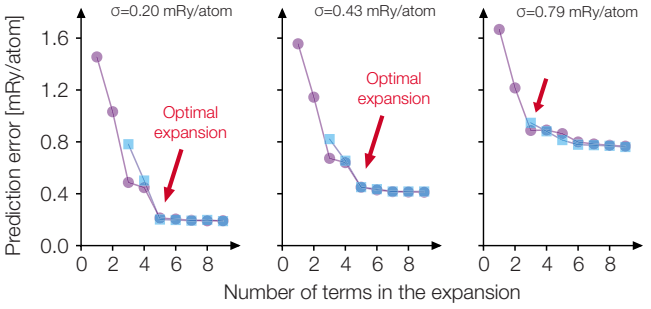

FIG. 8. (Color online) Prediction error (leave-one-out cross validation) as a function of the number of terms in the expansion as calculated by the VCX (circles) and GA (squares). The 80 database was randomly noised using transformation $G(F,\sigma)$ for $\sigma=0.20$, 0.43, and 0.79 mRy/atom.

rounded database does not bring different answers since it is always the same $R(F,\alpha)$. In any case, it is remarkable that one can recover the true Hamiltonian from a relative small database with such strong rounding off imprecisions (i.e., $\alpha=1$ in R).

2. Saturation

A drastic decrease in performance is expected when databases are saturated using Eq. (5). For large values of α, the data collapses into lower and upper bands corresponding to the lowest and highest values of F (see Fig. 3), thus compromising the selectivity power of any map-learning algorithm. It is important to emphasize that a database saturated with $S(F,\alpha=3)$ is already a very noisy training set.

For low levels of saturation (i.e., $0<\alpha\leq 2$) both the GA and VCX retrieve the true Hamiltonian when inverting the 80 database using leave-one-out cross-validatory estimation for the prediction error. In particular, for $\alpha=2$, the ECI-MSE is 0.147 and 0.370 mRy/atom for the VCX and GA, respectively. Both the GA and VCX predicted optimal expansions having the same cross-validation score of 0.64 mRy/atom. The behavior of the prediction error vs the number of terms in the expansion is the same for long expansions. The predicted expansions of the VCX and GA branch out at the true expansion (five terms) as seen in Fig. 7, similarly to what happens in the case of rounding imprecisions (cf. Fig. 6).

Interestingly enough, the leave-one out ECI-MSE increased from 0.147 to 0.170 to 0.184 to 0.201 mRy/atom as the database size increased from $N_s=80$ to 160 to 320 to 629, respectively. For a high saturation level ($\alpha=3$), the inversion process by the VCX did not yield the true expansion for the 80 database. In this case, the true expansion was retrieved by the VCX only when using larger (i.e., the 160 database or a bigger) databases.

It is important to note that the GA approach identified a five-term expansion with the true cluster figures for all saturated databases investigated in this paper. In this sense, the GA approach is superior to the VCX for extremely saturated databases. However, due its native stochastic nature, the GA approach renders almost a continuously evolving family of solutions, i.e., there are no sharp jumps in the prediction error vs the number of terms in the expansion (see Fig. 7). This fact complicates the selection of the optimal expansion. We shall discuss this point further in Sec. V C.

Saturation errors can be considered as an extreme case of rounding, i.e., where the data is rounded up and down to the extremal values of the database. This explains why increasing N_v in the cross-validatory estimation of the prediction error has no effect on the quality of the expansions for either the VCX or GA approaches. The fact that the upper and lower saturation bands depend on the system at hand, i.e., on the distribution of F as a function of the configuration $\boldsymbol{\sigma}$, makes further analytical treatment difficult to achieve (function S is, in general, unknown for real databases). Databases with a higher (lower) density of structures in the middle of the upper and lower saturation bands will saturate less (more) as function of α. One can envisage alternative approaches to handle saturated databases where the data is partitioned into bins and analyzed independently.

B. Random errors

The working databases have been transformed using Eq. (6) thus producing representative random imprecisions by adding Gaussian-distributed noise (see Fig. 4). The additive Gaussian noise is characterized by standard deviation σ. However, it is illustrative to compare such σ value with physical quantities. A straightforward comparison can be drawn by considering that our databases are produced from the effective Hamiltonian (3) that shows the lowest-energy value of −8.0 mRy/atom for the B2 CsCl-type structure. Therefore, adding Gaussian noise characterized by $\sigma=0.20$, 0.43, 0.79, and 1.66 mRy/atom represents fluctuations on the order of 5% (2.5%), 11% (5.5%), 20% (10%), and 42% (21%) for the highest ECI (lowest-energy value), respectively.

The true Hamiltonian (i.e., cluster figures and ECIs) was recovered for $0<\sigma\leq 0.43$ using the relative small 80 database together with a leave-one-out cross-validatory estimation of the prediction error (see Fig. 8). In particular, for $\sigma=0.43$ mRy/atom the prediction error of the optimal expansion selected by the VCX (GA) is 0.45 mRy/atom (0.45 mRy/atom) with an ECI-MSE of 0.024 mRy/atom (0.078 mRy/atom). This is a very satisfactory result, considering that typical first-principles formation enthalpies are expected

to be accurate within few tenths of mRy/atom.

Our results for the VCX and GA showed that small databases with larger imprecisions (i.e., $\sigma=0.79$ or 1.66 Ry/atom) do not contain enough information to resolve the true Hamiltonian. For example, the smallest database that rendered the true expansion for $\sigma=0.79$ contained 160 structures and for $\sigma=1.66$ it was necessary to contemplate inverting a database as large as 629 structures.

Contrary to our initial expectation and widespread common assumption, the inversion process appears to be insensitive to the level of cross validation, i.e., the improvement is only *quantitative* and within few percent when increasing N_v from 1 to 40 (with L_v's as large as 5×10^5). However, all the figures of merit, i.e., ECIs-MSE and residual of the expansion, converged steadily as we systematically increased the database size from 80 to 160 to 320 to 629 structures. In particular, the residual of the expansion (that is, the fitting error of the optimal expansion) converged smoothly to the noise level in every case, thus providing a selection criterion for the optimal expansion.

V. DISCUSSION

A. How large should the database be?

An important question when producing cluster expansions from first-principles is "how large should the database be?" The conventional answer to this question is to cluster expand a small trial database and to use the cluster-expansion Hamiltonian to search for the ground-state line. Whenever new structures are predicted, they are directly calculated and incorporated into the original database. A new expansion is then performed followed by a second prediction for the ground-state line. This process is to be continued until no new ground-state structures are found. This iterative procedure has been successfully applied to the determination of ground and near-ground-state structures and to the prediction of finite-temperature properties. However, it has been noted that such a process cannot be fully applied for physical observables other than the energy. For instance, there is not a direct (simple) correlation between extrema in the magnetization or semiconductor band gap and the ground states in the system.[22] This notion has been recently reinforced by Seko *et al.*,[47] who have pointed out that converging a cluster expansion using the conventional iterative procedure does not necessarily produce an optimal expansion since the prediction error is minimized only for the ground- and near-ground-state parts of the configurational space. The proposal of Seko *et al.* is to choose a database that samples the configurational space as much as possible. Although, in principle, this is the correct answer, the question still persists, now in the form of "how dense does this sampling of the configurational space need to be?"

In this paper, we have approached this question using two different methods of sampling the configurational space. The first one collected the naturally appearing structures with unit cells of moderate size (the largest unit cell considered in this approach contained 20 atoms) and it was complemented with low-Miller-index superlattice structures. Our rationale was that nature somehow already sampled the configurational space and rendered the most relevant cases. Of course, this approach is limited by the experimental success in characterizing binary compounds. Nevertheless, the first approach yielded a database with 80 entries that nicely spanned all the concentration range of bcc-based alloys. The second way of sampling the configurational space was a mathematical one, generating all irreducible derivative structures consistent with a given size of the unit cell (629 for a bcc-based binary alloy with unit cells containing up to eight atoms). Our underlying idea was that, contrary to the former case, the databases produced in this way were completely unbiased, limited only by the maximum number of atoms in unit cell that defines the mesh size in configurational space.

We found that the answer to the question posed here depends on how accurate, numerically speaking, the database is. In other words, optimal expansions can be easily attained with rather small *accurate* databases. However, we also found that applying modern cluster expansion techniques could recover the true Hamiltonian underlying a database even when such database contains large numerical imprecisions.

By introducing controlled systematic and random numerical noise in the working databases generated from a known Hamiltonian, we were able to gauge the performance of the inversion process and error propagation thereof. In particular, we found that databases treated with systematic rounding and saturation noise can be successfully inverted using the VCX or GA approaches even when using a least-square fit (LSF) to estimate the goodness-of-fit and prediction errors. This is a remarkable result because LSFs should work, in principle, only for a Gaussian distribution of errors. Databases containing errors as large as 10% of the largest value of the formation enthalpy were inverted successfully, yielding the true Hamiltonian, that is, the correct cluster figures and effective cluster interactions.

An interesting and quite unexpected finding was the insensitivity of the inversion process and quality thereof to the number of outs N_v. Contrary to the widespread assumption, we found that increasing N_v only translates into *quantitative* changes, that is, it reduces the mean-square error of the ECIs and converges the fitting error toward the noise level but it does not change the *qualitative* behavior of the inversion process. In other words, during our extensive analysis with the VCX and GA, whenever the true Hamiltonian was not retrieved for small N_v, increasing N_v (up to half of the database size) did not improve the inversion process. However, we found that increasing the database size in a systematic way did improve both qualitatively and quantitatively the cluster expansion.

B. Prediction error, noise level, and truncating the expansion

There is a second question that relates to the selection of the optimal expansion once a proper database has been inverted: is the optimal expansion one with the lowest possible prediction error? The conventional wisdom will point to an affirmative answer; after all whichever computational tool used to produce the database will be used for *a posteriori* validation of the model. The justification for this "cross

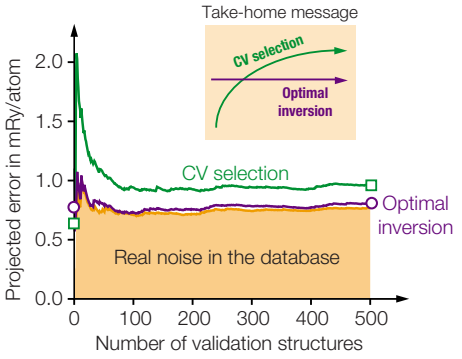

FIG. 9. (Color online) Real prediction error as a function of the validation set size. The projected expansion was determined using a Gaussian-noised database with σ=0.79 mRy/atom. The as-expanded prediction errors for the cross-validation and optimal selections are marked with a square and a circle, respectively. The evolution of these two expansions is followed as we increase the size of a validation database (treated with the same noise level).

validation-selection" (cv) is that an expansion that minimizes the prediction error even below the noise level, will only reproduce further fluctuations making this a suitable expansion to explore the full configurational space either by exhaustive enumeration or temperature effects.

The results obtained in this paper show the contrary, i.e., that an expansion with an error below the noise level has, in the long run, lower prediction power than an optimal expansion selected just above the noise level (with higher error). Figure 9 shows two possible expansions for a database treated with random noise. The cv-selection picks an expansion with the lowest cross-validation score (square) whereas the optimal pick (circle) corresponds to an expansion with the lowest prediction error above the noise level (i.e., 0.79 mRy/atom). The evolution of these two expansions is followed as we increase the size of a validation database (treated with the same noise level). Notice that after a few hundred structures the prediction error of the cv-selection surpasses that of the optimal selection. In fact, an important characteristic of the optimal selection is its constant performance, i.e., it is a transferable expansion, truly following the real noise in the validation databases.

It is important to point out that the behavior of the prediction error for both the cv and optimal selections for small databases is incidental, i.e., corresponds to a particular choice of structures. The relevant message, indicated in the inset, is the following: the performance of a cv-selected expansion will change as the number of sampled configurations increases thus compromising the prediction power to describe alloy thermodynamics. On the other hand, the optimal expansion, that is, an expansion that explicitly acknowledges and accounts for numerical noise in the database, has a constant performance. The practical consequences of our work are similar of those of Seko *et al.*[47] but for different reasons. While the work of Seko *et al.*, analyzed the consequences of having an expansion with an error higher than the optimal, we have analyzed the impact of using an expansion with an error below the noise level.

On the other hand, the problem of truncating of the expansion (that is, selecting the figures) has generated debate and several proposed solutions. The two main competing methods are hierarchical approaches (van de Walle,[45,46] and Zarkevich and Johnson[34]) and the variational cluster expansion. In principle, the hierarchal approach of Zarkevich and Johnson provides a formally complete and systematic truncation, much like Fourier Transform, that is variational (in the Rayleigh-Ritz sense), but because the important clusters are often late in the series, and this method requires that all subclusters of smaller size be included, achieving an accurate expansion typically requires hundreds of terms. The evolutionary approach explores all smaller subclusters but is not required to include them if they do not improve the predictive capabilities of the expansion. One reason that this question does not have consensus is the community yet is that the community does not have a common metric for determining the robustness of a fit. In our opinion, the most robust way to measure the goodness of a fit, and the method that seems to be most accepted for testing predictive models (e.g., see Refs. 48 and 49), is leave-many-out cross validation. We have tested the hierarchical approach against our approach and found that *under this metric* our evolutionary approach is more robust.

C. Synergy between the GA and VCX approaches

The concept of an undesigned cluster pool (UCP) is central to unbiased statistical approaches to cluster expansions. Such a cluster pool should contain as many as possible cluster figures. Different approaches to the cluster expansions, e.g., the GA and VCX, differentiate (among other technical details) in the way the UCP is sampled in search of the best expansion. In our paper, we have identified this as a subset-model-selection problem where the $2^{N_{\mathrm{CP}}}-1$ possible expansions (compatible with a cluster pool of size N_{CP}) are tested among themselves using the prediction error as the figure of merit.

The VCX samples the UCP in a "canonical" way, that is, optimizing the entire UCP and then decimating it in order to discriminate the irrelevant cluster figures that otherwise might contaminate the expansion with spurious terms arising from the numerical noise in the database. Once a cluster figure is removed, it will never appear in a shorter expansion. The GA, on the other hand, samples the UCP in a "grand canonical" way, i.e., even when a cluster has been discarded for an N-term expansion, it might appear again in an M-term expansion ($M<N$) as long as it helps to reduce the prediction error.

Whenever the UCP is large enough as to contain all the terms of the true configurational Hamiltonian, both the GA and VCX will produce similar results for expansions containing additional terms to the solution (see the Figs. 6–8 above). The GA and VCX will provide different expansions when the terms in the expansion are less than in the true Hamiltonian. The reason for this behavior is the following: the VCX approach search among the N possible solutions to a $N-1$ expansion based on a reduced CP corresponding to the true cluster figures. The GA, on the other hand, will look on the full UCP and provide an expansion with $N-1$ terms that minimizes the prediction error. If the UCP is large enough,

the VCX and GA will provide different solutions for the $N-1$ expansion. Therefore, the point at which their prediction error vs number of terms in the expansion plots split, singles out the true (best) configurational Hamiltonian.

The different ways of sampling the UCP between the VCX and GA is in fact a positive aspect since it can be used to select the optimal configurational Hamiltonian when the family of possibly expansions (consistent with the noise level in the database) precludes a selection by inspection. The systematic analyses in the previous sections indicate that for a given database, the step-by-step recipe is as follows:

Step (1) Run the GA on a undesigned cluster pool of thousands of cluster figures to decimate the pool to manageable size of several decades (say 20 or 30 cluster figures).

Step (2) Using the GA-decimated cluster pool submit the database to the VCX.

Step (3) Plot the prediction error vs the number of terms in the expansion for both the GA and VCX. The true configurational Hamiltonian corresponds to the most economical expansion where the GA and VCX prediction errors split up. If there is not such a forking point then the noise level in the database is such that the configurational Hamiltonian cannot be unambiguously resolved.

Step (4) Increase the size of the database.

Step (5) Repeat steps 1–4 until a forking point between the GA and VCX prediction errors is found.

It is worth to emphasize that the "forking criterion" can be applied only when the database has reached the critical size for a given configurational noise level. For instance, the 80 database is large enough as to contain sufficient information allowing both the GA and VCX to resolve the true configurational (five term) Hamiltonian. However, upon the increase in the noise level up to $\sigma=0.79$, the 80 database proves to be too small to be optimally inverted by either the VCX or GA methods. The fact that Fig. 8 shows no strict forking point for $\sigma=0.79$ signals a database below the critical size. This issue is solved by increasing the database size to $N_s=160$ for which the GA and VCX prediction error curves show a well-defined forking point. In a sense, the iterative application of the above step-by-step recipe, particularly steps 3 and 4, allows for the self-consistent determination of the configurational numerical imprecisions in a given database. In general, we expect such synergy between different approaches to cluster expansions not only for the VCX and GA methods, as long as such approaches provide statistically unbiased expansions.

VI. CONCLUSIONS

In order to retrieve an optimal and transferable expansion from a numerical database it is critical to analyze the error propagation. The optimal expansion is always an expansion that has a prediction error above the numerical error. The noise level can be determined self-consistently by the systematic increase in the validation set in a cross-validatory scheme, provided that database contains enough information as to resolve the relevant terms. This latter step can only be accomplished by increasing the size of the database. A trade-off between accuracy and size of the database is achieved by sampling all irreducible derivative structures consistent with a given size of the unit cell. Selecting an expansion with a prediction error below the noise level renders an expansion in which prediction power deteriorates as the number of validation structures increases.

Analyses of error propagation and trade-offs between the accuracy of a database and the information contained thereof are pertinent now that packages to perform high-throughput first principles of multicomponent alloys are readily available[50] thus making possible the calculations of databases containing hundreds or even thousands ordered structures in a reasonable amount of time.[2,51,52] In this light, our results underscore the importance of having a large database with moderate or even modest precision over a limited yet highly accurate database in producing optimal (i.e., transferable) expansions from first-principles data.

ACKNOWLEDGMENTS

A.D.O. acknowledges the support of the Alexander von Humboldt Foundation and Texas Advanced Computing Center of the University of Texas at Austin for computing resources. G.L.W.H. is grateful for support from the National Science Foundation (Grants No. DMR-0650406 and No. DMR-0908753) and use of the Fulton Supercomputing Laboratory at Brigham Young University.

*Present address: University of Ulm, Institute for Quantum Information Processing, Albert-Einstein-Allee 11 D-89069 Ulm, Germany, EU.

†Corresponding author; alejandro.diazortiz@gmail.com

[1] M. E. Eberhart and D. P. Clougherty, Nature Mater. **3**, 659 (2004).

[2] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder, Phys. Rev. Lett. **91**, 135503 (2003).

[3] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, Nature Mater. **5**, 641 (2006).

[4] O. A. von Lilienfeld, R. D. Lins, and U. Rothlisberger, Phys. Rev. Lett. **95**, 153002 (2005).

[5] M. Wang, X. Hu, D. N. Beratan, and W. Yang, J. Am. Chem. Soc. **128**, 3228 (2006).

[6] E. Borges, N. Lemes, and J. Braga, Chem. Phys. Lett. **423**, 357 (2006).

[7] J. M. Sanchez and D. de Fontaine, in *Modulated Structures*, edited by J. Cowley, J. Cohen, M. Salamon, and B. Wuensch (American Institute of Physics, New York, 1979), Vol. 53, pp. 133–145.

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

[9] J. M. Sanchez, Phys. Rev. B **48**, 14013 (1993).

[10] D. de Fontaine, Solid State Phys. **47**, 33 (1994).

[11] A. Zunger, in *Statics and Dynamics of Alloy Phase Transitions*, NATO Advanced Studies Institute, Series B:, edited by P. E. A. Turchi and A. Gonis (Plenum Press, New York, 1994), pp. 361–419.
[12] A. Franceschetti, S. V. Dudiy, S. V. Barabash, A. Zunger, J. Xu, and M. van Schilfgaarde, Phys. Rev. Lett. **97**, 047202 (2006).
[13] A. Díaz-Ortiz, R. Drautz, M. Fähnle, H. Dosch, and J. M. Sanchez, Phys. Rev. B **73**, 224208 (2006).
[14] A. van de Walle, Nature Mater. **7**, 455 (2008).
[15] Notice that in this case, the f_α above represent the many-body configuration averages of the potential V.
[16] K. Kawasaki, in *Phase Transitions and Critical Phenomena*, edited by C. Domb and M. S. Green (Academic, New York, 1973), Vol. 2, p. 465.
[17] A. Franceschetti and A. Zunger, Nature (London) **402**, 60 (1999).
[18] M. d'Avezac and A. Zunger, J. Phys.: Condens. Matter **19**, 402201 (2007).
[19] M. Stone, J. R. Stat. Soc. Ser. B (Methodol.) **36**, 111 (1974).
[20] J. Shao, J. Am. Stat. Assoc. **88**, 486 (1993).
[21] A. Díaz-Ortiz and H. Dosch, Phys. Rev. B **76**, 012202 (2007).
[22] A. Díaz-Ortiz, H. Dosch, and R. Drautz, J. Phys.: Condens. Matter **19**, 406206 (2007).
[23] G. L. W. Hart, V. Blum, M. Walorski, and A. Zunger, Nature Mater. **4**, 391 (2005).
[24] V. Blum, G. L. W. Hart, M. J. Walorski, and A. Zunger, Phys. Rev. B **72**, 165113 (2005).
[25] T. Saito, T. Furuta, J.-H. Hwang, S. Kuramoto, K. Nishino, N. Suzuki, R. Chen, A. Yamada, K. Ito, Y. Seno, T. Nonaka, H. Ikehata, N. Nagasako, C. Iwamoto, Y. Ikuhara, and T. Sakuma, Science **300**, 464 (2003).
[26] T. Li, J. W. Morris, Jr., N. Nagasako, S. Kuramoto, and D. C. Chrzan, Phys. Rev. Lett. **98**, 105503 (2007).
[27] V. Blum and A. Zunger, Phys. Rev. B **72**, 020104(R) (2005).
[28] R. Drautz, A. Díaz-Ortiz, M. Fähnle, and H. Dosch, Phys. Rev. Lett. **93**, 067202 (2004).
[29] G. L. W. Hart and R. W. Forcade, Phys. Rev. B **77**, 224115 (2008).
[30] G. L. W. Hart and R. Forcade (unpublished).
[31] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics: A Foundation for Computer Science* (Addison-Wesley, London, 1994).
[32] See also, E. W. Weisstein,"Floor Function," from MathWorld—A Wolfram Web Resource, http://mathworld.wolfram.com/FloorFunction.html
[33] D. B. Laks, L. G. Ferreira, S. Froyen, and A. Zunger, Phys. Rev. B **46**, 12587 (1992).
[34] N. A. Zarkevich and D. D. Johnson, Phys. Rev. Lett. **92**, 255702 (2004).
[35] M. H. F. Sluiter and Y. Kawazoe, Phys. Rev. B **71**, 212201 (2005).
[36] R. Kikuchi, Phys. Rev. **81**, 988 (1951).
[37] J. A. Barker, Proc. R. Soc. London, Ser. A **216**, 45 (1953).
[38] T. Morita, J. Phys. Soc. Jpn. **12**, 753 (1957).
[39] J. M. Sanchez and D. de Fontaine, Phys. Rev. B **17**, 2926 (1978).
[40] D. A. Vul and D. de Fontaine, Mater. Res. Soc. Symp. Proc. **291**, 401 (1993).
[41] In practice, we have determined the number of validation sets used in Eq. (9) by a random construction from the leave-one-out set, i.e., to warrant that all structures are at least once out. To avoid that ill-defined sets spoil the estimation of the prediction error, several random runs are taken at each step. For all cases studied here, this proved to be enough for both the variational and evolutionary approaches to the CE.
[42] A. J. Miller, *Subset Selection in Regression* (Chapman and Hall, London, 1990).
[43] C. R. Rao and Y. Wu, Inst. Math. Stat. Lecture Notes Monogr. Ser. **38**, 1 (2001).
[44] R. Drautz and A. Díaz-Ortiz, Phys. Rev. B **73**, 224207 (2006).
[45] A. van de Walle and G. Ceder, J. Phase Equilib. **23**, 348 (2002).
[46] A. van de Walle, M. Asta, and G. Ceder, CALPHAD: Comput. Coupling Phase Diagrams Thermochem. **26**, 539 (2002).
[47] A. Seko, Y. Koyama, and I. Tanaka, Phys. Rev. B **80**, 165122 (2009).
[48] J. U. Hjorth, *Computer Intensive Statistical Methods* (Chapman and Hall/CRC, London/Cleveland, 1994).
[49] S. M. Weiss and C. A. Kulikowski, *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems, Machine Learning* (Morgan Kaufmann, San Mateo, CA 1991).
[50] S. Curtarolo, AFLOW: *Software for High Throughput Calculation of Material Properties*, 2009, http://materials.duke.edu/aflow.html
[51] S. Curtarolo, D. Morgan, and G. Ceder, CALPHAD: Comput. Coupling Phase Diagrams Thermochem. **29**, 163 (2005).
[52] O. Levy, R. V. Chepulskii, G. L. W. Hart, and S. Curtarolo, J. Am. Chem. Soc. **132**, 833 (2010).