

<https://doi.org/10.1038/s41524-024-01338-y>

The cluster decomposition of the configurational energy of multicomponent alloys

Luis Barroso-Luque^{1,2}✉ & Gerbrand Ceder^{1,2}✉

The cluster expansion method (CEM) is a widely used lattice-based technique in the study of multicomponent alloys. Despite its prevalent use, a clear understanding of expansion terms is lacking. We present a modern mathematical formalism of the CEM and introduce the *cluster decomposition* — a unique and basis-independent decomposition for functions of the atomic configuration in a crystal. We identify the cluster decomposition as an invariant ANOVA decomposition; and demonstrate how functional analysis of variance and sensitivity analysis can be used to interpret interactions among species. Furthermore, we show how the mathematical structure of the cluster decomposition enables numerical evaluation that scales with the number of clusters and is independent of the number of species. Overall, our work enables rigorous interpretations of interactions among species, provides opportunities to explore parameter estimation beyond linear regression, introduces a numerical efficient implementation, and enables analysis of cluster expansions based on established mathematical and statistical principles.

Computational methods based on lattice models are used extensively in the applied physical sciences. The cluster expansion (CE) method provides a mathematical framework for representing and parameterizing generalized lattice models with discrete configurations^{1–3}. The CE method coupled with Monte Carlo (MC) sampling has become an established technique to compute thermodynamic properties of multicomponent crystals^{4,5}. The CE method plays an active role in materials science research, particularly in the study of metallic alloys^{6–8}, semiconductors^{9,10}, superionic conductors^{11,12}, battery electrodes¹³, and surface catalysis¹⁴. Moreover, researchers are continuously developing methodologies that are based on or utilize the CE method. Recent advances have introduced generative models as alternative ways to compute free energies^{15,16}. Additionally, the mathematical formalism of the CE has been used to develop methodological extensions to parameterize functions of continuous degrees of freedom^{17–20}, which can be used to represent vector and tensor material properties^{21,22}, and capture full potential energy landscapes through the recently proposed atomic CE²³.

The core of the CE method is the expansion of a function of configurational variables distributed over a crystallographic lattice. The mathematical formalism of the CE comprises a harmonic expansion of functions over a tensor-product domain²⁴. Intuitively, such a formalism leads to expansions that are generalizations of the

Ising model^{25,26}, which can be expressed as follows,

$$\begin{aligned} H(\boldsymbol{\sigma}) &= \sum_{\beta} J_{\beta} \sum_{\alpha \in \beta} \prod_{i \in [N]} \phi_{\alpha_i}(\boldsymbol{\sigma}_i) \\ &= \sum_{\beta} J_{\beta} \sum_{\alpha \in \beta} \Phi_{\alpha}(\boldsymbol{\sigma}), \end{aligned} \quad (1)$$

where $\boldsymbol{\sigma}$ is a string of occupation variables that represent the chemical species residing on each of N crystallographic sites. α are multi-indices of length equal to N , that are used to label the univariate site functions ϕ_{α_i} . β are sets of symmetrically equivalent multi-indices. And J_{β} are the expansion coefficients which are referred to in the literature as *effective cluster interactions*^{3,5}. The product of site functions over all sites is referred to as a product function or a *cluster basis function*⁵, which we compactly write as Φ_{α} .

The resemblance to the Ising model is evident when considering binary configuration variables $\sigma_i = \pm 1$, for which $\phi_0 = 1$ and $\phi_1(\sigma_i) = \sigma_i$ can be used as a basis. With such a choice, the cluster basis functions Φ_{α} are constructed from products of *spin* variables, and Equation (1) is a direct generalization of the Ising model to higher-degree interactions. Similarly, a binary CE using site indicator functions: $\phi_1(\sigma_i) = \mathbf{I}_{\sigma_i}(\sigma_i)$, is a generalization of the *lattice gas* model²⁷, or a generalization of the q -state Potts model when an over-complete representation is used^{28,29}.

¹Department, Materials Science, 2607 Hearst Ave, Berkeley, 94720 CA, USA. ²Division, Materials Science, 1 Cyclotron Way, Berkeley, 94720 CA, USA.

✉ e-mail: lbluque@berkeley.edu; gceder@berkeley.edu

The connections between classical lattice models and the CE method have been used by practitioners to directly interpret CEs. For example, it is common practice to evaluate the spatial decay of interactions^{30,31} and interpret the effects of chemical interactions between species^{3,12,29} by directly examining expansion coefficients. However, for complex systems with three or more components, coefficient values depend non-trivially on the choice of basis (The number of distinct basis in lattice gas CE sets grows with the number of components. In a Fourier CE, there are infinitely many basis set choices for 3 or more components. In an overcomplete representation²⁸ there are infinitely many expansion coefficient choices that represent a given function.), and so relying solely on intuition from classical lattice models to directly interpret coefficients can be ambiguous and even misleading. As a simple analogy, consider the case of elasticity theory. Although any choice of basis for the representation of stress and strain tensors is suitable for calculations of elastic deformation, much of our understanding of the mechanics of elasticity would be out of reach if we attempted to interpret tensor elements in an arbitrary basis instead of turning to the concept of stress and strain invariants. The CE method has so far been missing the latter.

In this work, we address these ambiguities by demonstrating that any CE that employs an orthonormal basis can be expressed as a unique and basis-independent decomposition, which call the *cluster decomposition* (CD). From the analysis of uniqueness and basis independence, we are able to extract *invariants*—which are properties of a physical system or our approximation thereof, but not of the particular choice of basis. We demonstrate a direct relationship between the CD and well-established expansions of random variables, known as ANOVA or Sobol decompositions³² among other names^{33,34}. We subsequently show that the CD has analytic properties that allow a much deeper understanding of the structure and enable formal interpretations of expansion terms. We then illustrate a practical use case of the CD and use relevant concepts from functional analysis of variance (fANOVA) and sensitivity analysis (SA) to gain mathematically rigorous insight from CD and MC simulations of a ternary alloy. Our presentation establishes the CD as a mathematically formal framework that enables one to analyze the contribution of the interactions among clusters of chemical species to the configurational energy of alloys in a rigorous and unambiguous way.

In Section 2.1, we provide a concise overview of the CE formalism and present the construction of CEs using orthonormal basis functions. In Section 2.1.3, we use the mathematical framework of the CE to motivate the development of a unique and basis-free representation. To do so, we analyze the geometry of orthonormal cluster basis sets and demonstrate that the norms of expansion coefficients associated with the same orbit of crystallographic sites are invariant to the choice of basis. In Section 2.2, we introduce the cluster decomposition (CD), which is a unique decomposition of the configurational energy into multi-body terms. In the CD each orbit of symmetrically equivalent site clusters is represented by a single term, independent of the number of allowed species, in contrast to a CE in which the number of correlation functions per orbit scales with the number of allowed species. In Section 2.2.1, we provide a formal interpretation of the multi-body terms by establishing the CD as a symmetrically invariant functional ANOVA decomposition. In Section 2.2.2, we leverage the nature of the CD to obtain a decomposition of the variance of the configurational energy and use the terms of this decomposition to define sensitivity indices that enable formal ranking and comparisons of the energy contributions of the terms included in a CD. Finally, in Section 2.3, we present illustrative applications of the concepts discussed by means of a brief example involving the CrCoNi medium entropy alloy.

Results

Fourier CEs

We begin by presenting the CE mathematical formalism^{1,35} used to create expansions of the form presented in Equation (2). We present a perspective on the CE formalism which highlights the underlying tensor-product nature of the formalism. Our perspective leads to a more concise, intuitive, and

computationally efficient formulation. We will then use the CE formalism presented to derive the cluster decomposition and prove its invariance and uniqueness in Section 2.2. A summary of the symbols and notation used throughout the paper can be found in Supplementary Note 1.

The CE formalism is concerned with the representation of symmetrically invariant functions of atomic configuration. The CE of a configuration Hamiltonian H is commonly written as follows,

$$H(\sigma) = N \sum_{\beta} m_{\beta} J_{\beta} \sum_{\alpha \in \beta} \frac{1}{m_{\beta} N} \Phi_{\alpha}(\sigma) = N \sum_{\beta} m_{\beta} J_{\beta} \Theta_{\beta}(\sigma), \quad (2)$$

where Θ_{β} are called *correlation functions* and m_{β} are the crystallographic multiplicities per normalizing unit, usually the number of sites in a primitive or unit cell of the disordered structure. N is the number of such units being considered. J_{β} are the *effective cluster interaction* parameters.

The correlation functions Θ_{β} ensure that the expansion in Equation (2) is symmetrically invariant. Furthermore, to enable tractable parameterizations and evaluations, a CE is truncated to include only a small number of relevant terms, including only correlation functions over clusters with a small number of sites and which are physically compact. Finally, expressing the expansion as a density by dividing by N results in the widely used CE of the configurational energy of bulk crystals^{1,5}.

Configuration space and cluster basis functions. The domain of a CE is the space of all possible atomic configurations over a given disordered crystal structure. The space of atomic configurations σ is formally a *product space*, i.e., the Cartesian product of sets Ω_i of allowed species—which we refer to as *site spaces*—associated with each of N crystallographic sites. Each configuration variable $\sigma_i \in \Omega_i$ represents the species occupying the i -th site. The function space over configurations σ is then a *tensor-product* space of the function spaces over each configuration variable σ_i ²⁴.

A set of basis functions Φ_{α} that span the tensor-product space over atomic configurations can be obtained by first constructing a basis for the function spaces over single configuration variables σ_i . Doing so requires finding a total of $n = |\Omega_i|$ linearly independent functions ϕ_{α_i} , $\alpha_i = 0, \dots, n - 1$. An obvious choice is the set of all n indicator functions \mathbf{I}_{σ_i} for each of the allowed species $\sigma_i \in \Omega_i$. However, such a choice does not result in a basis in which the configurational energy can be represented effectively using only a small subset of basis functions, i.e., a basis that allows a sparse or compressible representation.

Following the original CE formulation¹, we can obtain a suitable representation by requiring that (1) one of the basis functions is constant $\phi_0 = 1$, and (2) that the basis be orthonormal under the following inner product^{30,35},

$$\langle \phi_j, \phi_k \rangle = \sum_{\sigma_i \in \Omega_i} \rho_i(\sigma_i) \phi_j(\sigma_i) \phi_k(\sigma_i) \quad (3)$$

where $\rho_i(\sigma_i)$ is an a-priori probability mass function of finding each of the allowed $\sigma_i \in \Omega_i$ on the i -th site. A uniform probability is most often used, but formally it should be equal to the concentration of chemical species in the non-interacting limit³⁰. We call a site basis that satisfies the above two requirements a *standard site basis*.

The basis set over the configuration product space is then given by the tensor product of the single site basis sets^{36,37}. Equivalently, a basis function Φ_{α} is the tensor product of a site basis function from each of the N sites,

$$\Phi_{\alpha} = \bigotimes_{i=1}^N \phi_{\alpha_i}, \quad (4)$$

where the specific basis function taken for each site is indexed by the corresponding element α_i of the multi-index α .

In practice, the product basis functions can be evaluated as N -fold products involving a specified sequence of site basis functions evaluated at the corresponding configuration variable as follows^{1,37},

$$\Phi_\alpha(\sigma) = \prod_{i=1}^N \phi_{\alpha_i}(\sigma_i) \quad (5)$$

If all site basis sets are orthonormal with respect to the inner product given in Equation (3), then it follows that the resulting set of product basis functions Φ_α are orthonormal with respect to the following inner product (a proof is given in Supplementary Note 2)^{24,35},

$$\langle \Phi_\alpha, \Phi_\gamma \rangle = \sum_{\sigma} \rho(\sigma) \Phi_\alpha(\sigma) \Phi_\gamma(\sigma), \quad (6)$$

where the sum is over all possible configurations σ ; and $\rho(\sigma) = \prod_{i=1}^N \rho_i(\sigma_i)$ is the a-priori product probability distribution. The inner product in Equation (6) can be interpreted as an expectation value in the non-interacting limit.

Including $\phi_0 = 1$ in all site bases is necessary so that Equation (2) is a *hierarchical* expansion, and therefore allows sparse or compressible representations. In the resulting expansion, the effective domain of each function Φ_α is the space of occupation variables of a cluster of sites S given by the indices of non-zero elements of the multi-index α , which we refer to as the *support*, $\text{supp}(\alpha)$, of the multi-index. As a result, the product functions Φ_α are cluster functions if and only if $\phi_0 = 1$ in all site basis sets. With this requirement, a cluster function can be written solely in terms of clusters of sites $S = \text{supp}(\alpha)$ and the corresponding non-zero entries of the multi-index α , which we call a *contracted multi-index* $\hat{\alpha}$,

$$\Phi_\alpha(\sigma) = \hat{\Phi}_{\hat{\alpha}}(\sigma_S) = \prod_{i=1}^{|\hat{\alpha}|} \phi_{\alpha_i}(\sigma_{S_i}) \quad (7)$$

Expression (7) makes the effective domain of cluster functions explicit and separates the functional form of a cluster function from the particular cluster of sites it acts on, i.e., cluster functions that operate on symmetrically equivalent clusters have the same functional form (indicated by $\hat{\alpha}$), but differ in their effective domain (indicated by S). We will refer to cluster functions that are constructed using a standard site basis as *Fourier cluster functions*, and a resulting expansion as a *Fourier CE*, in order to distinguish from CEs that use $\phi_0 = 1$ but do not use orthogonal site basis functions.

Since we are working with a discrete and countable domain, site basis functions are nothing more than vectors $\phi_{\alpha_i} \in \mathbb{R}^{n_i}$ (where $n_i = |\Omega_i|$) if we simply treat Ω_i as a sequence by specifying an order for the allowed species. We can therefore represent cluster functions over a cluster S as a Cartesian tensor by computing them using a real vector tensor product (instead of using an N -fold product as in Equations (5) and (7)),

$$\hat{\Phi}_{\hat{\alpha}} = \bigotimes_{i=1}^{|\hat{\alpha}|} \phi_{\alpha_i} \quad (8)$$

$$\hat{\Phi}_{\hat{\alpha}} \in \mathbb{R}^{n_1 \times \dots \times n_{|\hat{\alpha}|}} \quad (9)$$

In doing so, we obtain a practical and numerically efficient implementation, that serves as a complimentary and arguably more intuitive depiction of the mathematical formalism presented thus far. Based on Equation (9) cluster functions can be represented as multi-dimensional arrays, where each dimension is associated with a site in a cluster S ; and the entries along a given dimension correspond to the allowed species at that site. Then, the value of a cluster function Φ_α evaluated at a configuration σ_S is given directly by the value of the corresponding array element.

Symmetry invariance and correlation functions. The final component necessary to construct Fourier CEs of crystalline materials is the

construction of expansion functions of configurations that are invariant to the crystallographic symmetry of the underlying structure. Constructing such symmetry invariant basis functions is achieved by averaging over the action of all symmetry operations in the space group of the underlying disordered crystal structure³⁷.

Symmetry operations transform configurations σ by permuting its elements. Similarly, the application of a symmetry operation on a cluster basis function Φ_α results in a permutation of site basis functions. The permutation of site basis functions in Φ_α can be suitably specified by the corresponding permutation of the elements of its multi-index α . The sets of symmetrically equivalent cluster basis functions can then be identified by the orbits β constructed from permutations of the multi-indices α . The *correlation functions* Θ_β used in the CE in Equation (2) are precisely the average of cluster basis functions Φ_α over orbits β .

However, we present an equivalent but simpler way to construct invariant basis functions by leveraging our approach of treating cluster basis functions $\hat{\Phi}_{\hat{\alpha}}$ explicitly as real space tensors. We introduce the concept of *reduced correlation functions*, which are obtained by averaging cluster basis functions over orbits of symmetrically equivalent contracted multi-indices $\hat{\alpha}$, i.e., only over permutations of site functions over a fixed cluster of sites S . Reduced correlation functions are given as follows,

$$\hat{\Theta}_\beta(\sigma_S) = \frac{1}{\hat{m}_\beta} \sum_{\hat{\alpha} \in \beta} \hat{\Phi}_{\hat{\alpha}}(\sigma_S), \quad (10)$$

where $\hat{\beta}$ is an orbit of symmetrically equivalent contracted multi-indices $\hat{\alpha}$ (permutations of site functions over a fixed cluster of sites S); or equivalently, the set $\hat{\beta}$ can also be obtained by converting all indices $\alpha \in \beta$ to contracted multi-indices $\hat{\alpha}$. $\hat{m}_\beta = |\hat{\beta}|$ is the total number of contracted multi-indices in $\hat{\beta}$.

Reduced correlation functions are also Cartesian tensors (since they are the sum of tensors), and for practical purposes can be precomputed and stored as multi-dimensional arrays for efficient evaluation. In doing so, correlation functions can be efficiently computed by taking averages of the array elements corresponding to the occupancy σ_S of each of the symmetrically equivalent site clusters $S \in B$. Evaluating correlation functions for subsequent prediction using array access is much more time efficient (time complexity $\mathcal{O}(1)$) than computing the $|\hat{\beta}|$ -fold products given in Equation (7) (time complexity $\mathcal{O}(|\hat{\beta}|)$).

The correlation functions used in Equation (2) can then be expressed in terms of reduced correlation functions by averaging over symmetrically equivalent clusters of sites S ,

$$\Theta_\beta(\sigma) = \frac{1}{m_B N} \sum_{S \in B} \hat{\Theta}_\beta(\sigma_S) \quad (11)$$

where B is an orbit of symmetrically equivalent clusters of sites $S \subseteq [N]$, $m_B = |B|/N$ is the site cluster orbit multiplicity per normalizing unit N .

Figure 1 shows a graphical illustration of the relationship amongst the various multi-indices and basis functions we have introduced using a representative triplet cluster of sites in a cubic rocksalt unit cell as an example. Figure 1 depicts the relationship between multi-indices α , contracted multi-indices $\hat{\alpha}$, and cluster basis functions $\hat{\Phi}_{\hat{\alpha}}$. Figure 1 also shows a graphical representation of orbits β of multi-indices α , the corresponding orbits $\hat{\beta}$ of contracted multi-indices $\hat{\alpha}$, and the resulting permutation invariant reduced correlation function $\hat{\Theta}_\beta$.

Invariants in Fourier CEs. We motivate the search for a basis-independent representation of a CE by making a geometric observation. By their orthogonality, standard site basis sets are related by rotations about the hyperplane normal to the function ϕ_0 . This observation is illustrated graphically for a ternary site space in Fig. 2a. Any ternary standard site basis must include two orthogonal basis functions that lie on the plane orthogonal to ϕ_0 .

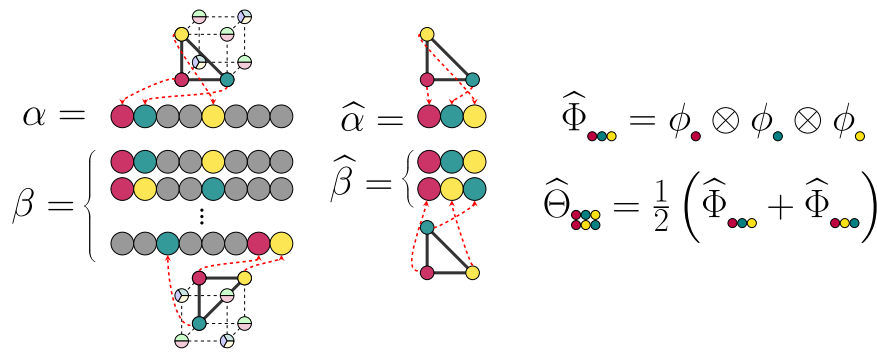


Fig. 1 | Graphical representation of a multi-index α , a corresponding contracted multi-index $\hat{\alpha}$, and the resulting triplet cluster basis function $\hat{\Phi}_\alpha$ over a three-point cluster of sites in a rocksalt unit cell. Graphical representation of an orbit of multi-indices β , a corresponding orbit of contracted multi-indices $\hat{\beta}$, and the resulting triplet reduced correlation function $\hat{\Theta}_\beta$. The different colored spheres

represent multi-index values. Gray spheres correspond to zero values, i.e., constant site functions $\phi_0 = 1$. Translucent color wedges on the remaining sites (those not included in the highlighted cluster) illustrate the partial occupancy, or equivalently, the site basis function choices at each site.

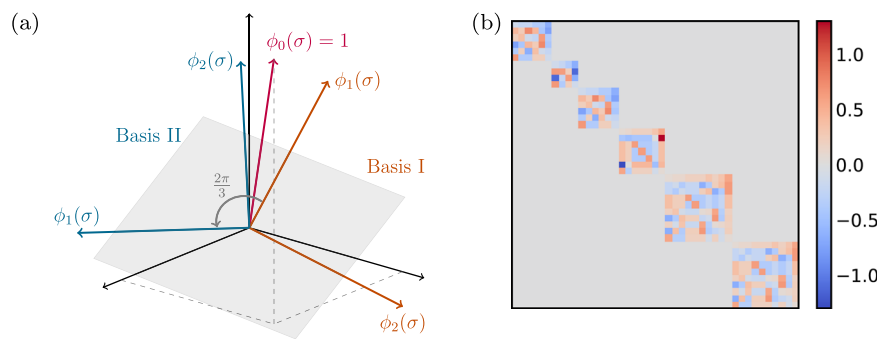


Fig. 2 | Geometry and transformations between Fourier basis sets. a Geometry of standard site basis sets for a ternary site space. Two standard site basis sets related by a rotation of $2\pi/3$ are shown. Both basis sets include the constant ϕ_0 . Any arbitrary

rotation about ϕ_0 results in a distinct standard site basis. **b** Block-diagonal change of basis matrix relating the two different sets of Fourier cluster basis functions up to quadruplets constructed using the site basis sets in **a**.

By considering the geometry of standard site bases, we can show that the change of basis matrix (CBM) between two Fourier cluster basis sets is block-diagonal, and any term connecting cluster functions over symmetrically distinct clusters is zero. A derivation is given in Supplementary Note 3. A visualization of the CBM between two Fourier cluster bases of a ternary system is shown in Fig. 2b. Since each of the diagonal blocks in the CBM is unitary, it follows that the norm of expansion coefficients within each block is invariant to any change of orthogonal basis. In other words, a change of orthogonal basis does not mix cluster basis functions over symmetrically distinct clusters. (In contrast, basis transformations involving a non-orthogonal basis set, such as those using site indicator functions, will mix across symmetrically distinct clusters.)

Formally, these invariance relations can be expressed as a sum of the squares of expansion coefficients associated with a single symmetrically distinct cluster of sites S as follows,

$$\sum_{\gamma: \text{supp}(\gamma)=S} J_\gamma^2 = \sum_{\alpha: \text{supp}(\alpha)=S} J_\alpha^2 \quad (12)$$

where J_γ^2 and J_α^2 are coefficients in the Fourier expansion of a function H using two distinct Fourier cluster basis sets.

Equation (12) applies to any function of configuration regardless of symmetry. When dealing with a symmetrically invariant function, we can group the sums by orbits of symmetrically equivalent site clusters B and

obtain the following invariance relation,

$$\sum_{\eta \in L(B)} \hat{m}_\eta J_\eta^2 = \sum_{\beta \in \hat{L}(B)} \hat{m}_\beta J_\beta^2 \quad (13)$$

where $L(B) = \{\beta: \text{supp}(\alpha) \in B \forall \alpha \in \beta\}$ are sets of orbits β of multi-indices α with symmetrically equivalent supports, i.e., corresponding to the group of correlation basis functions that operate over the same orbit B .

The cluster decomposition

Based on the invariance of expansion coefficients given in Equation (13), we rewrite the CE in Equation (1) in terms of reduced correlation functions grouped by orbits B of site clusters as follows,

$$H(\sigma) = \sum_B \sum_{\hat{\beta} \in \hat{L}(B)} \hat{m}_\beta J_\beta \sum_{S \in B} \hat{\Theta}_\beta(\sigma_S), \quad (14)$$

where $\hat{L}(B)$ are sets of orbits $\hat{\beta}$ of contracted multi-indices representing all symmetrically distinct permutations of site basis functions over the sites in the clusters S —we use the symbol “ L ” to stand for *labeling*, i.e., $\hat{L}(B)$ is the set of symmetrically distinct *labelings* of a cluster of sites $S \in B$ with site basis functions.

The two inner sums in Equation (14) are independent and can be rearranged to obtain a far more physically intuitive many-body expansion that includes only a single term per relevant cluster S irrespective of the number

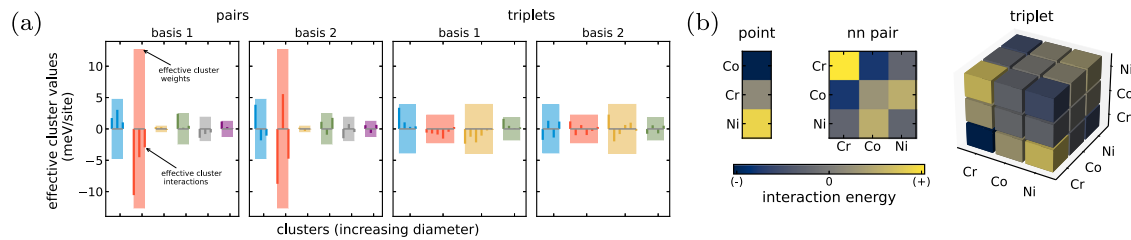


Fig. 3 | Invariance and visualization of cluster interactions. **a** Expansion coefficients (effective cluster interactions) grouped by orbits B of site clusters for two Fourier cluster expansions of the same function of the configurational energy of a CrCoNi alloy using two different basis sets (stems), and the square root of the corresponding effective cluster weights (translucent bars) as defined in Equation (18).

b Visualization of the main effect (point), nearest neighbor pair, and triplet cluster interactions as multi-dimensional arrays for the resulting cluster decomposition of the same CrCoNi alloy used in **a**. Each array entry represents the energy contribution of a single cluster occupied accordingly, i.e., blue (yellow) represents favorable (unfavorable) interactions.

of allowed species,

$$H(\sigma) = \sum_B \sum_{S \in B} \sum_{\beta \in \widehat{L}(B)} \widehat{m}_\beta J_\beta \widehat{\Theta}_\beta(\sigma_S) \quad (15)$$

$$= \sum_B \sum_{S \in B} \widehat{H}_B(\sigma_S) \quad (16)$$

where the n -body terms $\widehat{H}_B(\sigma_S)$ account for the energy originating from the interactions between the species that reside on a cluster $S \in B$. For clusters S with more than one site, $|S| > 1$, we will therefore call these terms *cluster interactions*. (Not to be confused with the expansion coefficients J_β which were named *effective cluster interactions* based on a binary Fourier CE.)

Following the original CE formalism, Equation (16) can also be written as a density by averaging cluster interactions \widehat{H}_B over symmetrically equivalent clusters $S \in B$,

$$H(\sigma) = N \sum_B m_B \left(\frac{1}{m_B N} \sum_{S \in B} \widehat{H}_B(\sigma_S) \right) \quad (17)$$

$$= N \sum_B m_B H_B(\sigma)$$

We will refer to the terms H_B with $|S| > 1$ for all $S \in B$ as *mean cluster interactions*, and as *composition effects* for point clusters ($|S| = 1$). We use P to represent an orbit of single sites ($|S| = 1, \forall S \in P$) when an explicit distinction between point terms and cluster interactions is needed.

We call Equations (16) and (17) the *cluster decomposition* (CD) of the configuration energy $H(\sigma)$. Note that we have purposefully written the CD in terms of reduced correlation functions $\widehat{\Theta}_\beta$ such that the *cluster interactions* \widehat{H}_B are in fact also Cartesian tensors and can be suitably represented as arrays with the same dimensions as the underlying cluster basis functions given in Equation (9).

This tensor representation also facilitates efficient computations and manipulations of the CD. By representing cluster interactions as multi-dimensional arrays, the computational complexity of evaluating the energy of a structure is substantially reduced since there is no longer need to compute the values of correlation functions individually. This practical benefit is significant, as Equations (16) and (17) have the same number of terms for a given crystal structure regardless of the number of species or components. When using a CD for inference, such as MC sampling, the computational time complexity is only a function of the cluster orbits $\mathcal{O}(|B|)$, and is independent of the number of allowed species (see Equation (16)). In contrast, evaluating a CE scales with the number of species $\mathcal{O}(\beta) = \mathcal{O}(\widehat{L}(B) \times B)$ (see Equation (2)), where the number of site basis permutations $|\widehat{L}(B)|$ scales as with the number of species n as $n^{|S|}$ excluding symmetry.

While it is possible to obtain an expression similar to Equation (16) that retains the practical benefits for any choice of site basis, whether orthogonal or not, a true CD can only be obtained from a Fourier CE. This distinction is fundamental because decompositions obtained with CEs that use non-orthogonal basis lack the analytical properties that we discuss in the rest of this work.

It follows directly from Equation (13) that the norm of the cluster interactions \widehat{H}_B are invariant to a change of standard basis, i.e., they are invariant to arbitrary basis rotations orthogonal to ϕ_0 ,

$$\|\widehat{H}_B\|_2^2 = \sum_{\beta \in \widehat{L}(B)} \widehat{m}_\beta J_\beta^2, \quad (18)$$

We call the squared norm of a cluster interaction $\|\widehat{H}_B\|_2^2$ the *effective cluster weight* of a cluster $S \in B$. In addition, we define the *total cluster weight* as the effective cluster weight multiplied by the multiplicity of its orbit, $m_B \|\widehat{H}_B\|_2^2$.

Figure 3a shows CE coefficients grouped by site cluster orbits for two Fourier CEs of the same fit of the configuration energy of a CrCoNi alloy. Figure 3a also shows the square root of the resulting effective cluster weights for both expansions. Unsurprisingly, these values are exactly the same for both expansions since they are precisely the invariants given in Equations (13). Figure 3b also shows a visualization of a point term, nearest neighbor pair, and triplet cluster interactions as multi-dimensional arrays for a CD of the same CrCoNi fit. This compact representation of cluster interactions as multi-dimensional arrays, allows one to read off the interaction energy for a particular cluster occupancy directly from the corresponding array element. For example, we can readily determine which interactions are favorable (negative) and which are unfavorable (positive). Most importantly, since it is basis-independent such interpretations are unambiguous, in contrast to attempting to do so using the expansion coefficients of particular basis sets shown in Fig. 3a.

In addition to their cluster weight invariance, cluster interactions have the following mathematical properties (derivations and proofs are given in Supplementary Note 4):

1. $\langle \widehat{H}_B \rangle = 0$ (zero mean)
2. $\langle \widehat{H}_B, \widehat{H}_D \rangle = 0$ for $B \neq D$ (orthogonal)
3. $\langle \widehat{H}_B, F_D \rangle = 0$ for any set of orbits \mathcal{D} such that $B \notin \mathcal{D}$ and any function $F_{\mathcal{D}}$ that can be expanded using Fourier basis functions Φ_α with $\text{supp}(\alpha) \in \mathcal{D}$ for $D \in \mathcal{D}$. (irreducible)

From properties (1) and (2) it follows that the CD of a given function H is *unique*³⁸, meaning there exists one and only one set of cluster interactions \widehat{H}_B for any given H (a proof is given in Supplementary Note 5). As a result, the cluster weights and cluster interaction values plotted in Fig. 4 are intrinsic properties of the function H and can therefore be considered properties of the physical system (to within the validity of the regression approximation), irrespective of the choice of basis. Furthermore, uniqueness and the properties listed above imply that Equations (16) and (17) are

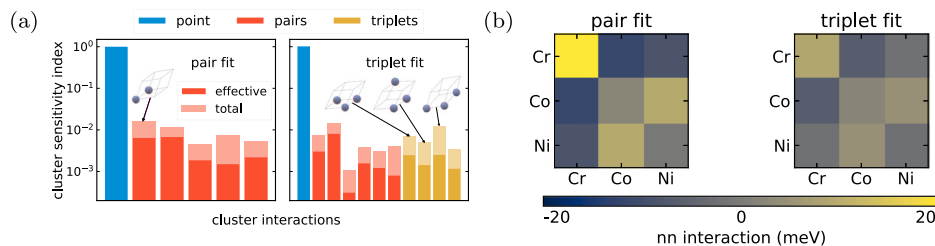


Fig. 4 | Estimated sensitivity indices and cluster interactions of cluster decompositions truncated at pairs and triplets. **a** Cluster sensitivity indices of two fitted CrCoNi CDs (one including only pairs, and another including pairs and triplets) sorted by cluster diameter. Effective (total) cluster sensitivity indices are shown with

solid (translucent) colors. **b** Nearest-neighbor cluster interactions as two-dimensional arrays. Each entry of the array represents the interaction energy between the corresponding nearest neighbor species.

ANOVA representations of $H(\sigma)$ ^{32,38}. In fact, re-written in such a form, a Fourier CE is nothing more than a fANOVA representation, in which by symmetry, interactions among equivalent clusters $S \in B$ are given by the same function \hat{H}_B . By extension, using a CD as an effective Hamiltonian to define a Boltzmann distribution is equivalent to log-density ANOVA estimation of a probabilistic graphical model^{39–41}.

Statistical interpretation of cluster interactions. Using the properties of ANOVA representations, we can obtain a much deeper understanding of the terms in a CD. ANOVA terms are constructed from hierarchical inclusion-exclusion of expectation values conditioned on the occupancy of clusters. For example, as already known from the original CE formalism¹, the constant term is equal to the mean energy in the non-interacting limit, $J_\emptyset = H_\emptyset = \langle H(\sigma) \rangle$. In the statistics literature, J_\emptyset is usually referred to as the *grand mean*⁴². The point terms, $\hat{H}_p(\sigma_i)$ are the difference between the mean conditioned on the i -th site and the grand mean, $\hat{H}_p(\sigma_i) = \langle H(\sigma) | \sigma_i \rangle - \langle H(\sigma) \rangle$. The point terms of an ANOVA representation are called *main effects*⁴². The main effects are the mean contribution that a specific species σ_i residing on the i -th site has on the total energy, and the average of *main effects* in the CD represents the portion of the energy that depends on composition only.

The remaining terms involving clusters S with more than one site are referred to in the statistics literature as *interactions*⁴², which further motivates our terminology. The cluster interaction $\hat{H}_B(\sigma_S)$ of a cluster S is the mean conditioned on the sites in cluster S minus the cluster interactions of all its sub-clusters $T \subset S$,

$$\hat{H}_B(\sigma_S) = \langle H(\sigma) | \sigma_S \rangle - \sum_{T \subset S} \hat{H}_C(\sigma_T) \quad (19)$$

Equation (19) clarifies the nature of a cluster interaction as the mean contribution (under the a-priori product distribution) to the total energy that originates solely from a single cluster $S \in B$ and none of its sub-clusters. Hence, the terms in the CD represent energetic interactions among species occupying the sites of a cluster that cannot be captured by any lower-order interactions. Moreover, we have an exact interpretation of the elements in the cluster interactions tensor shown in Fig. 3; they are the mean energy conditioned on the specific occupation σ_S of that cluster, with all interactions of sub-clusters $T \subset S$ conditioned on the corresponding occupation σ_T removed.

In our exposition, we started with a representation of a CD using a Fourier CE. However, since the CD is basis-independent, we can discard the concept of a basis altogether. In fact, in fANOVA and related literature, a function is decomposed into its ANOVA representation by directly appealing to Equation (19)^{32,38}. This approach has been used in concurrent work⁴³ that provides an axiomatic exposition of the CE and the CD. This exposition is equivalent to the formalism of tensor-product fANOVA decompositions⁴⁰ that we have presented here.

Variance decomposition & cluster sensitivity indices. As the name *analysis of variance* suggests, a CD also comprises a decomposition of the variance of the configuration energy under the a-priori non-interacting product measure $P(\sigma) = \prod_i p_i(\sigma_i)$ ³⁸. The total variance of a Fourier CE can be conveniently computed from the expansion coefficients as follows,

$$\text{Var}[H(\sigma)] = \langle H^2 \rangle - \langle H \rangle^2 \quad (20)$$

$$= \sum_{\alpha} J_{\alpha}^2 - J_{\mathbf{0}}^2 = \sum_{\alpha \neq \mathbf{0}} J_{\alpha}^2 \quad (21)$$

where $\mathbf{0}$ is the multi-index of all zeros, and where we have used the orthonormality of Fourier cluster functions.

By grouping terms by multi-indices with the same support and subsequently by symmetrically equivalent multi-indices, Equation (21) can be re-written as,

$$\text{Var}[H(\sigma)] = \sum_{B \neq \emptyset} \sum_{S \in B} \sum_{\beta \in L(B)} \hat{m}_{\beta} J_{\beta}^2 \quad (22)$$

$$= N \sum_{B \neq \emptyset} m_B \|\hat{H}_B\|_2^2 \quad (23)$$

Where we identify the innermost sum in Equation (22) as the effective cluster weights.

From on Equation (21), we see that the variance of a single cluster interaction is equal to its effective cluster weight, $\text{Var}[\hat{H}_B] = \|\hat{H}_B\|_2^2$. Equation (22) can be explicitly expressed as a variance decomposition in terms of individual cluster variances as follows,

$$\text{Var}[H(\sigma)] = N \sum_{B \neq \emptyset} m_B \text{Var}[\hat{H}_B(\sigma_S)] \quad (24)$$

By computing the variance of Equation (19), we can show that the effective cluster weights correspond to the conditional variance associated with a cluster with all lower-order variances subtracted,

$$\text{Var}[\hat{H}_B(\sigma_S)] = \text{Var}[H(\sigma) | \sigma_S] - \sum_{T \subset S} \text{Var}[H(\sigma) | \sigma_T] \quad (25)$$

In other words, an effective cluster weight is the variance in the non-interacting limit that can be attributed to a single cluster only and to none of its sub-clusters.

The variance decomposition resulting from the CD provides motivation and interpretation for the choice of regularization used when fitting. For example, Ridge regularization can be interpreted as setting an upper cutoff to the total variance. The use of Tikhonov regularization can be used to more finely set variance cutoffs for specific correlation functions. Group-wise regularization² can be directly motivated as a judicious form to regularize cluster interactions H_B directly and tune the resulting relative variance

contributions. Finally, estimation algorithms with hierarchical inclusion/exclusion of clusters^{3,4,44–46}, can be motivated by appealing to statistical concepts of *hierarchically well-formulated models*⁴⁷ that satisfy *marginality constraints*⁴⁸, or that abide by *heredity principles* under so-called weak or strong hierarchy constraints^{49,50}.

The variance decomposition inherent in the CD also allows one to formally rank the importance of each cluster interaction by following the prescription of Sobol's sensitivity indices³², which are used extensively in variance-based sensitivity analysis. Accordingly, we define the effective cluster sensitivity index $\hat{\tau}_B$ as the fraction of the total variance of H carried by the interactions of a cluster $S \in B$,

$$\hat{\tau}_B = \frac{\text{Var}[\hat{H}_B(\sigma_S)]}{\text{Var}[H(\sigma)]} \quad (26)$$

We define the cluster sensitivity index τ_B as the normalized fraction of the total variance of $H(\sigma)$ contributed by the cluster interaction \hat{H}_B per normalizing unit, $\tau_B = m_B \hat{\tau}_B$. Cluster sensitivity indices provide a mathematically formal and basis-independent route for evaluating trends in the strength of interactions, such as their decay with respect to the number of sites and increasing spatial diameter. Cluster sensitivity indices can be directly computed from a CE by using Equation (18). Furthermore, global sensitivity indices of a cluster S , can be computed by summing effective cluster indices over all sub-clusters or over all clusters with non-zero intersections with S ^{32,51}.

Example—CrCoNi medium entropy alloy

As a brief example demonstrating a practical use case of the CD, we fit two expansions of a CrCoNi medium entropy alloy: (1) a less accurate expansion (in terms of cross-validation error) that includes pairs terms only (pair fit), and (2) a more accurate expansion including pairs and triplets (triplet fit). Details of the DFT calculations used to generate training data and the regression procedure to obtain the final fits are given in Section 4. Our approach follows a recent study of the CrCoNi alloy³¹ which uses the CE method and Wang–Landau sampling to explore the role of pair and triplet interactions in tuning short-range order (SRO) and the resulting experimentally observed peak in the heat capacity around 940 K⁵². We will use the cluster decomposition formalism to carry out the corresponding calculations and analysis in order to illustrate how a richer and deeper insight is obtained directly from comparing the thermodynamic results obtained from each CD without the need for ad-hoc models to analyze interactions between species. We aim only to demonstrate the possibility of more in-depth analysis via the CD as an illustration, but we do not attempt to make any novel scientific claims about this particular alloy.

The cluster sensitivity indices for the resulting cluster decompositions are shown in Fig. 4a in order of increasing number of sites and cluster diameter. The selected pair clusters between the two fits are not exactly the same because they result from statistical fits using different feature matrices and a regression algorithm with inherent feature selection. Accordingly, obtaining different selected pair cluster interactions and values for the two different fits is expected. We also note that our rigorous claims of uniqueness, invariance, and interpretability deal with a *given* function and not the process of statistical approximation used to obtain it—our work opens worthwhile ways to develop basis-independent estimation, but that is beyond the scope of this work.

Although the same starting pool of pair terms is used in both fits, five pair interactions are selected in the pair-only fit, and six pair interactions are selected in the fit that includes triplets. Both fits include the first, second, and fourth nearest neighbor pair interactions, as well as a pair interaction with a 7.48 Å diameter. The pair-only fit includes an additional pair interaction with a diameter of 6.6 Å, while the triplet fit includes the third nearest neighbor (though with a noticeably lower sensitivity index) and an additional pair interaction with an 8.98 Å diameter. These differences in the selection of pairs for the two fits are attributed to the statistical-based feature selection of the regularized regression we have used.

The cluster sensitivity indices for the nearest neighbor pair interaction and three of the four included triplet interactions in Fig. 4 are labeled with figures of their representative clusters S . We observe that the nearest neighbor pair is a sub-cluster of all three labeled triplet clusters. This is not unexpected given that we obtain the fits using hierarchical constraints, i.e., for any triplet cluster interaction included, the interactions of all its sub-clusters must also be included. However, there is a clear selection of triplet clusters that include the nearest neighbor pair over other selected pair interactions for which triplets that include these were also present and fully sampled (full rank) in the initial pool of clusters. Furthermore, we observe that the cluster sensitivity index for the nearest neighbor pair interaction (first red bar in both fits) is lower in the triplet fit than in the pair-only fit. This behavior can be rationalized as an effect of the possible renormalization of interactions⁵³ from the different levels of truncation used in the two fits.

By inspecting Fig. 4a, b, we see that the interactions shown in Fig. 3 are of different orders of magnitude. The cluster sensitivity index of the main effect is about 2 orders of magnitude larger than that of higher-degree interactions. In Fig. 4b, we see that nearest neighbor interactions are of the order of 10 meV, and therefore main effects are of eV magnitude. We can identify the most important cluster interactions, rank their importance, and compare different fits on rigorous grounds by using cluster sensitivity indices. The point terms account for the overwhelming majority of the variance, however their contributions, are only relevant when the composition is not fixed (i.e., in the grand canonical ensemble). In both fits, the nearest neighbor and second nearest neighbor interactions are the most important (largest sensitivity), with significant contributions coming from triplet interactions in the triplet fit.

Figure 4b shows the nearest neighbor pair interactions for both fits. The energy contributions arising from the interaction between specific species can be obtained by directly inspecting the values of the cluster interaction elements. For example, a Cr–Cr nearest neighbor interaction in the pair fit contributes ~ 20 meV to the total energy, and a Cr–Co contributes ~ -20 meV. Thus, we can expect SRO to favor Cr–Co nearest neighbor pairs over Cr–Cr; though this behavior will be adjusted by the triplet interactions that includes nearest neighbor pairs. Comparing the nearest neighbor pair interactions between the two fits, we see that the attraction/repulsion and of relative magnitude trends are the same. For example, in both fits Cr–Co nearest neighbor pairs have the strongest attractive interaction, and Cr–Ni have the second strongest attractive interaction. However, we observe that the overall magnitude of nearest neighbor pair interaction values in the triplet fit is lower than those of the pair-only fit, in accordance with their cluster sensitivity indices. These observations suggest that the inclusion of triplet interactions modifies the strength but not the trends among nearest neighbor pair interactions, and as a result, triplet interactions will indeed tune the resulting SRO and heat capacity without changing the overall trend.

Finally, the relative trends of the nearest-neighbor interactions obtained directly from the CD (Fig. 4b) agree with previous results that had to be obtained via an ad-hoc and less accurate nearest-neighbor pair model³¹. Using the cluster decomposition we are able to directly obtain the interaction energy from both fits and compare the overall importance of interactions arising from different clusters by using their cluster sensitivity indices. Hence the CD enables mathematically rigorous analysis and interpretation of the interactions among chemical species without the need to resort to auxiliary models—an essential component so far absent in the CE formalism.

SRO and the K-state phenomenon. Configuration entropy has been found to be the main driver of the K-state phenomenon characterized by a peak in the heat capacity of the CrCoNi alloy, specifically by way of an ordering transition that favors CrNi and CrCo nearest neighbor pairs. The ordering transition is observed to be driven predominantly by nearest neighbor pair interactions but substantially tuned by many-body interactions³¹. This behavior is foreshadowed by our previous analysis of nearest-nearest neighbor cluster interactions and cluster sensitivity indices. Nonetheless, to further illustrate the use of the CD, we

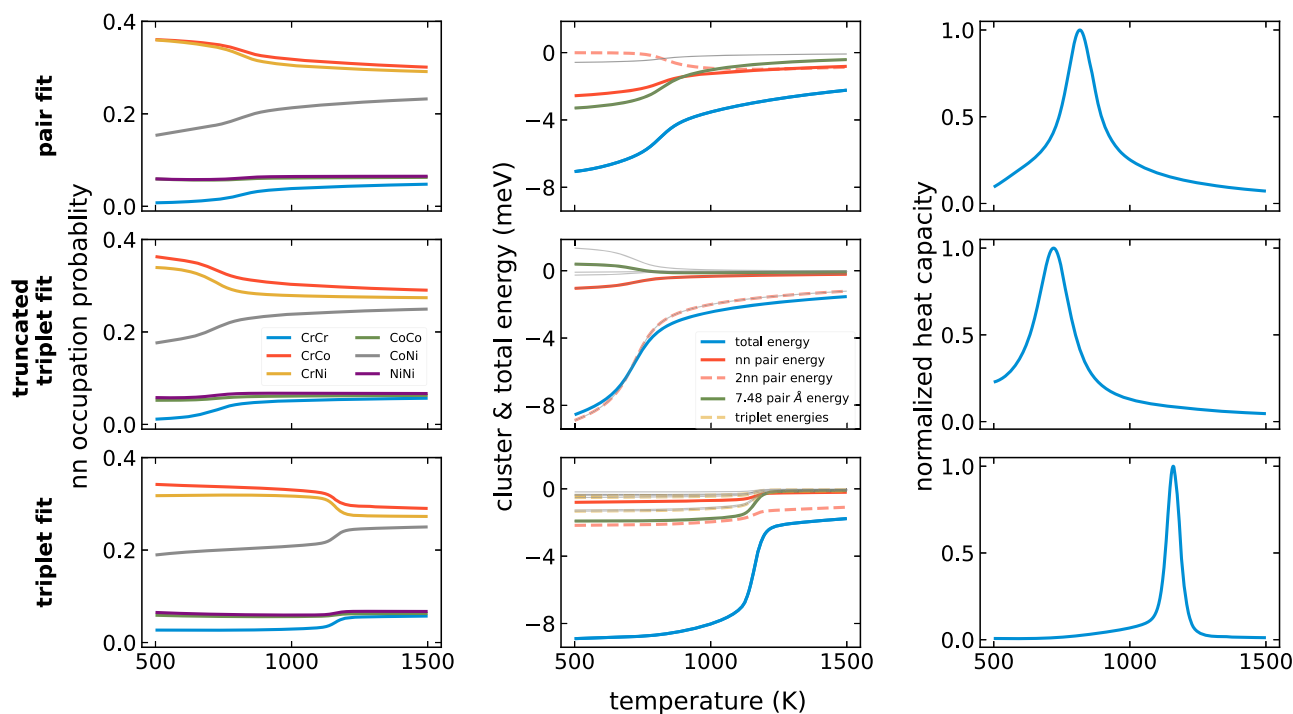


Fig. 5 | Nearest-neighbor pair probabilities (SRO), internal energy, nearest neighbor pair energy, and normalized heat capacity. The nearest neighbor pair energy is plotted with a solid blue curve, and the internal energy with a solid red

curve. Pair fit results (top), triplet fit results (bottom, solid), truncated triplet fit to pairs (bottom, translucent/dot-dash).

corroborate the origin of the peak in the heat capacity, explicitly compute SRO behavior, and further substantiate the predicted roles of higher-order interactions using a Wang–Landau estimation of the density of states.

We computed nearest-neighbor pair SRO, internal energy, and the resulting heat capacity via equiatomic canonical Wang–Landau sampling. Figure 5 shows the nearest-neighbor pair SRO, the internal energy, and the heat capacity computed for the pair fit, the triplet fit, as well as a truncated CD that includes only the pair interactions from the triplet fit (the six pair interactions in the rightmost plot in Fig. 4a). The middle column in Fig. 5, also shows the contribution to the internal energy coming from nearest-neighbor interactions, second nearest neighbor interactions, the 7.48 Å interaction, and triplet interactions. All other interactions are plotted in gray curves. Since the sum of all interactions in each simulation gives the total internal energy, the CD allows us to obtain a breakdown of the finite temperature interactions arising from different clusters.

For all cases, we observe that the start of the ordering transition and the inflection of internal energy are congruent with the location of the peak in the heat capacity; suggesting that, as previously reported⁵¹, the ordering transition is the main cause for the peak in heat capacity associated with the K-state phenomenon. Specifically, the peak in the heat capacity predicted with the pair fit occurs at 620 K and the peak predicted with the triplet fit happens at 1300 K. We observe favorable ordering tendencies between CrCr and CrNi nearest neighbor pairs and a decrease in the probability of observing CoNi pairs as foreshadowed by directly inspecting their cluster interactions shown in Fig. 4b.

We obtain a much richer picture by observing the contributions to the internal energy arising from the included cluster interactions. Above the transition temperature, the magnitude of contributions of each cluster follows the order set by the cluster sensitivity indices shown in Fig. 4a, which we rationalize by interpreting the expectation operation used in their ANOVA statistical interpretation—such as in Equation (19)—as the ensemble average in the random (high temperature) limit. However, we observe that the order of contributions to the internal energy changes at the phase transition, which we attribute to the covariance that emerges between cluster

interactions at finite temperatures. Notably, the contribution from 7.48 Å pair interaction in the pair and triplet fit begins to overcome all other contributions as temperature decreases. In contrast, the contribution of the second nearest neighbor interaction decays to zero for the pair fit, but remains the most important in the triplet fit. These observations point to the importance of considering both cluster sensitivity indices and effects at finite temperatures when studying the role of different interactions in determining thermodynamic properties.

The agreement between the nearest-neighbor pair energies and SRO of the two CDs that include only pair interactions shown in Fig. 5, suggests that the overall SRO and total internal energy trends are set predominantly by the first and second nearest-neighbor pair interactions (those with the highest cluster sensitivity from Fig. 4b). Furthermore, the qualitatively similar trend of the internal energy and nearest-neighbor pair energy in all CDs further corroborates this conclusion.

However, the role of the second nearest neighbor pair below the transition temperature is markedly different when triplets are included in the fit. As we noted, the contributions of the second nearest neighbor pair decay to zero for the pair-only fit, but remain the most important for the triplet fit as well as the truncated triplet fit. This observation may be indicative of a renormalization⁵³ of excluded interactions in the pair fit, since the second nearest neighbor contribution does not decay when triplets are removed after fitting in the truncated triplet fit. In contrast, the internal energy contributions of the 7.48 Å pair do not decay and are qualitatively similar in both pair and triplet fit. This is indicative that this cluster interaction is not affected by renormalization, which we can further motivate by positing that all higher-order interactions that include the long-range 7.48 Å that are excluded from the fit have negligible contributions.

Considering the similarity of the resulting heat capacity from the CDs with pairs only suggests that the transition is indeed driven predominantly by nearest-neighbour pair interactions. It is interesting to note that the transition temperature for the CD obtained from truncating the triplet fit is lower than that of the pair fit. This behaviour can also be attributed to the removal of higher-order contributions from triplets that were not renormalized into pair interactions during fitting.

By comparing the results from the CD that includes triplet interactions with the pair-only CDs, we can observe that triplet terms tune the SRO values, reduce the fraction of internal energy attributed to pair terms, and raise the transition temperature. Additionally, the triplet interactions qualitatively change the nature of the transition, as observed by the more abrupt changes in SRO values and internal energy and the narrower heat capacity peak compared to the corresponding values obtained from CDs with pairs only. To delve deeper, one could inspect the triplet interaction values from their array representations (as shown in Fig. 3) to better understand the specific roles in tuning the ordering transition.

The computed internal energy and heat capacity, which are total system values, could have been suitably obtained from the CE formalism only. However, the formal breakdown of contributions from the different cluster interactions that result from the CD formalism permits a richer analysis of the role of these interactions. For this particular example, we have leveraged the CD formalism, in particular, the breakdown of energy into statistically independent cluster contributions and their cluster sensitivity indices, to faithfully explain the primary origin of the K-state phenomenon in the CrCoNi medium entropy alloy. We also characterized the primary role of nearest-neighbour interactions in setting ordering trends and the transition temperature and clarified the role of longer-range pairs and triplet interactions in tuning both of these quantities. All of the presented conclusions based on our results agree with those reported previously³¹; however, by using the CD, we have shown how results leading to these conclusions can be more effectively calculated and further substantiated with the mathematically formal analysis resulting from the CD.

Discussion

In this work, we introduce the CD as a versatile framework for representing the configurational energy of multicomponent materials. Unlike CEs that rely on an arbitrary choice of basis, the CD offers a basis-independent representation in terms of cluster interactions. In contrast to correlation functions, cluster interactions are physically intuitive, numerically more efficient, and mathematically meaningful. To achieve this, we revisited the mathematical formalism of the CE method and provided a modern exposition that emphasizes its tensor-product structure. This enables us to obtain a more intuitive, compact, and numerically efficient representation of cluster interactions.

We further established a direct connection between the CD and functional analysis of variance (ANOVA). Specifically, we showed that the cluster interaction terms in the CD can be interpreted as conditional expectation values in the non-interacting limit, with lower-order effects removed. We also discussed the nature of the CD as a variance decomposition and introduced cluster sensitivity indices as a formal measure to rank and compare the importance of the contributions of different cluster interactions. This provides a mathematically rigorous interpretation of the CD and the nature of the cluster interactions.

As an illustrative example, we used the CD formalism to explain the role of SRO in the K-state phenomenon of the CrCoNi medium entropy alloy. We demonstrated how the CD can be used to analyze the contributions arising from the interactions among species in different clusters and explain their effects in setting and tuning ordering tendencies and transition temperatures.

We believe that substantially more insight, use cases, and parameter estimation methods beyond what we have presented can be developed using the CD and its formal statistical properties. The statistical literature is ripe with analysis techniques and methodology—such as aforesaid log-density ANOVA models^{40,41} and sensitivity analysis^{32,51}—that can be directly leveraged in applications using parameterized lattice models. Methods that can be used for direct estimation of cluster interactions and cluster indices in fully basis-independent manners already exist in the statistics literature^{32,39,41,51}. Moreover, the formalism of the CD is not limited to scalar functions of discrete degrees of freedom as presented here. In fact, a CD can

be obtained for any representation of scalar, vector, or tensor-valued function over a tensor-product space by following the same approach we have presented. Related expansions and generalizations such as spin CEs, expansions of elastic properties, and the atomic CE^{17–19,21,22} can be recast as a CD and thus open the door to continued and significant developments based on rigorously established mathematical and statistical grounds.

Methods

Density functional theory calculations

The pair and triplet expansions were fit using a set of 500 training structures with up to 12 atoms per super-cell. The energy of the training structures was computed with density functional theory (DFT) using the Vienna ab initio simulation package (VASP) with the projector-augmented wave method^{54,55}. The DFT calculations were performed using a plane-wave basis set with an energy cutoff of 520 eV, and a reciprocal space discretization of 200 k -points per Å. Electronic exchange-correlation effects are described using the Perdew-Burke-Ernzerhof (PBE) generalized gradient approximation exchange-correlation functional⁵⁶. All calculations were converged to 10^{-5} eV in total energy for electronic loops and 0.01 eV/Å for forces. All DFT calculations are performed following the Materials Project⁵⁷ MetalRelaxSet settings defined in the pymatgen Python package⁵⁸.

Mixed-integer quadratic programming hierarchically constrained structured sparsity fits

The two fits were done using a mixed-integer quadratic problem formulation of a grouped ℓ_0 pseudo-norm and ℓ_2 norm regularization to obtain hierarchically constrained structured sparsity between cluster interactions. The regression method used is a simple generalization of a recently proposed method used to obtain fits with hierarchical constraints between correlation functions^{3,46}. Our choice of regularization results in efficient cluster selection and ensures that strong hierarchy constraints are respected in the resulting CDs^{47,49,50}. The regression optimization problem used is given as follows,

$$\begin{aligned} \min_{\mathbf{J}} \quad & \mathbf{J}^T (\mathbf{\Pi}^T \mathbf{\Pi} + \lambda_1 \mathbf{I}) \mathbf{J} - 2 \mathbf{E}^T \mathbf{\Pi} \mathbf{J} + \lambda_0 \sum_B z_B \\ \text{subject to} \quad & -M z_B \mathbf{1} \leq \mathbf{J}_B \leq M z_B \mathbf{1} \\ & z_B \in \{0, 1\} \\ & z_B \leq z_D \quad \forall D \text{ s.t. } D \supset B \end{aligned} \quad (27)$$

where \mathbf{J} is a vector of all expansion coefficients, \mathbf{J}_B are the coefficients corresponding to a single orbit B ; $\mathbf{\Pi}$ is a matrix of the correlation vectors of all training structures; \mathbf{E} is a vector of DFT computed energies; \mathbf{I} is the identity matrix; $\mathbf{1}$ are vectors of all ones, $\lambda_0, \lambda_1 \in \mathbb{R}_+$ are hyper-parameters; $M \in \mathbb{R}_+$ is a fixed parameter; and z_B are slack variables that determine whether a group of coefficients \mathbf{J}_B associated with a single cluster interaction is zero or non-zero, i.e., active ($z_B \neq 0$) or inactive ($z_B = 0$). The notation $D \supset B$ means that any cluster $T \in D$ is a sub-cluster $T \subset S$ of some cluster $S \in B$. We refer the reader to the aforementioned references^{3,46} for further details on this regression methodology.

The final expansion fits are converged to a fivefold root mean squared cross-validation of 12.8 meV/atom (triplet fit) and 14.9 meV/atom (pair fit). The final fits include non-zero pair and triplet interactions with diameters up to 9 Å and 4.3 Å, and non-zero pair interactions with diameters up to 7.5 Å, respectively, both based on a 2.49 Å primitive lattice constant.

Wang–Landau density of states calculations

The canonical equiatomic Wang–Landau sampling used to directly approximate the density of states was done using supercells with 216 sites ($6 \times 6 \times 6$ supercells of a primitive face-centered cubic structure.). Wang–Landau simulations are converged to a fill factor value $< 10^{-6}$ ⁵² as shown in Fig. 6.

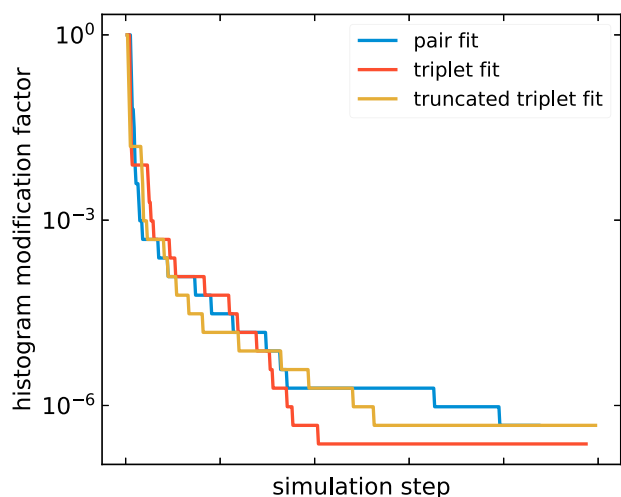


Fig. 6 | Convergence of Wang–Landau density of states calculations. Convergence curves of histogram modification (fill) factors with respect to the number of simulation steps taken.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

An implementation of the cluster decomposition, Fourier CEs, the Wang–Landau sampling algorithm, and all other code used in this work is available in the *Statistical Mechanics on Lattices* (smo1) Python package⁵⁹.

Received: 29 August 2023; Accepted: 28 June 2024;

Published online: 19 July 2024

References

- Sanchez, J. M., Ducastelle, F. & Gratias, D. Generalized cluster description of multicomponent systems. *Phys. A: Stat. Mech. Appl.* **128**, 334 (1984).
- Barroso-Luque, L. et al. Cluster expansions of multicomponent ionic materials: formalism and methodology. *Phys. Rev. B* **106**, 144202 (2022).
- Xie, J.-Z., Zhou, X.-Y. & Jiang, H. Perspective on optimal strategies of building cluster expansion models for configurationally disordered materials. *J. Chem. Phys.* **157**, 200901 (2022).
- van de Walle, A. & Ceder, G. Automating first-principles phase diagram calculations. *JPE* **23**, 348 (2002).
- Van der Ven, A., Thomas, J., Puchala, B. & Natarajan, A. First-principles statistical mechanics of multicomponent crystals. *Annu. Rev. Mater. Res.* **48**, 27 (2018).
- Hart, G. L. W., Mueller, T., Toher, C. & Curtarolo, S. Machine learning for alloys. *Nat. Rev. Mater.* **6**, 730 (2021).
- Sutton, C. and Levchenko, S. V. First-principles atomistic thermodynamics and configurational entropy. *Front. Chem.* **8**, 757 (2020).
- Nataraj, C., Borda, E. J. L., van de Walle, A. & Samanta, A. A systematic analysis of phase stability in refractory high entropy alloys utilizing linear and non-linear cluster expansion models. *Acta Mater.* **220**, 117269 (2021).
- Xu, X. & Jiang, H. Cluster expansion based configurational averaging approach to bandgaps of semiconductor alloys. *J. Chem. Phys.* **150**, 034102 (2019).
- Han, G., Yeu, I. W., Ye, K. H., Hwang, C. S. & Choi, J.-H. Atomistic prediction on the composition- and configuration-dependent bandgap of Ga(As,Sb) using cluster expansion and ab initio thermodynamics. *Mater. Sci. Eng. B* **280**, 115713 (2022).
- Richards, W. D., Wang, Y., Miara, L. J., Kim, J. C. & Ceder, G. Design of $\text{Li}_{1+2x}\text{Zn}_{1-x}\text{PS}_4$, a new lithium ion conductor. *Energy Environ. Sci.* **9**, 3272 (2016).
- Deng, Z. et al. Phase behavior in rhombohedral NaSiCON electrolytes and electrodes. *Chem. Mater.* **32**, 7908 (2020).
- Van der Ven, A., Deng, Z., Banerjee, S. & Ong, S. P. Rechargeable alkali-ion battery materials: theory and computation. *Chem. Rev.* **120**, 6977 (2020).
- Chen, B. W. J., Xu, L. & Mavrikakis, M. Computational methods in heterogeneous catalysis. *Chem. Rev.* **121**, 1007 (2021).
- Wu, D., Wang, L. & Zhang, P. Solving statistical mechanics using variational autoregressive networks. *Phys. Rev. Lett.* **122**, 080602 (2019).
- Damewood, J., Schwalbe-Koda, D. & Gómez-Bombarelli, R. Sampling lattices in semi-grand canonical ensemble with autoregressive machine learning. *npj Comput. Mater.* **8**, 1 (2022).
- Drautz, R. & Fähnle, M. Spin-cluster expansion: parametrization of the general adiabatic magnetic energy surface with ab initio accuracy. *Phys. Rev. B* **69**, 104404 (2004).
- Singer, R., Dietermann, F. & Fähnle, M. Spin Interactions in bcc and fcc Fe beyond the Heisenberg model. *Phys. Rev. Lett.* **107**, 017204 (2011).
- Thomas, J. C. & Van der Ven, A. The exploration of nonlinear elasticity and its efficient parameterization for crystalline materials. *J. Mech. Phys. Solids* **107**, 76 (2017).
- Thomas, J. C., Bechtel, J. S. & Van der Ven, A. Hamiltonians and order parameters for crystals of orientable molecules. *Phys. Rev. B* **98**, 094105 (2018).
- van de Walle, A. A complete representation of structure–property relationships in crystals. *Nat. Mater.* **7**, 455 (2008).
- Drautz, R. Atomic cluster expansion of scalar, vectorial, and tensorial properties including magnetism and charge transfer. *Phys. Rev. B* **102**, 024104 (2020).
- Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **99**, 014104 (2019).
- Ceccherini-Silberstein, T., Scarabotti, F. & Tolli, F. *Discrete harmonic analysis: representations, number theory, expanders, and the fourier transform*, Cambridge Studies in Advanced Mathematics <https://doi.org/10.1017/9781316856383> (Cambridge University Press, Cambridge, 2018).
- Wolverton, C. & Zunger, A. Ising-like description of structurally relaxed ordered and disordered alloys. *Phys. Rev. Lett.* **75**, 3162 (1995).
- BRUSH, S. G. History of the Lenz-Ising model. *Rev. Mod. Phys.* **39**, 883 (1967).
- Stampfl, C., Kreuzer, H. J., Payne, S. H., Pfnür, H. & Scheffler, M. First-principles theory of surface thermodynamics and kinetics. *Phys. Rev. Lett.* **83**, 2993 (1999).
- Barroso-Luque, L., Yang, J. H. & Ceder, G. Sparse expansions of multicomponent oxide configuration energy using coherency and redundancy. *Phys. Rev. B* **104**, 224203 (2021).
- Kim, N., Blankenau, B. J., Su, T., Perry, N. H. & Ertekin, E. Multisublattice cluster expansion study of short-range ordering in iron-substituted strontium titanate. *Comput. Mater. Sci.* **202**, 110969 (2022).
- Sanchez, J. M. Cluster expansion and the configurational theory of alloys. *Phys. Rev. B* **81**, 224202 (2010).
- Pei, Z., Li, R., Gao, M. C. & Stocks, G. M. Statistics of the NiCoCr medium-entropy alloy: novel aspects of an old puzzle. *Npj Comput. Mater.* **6**, 1 (2020).
- Sobol' I. M. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simulat.* **55**, 271 (2001).
- Hoefding, W. A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.* **19**, 293 (1948).

34. Efron, B. & Stein, C. The Jackknife estimate of variance. *Ann. Stat.* **9**, 586 (1981).
35. Sanchez, J. M. Cluster expansions and the configurational energy of alloys. *Phys. Rev. B* **48**, 14013 (1993).
36. Lang, S. Integration and measures on locally compact spaces, In *Real and Functional Analysis*, Graduate Texts in Mathematics, Eds. Lang, S. https://doi.org/10.1007/978-1-4612-0897-6_9 (Springer, New York, NY, 1993).
37. Janson, S. and Janson, P. O. M. S. *Gaussian Hilbert Spaces* (Cambridge University Press, 1997).
38. Hooker, G. Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *J. Comput. Graph. Stat.* **16**, 709 (2007).
39. Jeon, Y. & Lin, Y. An effective method for high-dimensional log-density anova estimation, with application to nonparametric graphical model building. *Stat. Sin.* **16**, 353 (2006).
40. Jeon, Y. A characterization of the log-density smoothing spline ANOVA model. *Commun. Stat. - Theory Methods* **41**, 2081 (2012).
41. Gu, C. Regression with responses from exponential families. In *Smoothing Spline ANOVA Models*, Springer Series in Statistics, edited by Gu, C. https://doi.org/10.1007/978-1-4614-5369-7_5 (Springer, New York, NY) pp. 175–214 (2013).
42. Gelman, A. Analysis of variance—why it is more important than ever. *Ann. Stat.* **33**, 1 (2005).
43. Lammert, P. E. and Crespi, V. H. Cluster expansion methods from physical concepts, *arxiv* <https://doi.org/10.48550/arXiv.2210.10937> (2022).
44. Zarkevich, N. A. & Johnson, D. D. Reliable first-principles alloy thermodynamics via truncated cluster expansions. *Phys. Rev. Lett.* **92**, 255702 (2004).
45. Leong, Z. & Tan, T. L. Robust cluster expansion of multicomponent systems using structured sparsity. *Phys. Rev. B* **100**, 134108 (2019).
46. Zhong, P., Chen, T., Barroso-Luque, L., Xie, F. & Ceder, G. An ℓ_2 -norm regularized regression model for construction of robust cluster expansions in multicomponent systems. *Phys. Rev. B* **106**, 024203 (2022).
47. Peixoto, J. L. A Property of Well-Formulated Polynomial Regression Models. *Am. Stat.* **44**, 26 (1990).
48. McCullagh, P. & Nelder, J. A. *Generalized Linear Models*, 2nd ed. <https://doi.org/10.1201/9780203753736> (Routledge, New York, 2019).
49. Hamada, M. & Wu, C. F. J. Analysis of designed experiments with complex aliasing. *J. Qual. Technol.* **24**, 130 (1992).
50. Chipman, H. Bayesian variable selection with related predictors. *Can. J. Stat. Rev. Can. Stat.* **24**, 17 (1996).
51. Iooss, B. & Lemaître, P. A review on global sensitivity analysis methods. In *Uncertainty Management in Simulation-Optimization of Complex Systems: Algorithms and Applications*. Operations research/computer science interfaces series, edited by Dellino, G. and Meloni, C. https://doi.org/10.1007/978-1-4899-7547-8_5 (Springer US, Boston, MA) pp. 101–122 (2015).
52. Wang, F. & Landau, D. P. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* **86**, 2050 (2001).
53. Sanchez, J. M. Renormalized interactions in truncated cluster expansions. *Phys. Rev. B* **99**, 134206 (2019).
54. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15 (1996).
55. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758 (1999).
56. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
57. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
58. Ong, S. P. et al. Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314 (2013).
59. Barroso-Luque, L. et al. Smol: a python package for cluster expansions and beyond. *J. Open Source Softw.* **7**, 4504 (2022).

Acknowledgements

This work was primarily funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under Contract No. DE-AC02-05-CH11231 (Materials Project program KC23MP). This research also used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05-CH11231 using NERSC award BES-ERCAP0020531.

Author contributions

L.B.L. conceived the project idea, wrote the proofs and derivations, implemented the code, and generated the data. G.C. supervised the project. Both authors analyzed the results and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-024-01338-y>.

Correspondence and requests for materials should be addressed to Luis Barroso-Luque or Gerbrand Ceder.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024