



This is the accepted manuscript made available via CHORUS. The article has been published as:

Short-range order parameter expansions for simple configurational energetics

Flynn Walsh

Phys. Rev. Materials **8**, L100801 — Published 3 October 2024

DOI: [10.1103/PhysRevMaterials.8.L100801](https://doi.org/10.1103/PhysRevMaterials.8.L100801)

Short-range order parameter expansions for simple configurational energetics

Flynn Walsh

Lawrence Livermore National Laboratory, Livermore, California 94550, USA

Local chemical environments of multicomponent crystals are parametrized according to basis vectors of their irreducible invariant subspaces. These short-range order parameters enable Taylor expansions of configurational properties that easily incorporate additional couplings. As one example, expansions including long-range composition can describe the energies of many alloys more accurately than conventional cluster expansions while using far fewer neighbors and basis functions.

Configurational degrees of freedom in crystals are typically studied using Ising-like models of interactions among abstracted sites. Most commonly, multi-body cluster expansions (CEs) project configurational energies onto polynomials in integers representing crystal site occupancies [1, 2]. Despite their continuing popularity [3–7], standard CEs poorly describe properties that depend nonlinearly on composition [8], including the ordering energies of many materials. Composition-dependent versions of the formalism resolve this issue [8–11], but are not widely employed.

This study explores another approach to configurational modeling, with parallels to both Landau free-energy expansions [12] and the Atomic Cluster Expansion (ACE) [13]. The basic idea is to Taylor expand local properties in terms of symmetry-adapted order parameters, at first for short-range chemical environments but ultimately for any degree of freedom. Local constructions demonstrate modest improvement over conventional CEs for some systems, which is attributed to a less inefficient representation of composition. Explicitly including composition as an expansion parameter then leads to very simple but accurate descriptions of energy.

Fundamentally, configurational models express properties, particularly energy (U), as functions of the chemical occupancies of abstracted sites in a crystal basis. Contributions from local structural relaxations can be implicitly accounted as functions of the chemical environment. Practical modeling requires some assumption of locality, with the implication that quantities can be decomposed onto periodic subunits of the crystal, labeled i such that $U = \sum_i U_i$. Though not obligatory, projecting U onto basis sites as atomic energies is both intuitive and effective; atomic decomposition underlies interatomic potentials and can also be used to represent CEs [14].

Assumed locality implies that the energy associated with site i is a function of the chemical occupancy of the N sites within some finite distance. For a system of M chemical species, the occupancy of site j consists of M fractional concentrations, $x_j^A, x_j^B, \dots, x_j^M$, such that the local chemical environment of site i is described by the $M(N+1)$ -dimensional vector

$$\mathbf{x}_i = [x_{i0}^A, x_{i1}^A, \dots, x_{iN}^A, \dots, x_{i0}^M, x_{i1}^M, \dots, x_{iN}^M], \quad (1)$$

where site 0 describes i and sites 1 to N indicate its neighbors. Components are typically 0 or 1, with $\sum_A x_j^A = 1$, but mixed occupancies are theoretically admissible and

TABLE I. SROPs describing the NN chemical environment of an fcc site, as depicted in Fig. 1(a). Each chemical subspace is divided into IISs, for which symmetry operations form irreducible representations of the local point group (O_h), which are noted by the symbols Γ . IISs are indexed by α , with basis vector κ corresponding to SROP $\chi_{\kappa}^{A\alpha}$ for each species A .

Γ	α	κ	χ_{κ}^{α}
A_1^+	0	1	x_0
A_1^+	1	1	$(x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{11} + x_{12})/\sqrt{12}$
E^+	2	1	$(-x_1 + x_2 + x_3 - x_4 - x_9 + x_{10} + x_{11} - x_{12})/\sqrt{8}$
		2	$(x_1 + x_2 + x_3 + x_4 - 2x_5 - 2x_6 - 2x_7 - 2x_8 + x_9 + x_{10} + x_{11} + x_{12})/\sqrt{24}$
T_2^+	3	1	$(x_5 - x_6 - x_7 + x_8)/2$
		2	$(x_1 - x_4 - x_9 + x_{12})/2$
		3	$(x_2 - x_3 - x_{10} + x_{11})/2$
T_1^-	4	1	$(-x_2 + x_3 - x_5 + x_6 - x_7 + x_8 - x_{10} + x_{11})/\sqrt{8}$
		2	$(-x_1 + x_4 - x_5 - x_6 + x_7 + x_8 - x_9 + x_{12})/\sqrt{8}$
		3	$(-x_1 - x_2 - x_3 - x_4 + x_9 + x_{10} + x_{11} + x_{12})/\sqrt{8}$
T_2^-	5	1	$(x_1 - x_2 - x_3 + x_4 - x_9 + x_{10} + x_{11} - x_{12})/\sqrt{8}$
		2	$(x_2 - x_3 - x_5 + x_6 - x_7 + x_8 + x_{10} - x_{11})/\sqrt{8}$
		3	$(-x_1 + x_4 + x_5 + x_6 - x_7 - x_8 - x_9 + x_{12})/\sqrt{8}$

could prove useful in certain contexts. Local energies can be expressed as $U_i = U_i(\mathbf{x}_i)$, though $\mathbf{x}_i$ is not directly useful as its components are permuted by crystal symmetry operations that leave U_i invariant. Formally, $U_i(\mathbf{x}_i) = U_i(\mathbf{S}(g)\mathbf{x}_i)$, where $\mathbf{S}(g)$ is a permutation matrix representing local symmetry operation g . As site centering ensures translational invariance, the matrices $\mathbf{S}(g)$ form an $M(N+1)$ -dimensional permutation representation of the point group of site i , denoted $\mathcal{G}_i$.

Physical $U_i(\mathbf{x}_i)$ therefore must comprise functions of $\mathbf{x}_i$ that are invariant under $\mathcal{G}_i$. For example, CE basis functions intuitively sum products of site occupancies in symmetrically equivalent clusters. Alternatively, if local environments are reparametrized in terms of a maximally invariant basis, the machinery of group theory can directly provide invariant polynomials in $\mathbf{x}_i$. While few vectors in local concentration space are individually invariant, certain subspaces are preserved by all symmetry operations. The basis vectors of these invariant subspaces (ISs) are reoriented by symmetry operations, but their collective spans remain constant.

For instance, as symmetry operations do not transmute elements, the chemical space of a given species forms an

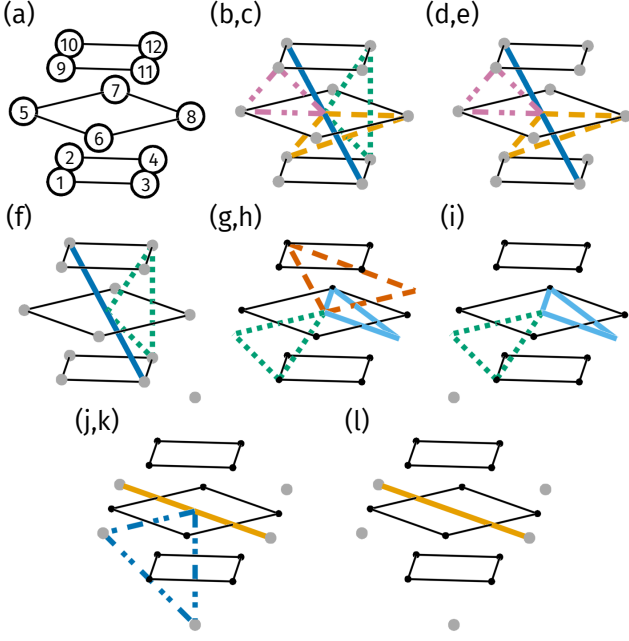


FIG. 1. (a) Site indices for the fcc NNs referenced in Table I. The same structure is used to depict the triplet clusters represented in the (b-f) 5 NN and (g-l) 6 second-NN third-order SROPE polynomials. Panels with two labels, such as (b,c), represent two different basis functions, which distinctly combine the same cluster components. Many polynomials also include self-interaction monomials, specifically second powers of individual site concentrations of the form $x_0^A \cdot x_j^B \cdot x_j^B$, which are represented by gray circles at sites j .

$N + 1$ -dimensional IS. Any $\mathcal{S}(g)$ simply permutes the components $x_{i0}^A, x_{i1}^A, \dots, x_{iN}^A$, which form the IS $\mathbf{x}_i^A$. All chemical subspaces share a common symmetric structure that can be decomposed in parallel, first into the neighbor shells of site i . These ISs are usually further separable into irreducible invariant subspaces (IISs), which can be obtained from the procedure described in App. A. For d^α -dimensional IIS α , basis vectors are noted as $\mathbf{e}_\kappa^\alpha$ for $\kappa = 1, \dots, d^\alpha$.

The projection of site concentrations onto IIS basis vectors defines order parameters,

$$\chi_{i\kappa}^{A\alpha} = \mathbf{x}_i^{A\top} \cdot \mathbf{e}_\kappa^\alpha, \quad (2)$$

that provide a maximally invariant basis for the space of $\mathbf{x}_i$. In contrast to the long-range order parameters that describe thermodynamic phases, $\chi_{i\kappa}^{A\alpha}$ is a short-range order parameter (SROP). As an illustrative example, Table I gives possible SROPs for the nearest-neighbor (NN) chemical environment of a face-centered cubic (fcc) lattice, which is depicted in Fig. 1(a).

SROPs are not generally as small as long-range order parameters near phase transitions, but their limited range nonetheless motivates a multivariable Taylor expansion of U_i that resembles the construction of Landau free energies [12]. To simplify the choice of reference

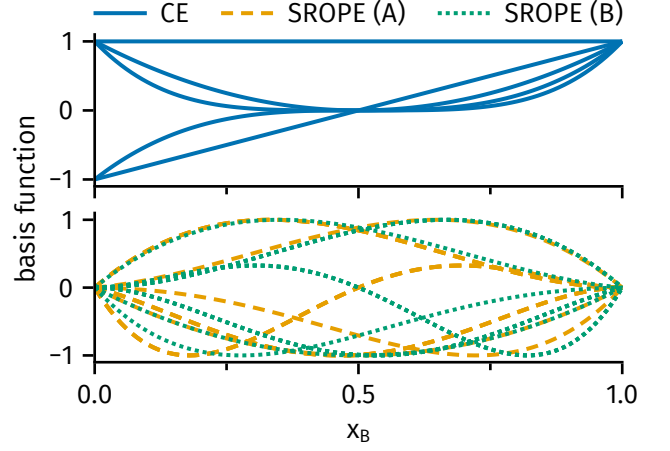


FIG. 2. 9 CE and 24 SROPE basis functions for random bcc A-B alloys with varying composition. Clusters are equivalent to those represented in the fourth-order nearest-neighbor SROPE. The latter contains more possible terms due to the addition of self-interactions, overcompleteness of fourth-order polynomials, and the definition of basis functions for each site element (A or B). In contrast to CEs, the SROPE formalism involves no intercept as polynomials go to zero at limits of the composition range. (For clarity, SROPE values are displayed relative to maximum magnitudes.)

state, properties of site i are expressed separately for each chemical species and weighted as

$$U_i = \sum_A x_{i0}^A U_i^A = \sum_A \chi_{i1}^{A0} U_i^A, \quad (3)$$

where $\chi_{i1}^{A0} = x_{i0}^A$ as in Table I. Each U_i^A is then expanded with respect to a pure A crystal, for which only $\chi_{i\kappa}^{A\alpha}$ are nonzero. Under the constraint of unity total concentration, only $M - 1$ chemical subspaces are independent, allowing expansions of U_i^A to drop $\chi_{i\kappa}^{A\alpha}$ in favor of SROPs that are zero in the pure reference state. Then, for specie indices $B, C, D \neq A$, with IISs respectively indexed by β, γ, δ ,

$$U_i \approx \sum_A \chi_{i1}^{A0} \left(U_i^A(\mathbf{0}) + \sum_B \sum_\beta \sum_\lambda \frac{\partial U_i^A}{\partial \chi_{i\lambda}^{B\beta}} \chi_{i\lambda}^{B\beta} + \frac{1}{2!} \sum_{BC} \sum_{\beta\gamma} \sum_{\lambda\mu} \frac{\partial^2 U_i^A}{\partial \chi_{i\lambda}^{B\beta} \partial \chi_{i\mu}^{C\gamma}} \chi_{i\lambda}^{B\beta} \chi_{i\mu}^{C\gamma} + \frac{1}{3!} \sum_{BCD} \sum_{\beta\gamma\delta} \sum_{\lambda\mu\nu} \frac{\partial^3 U_i^A}{\partial \chi_{i\lambda}^{B\beta} \partial \chi_{i\mu}^{C\gamma} \partial \chi_{i\nu}^{D\delta}} \chi_{i\lambda}^{B\beta} \chi_{i\mu}^{C\gamma} \chi_{i\nu}^{D\delta} + \dots \right), \quad (4)$$

where λ, μ, ν index SROPs and $U_i^A(\mathbf{0})$ is the reference energy of a pure A crystal. This expression will be referred to as a SROP expansion (SROPE).

The partial derivatives in Eq. (4) are unknown, but physics requires that all nonzero monomials form invariant polynomials. In the following, valid polynomials

are identified using the approach described in App. B, which resembles that of Ref. [13]. In brief, for some β and γ , the coupling coefficients $C_{\kappa,\lambda\mu}^{\alpha,\beta\gamma}$ combine the monomials $\chi_\lambda^\beta \chi_\mu^\gamma$ into a polynomial with the symmetry of χ_κ^α . Most simply, $C_{1,\lambda\mu}^{0,\beta\gamma}$ select invariant second-order polynomials, which require $\beta = \gamma$. To construct higher-order polynomials, multiple $C_{\kappa,\lambda\mu}^{\alpha,\beta\gamma}$ combine through intermediate couplings, which are denoted β' , γ' , etc.

Eq. (4) is thus rearranged into unique invariant polynomials, with partial derivatives absorbed into interaction parameters, J :

$$\begin{aligned}
 U_i \approx & \sum_A \chi_{i1}^{A0} \left(U_i^A(\mathbf{0}) + \sum_B \sum_\beta J_{0\beta}^{AB} \sum_\lambda C_{1,1\lambda}^{0,0\beta} \chi_{i\lambda}^{B\beta} \right. \\
 & + \sum_{BC} \sum_{\beta \geq \gamma} J_{0\beta\gamma}^{ABC} \sum_{\lambda\mu} C_{1,\lambda\mu}^{0,\beta\gamma} \chi_{i\lambda}^{B\beta} \chi_{i\mu}^{C\gamma} \\
 & + \sum_{BCD} \sum_{\beta'} \sum_{\beta \geq \gamma \geq \delta} J_{0\beta\beta'\gamma\delta}^{ABCD} \sum_{\lambda\lambda'\mu\nu} C_{1,\lambda\lambda'}^{0,\beta\beta'} C_{\lambda',\mu\nu}^{\beta',\gamma\delta} \chi_{i\lambda}^{B\beta} \chi_{i\mu}^{C\gamma} \chi_{i\nu}^{D\delta} \\
 & \left. + \dots \right). \quad (5)
 \end{aligned}$$

In the second term, $C_{1,1\lambda}^{0,0\beta}$ selects invariant $\chi_{i\lambda}^{B\beta}$. The prefactors $\chi_{i1}^{A0} = x_{i0}^A$ do not affect the invariance of the polynomials in parentheses, though do increase their order by one.

If SROPs are expanded as sums of site concentrations, Eq. (5) is transformed into polynomials of occupancies, not unlike a CE. As in ACE [13], these bear some resemblance to cluster averages, though the correspondence is not exact. In contrast to a cluster basis, atom-centered terms take the form $\sim x^A(x^B)^n$, which is zero at either end of the composition range. Besides reference $U_i^A(\mathbf{0})$, which are typically zeroed as formation energies, SROPEs contain no zeroth- or first-order terms, i.e. an intercept or linear concentration. Fig. 2 compares basis functions in a random bcc alloy; while order- n CE terms average to $(2x_B - 1)^n$, SROPE polynomials vary significantly. From a clusters-perspective, defining basis functions for each species creates redundancies that could be eliminated by skipping one element in the first sum of Eq. (5). As shown in the Supplemental Material [15], simultaneously considering all elements is practically advantageous.

As another distinction, most SROPE polynomials represent multiple clusters, which may differ from those included in a comparable CE, where clusters are usually enumerated up to a maximum distance between any two sites. Fig. 1(b-l) depict examples of the clusters contained in the 11 third-order SROPE polynomials for first- and second-NNs in an fcc lattice. These include cross-shell pairs separated by twice the nominal interaction distance, though the mean distance from the geometric center of these clusters is comparable to other terms. By this metric, the clusters contained in SROPEs are generally more compact than those provided by pair-based truncation.

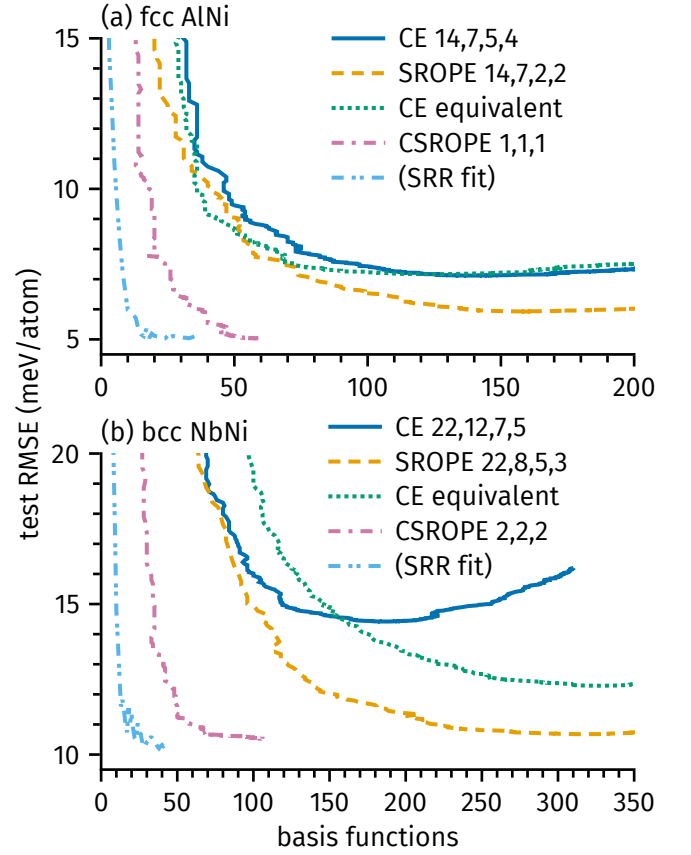


FIG. 3. Comparison of lattice models for (a) fcc NiAl and (b) bcc NbNi datasets [20]. Legends indicate the furthest neighbors pairs (CEs) or numbers of neighbor shells (SROPEs) used in second-, third-, fourth-, and fifth-order polynomials. CSROPEs comprise up-to-fourth-order polynomials of first- or second-NN SROPs and long-range composition. Test error is plotted against the number of basis functions selected by varying degrees of ℓ_1 regularization. CSROPEs were also fit using stepwise regression with replacement (SRR) [21], which provides significantly sparser solutions.

Furthermore, many SROPE polynomials include powers of individual site concentrations, or self-interactions, that are absent from CEs. For example, in the notation of Table I, the expression $\chi_1^{A0} \chi_1^{B1} \chi_1^{B1}$ contains the terms,

$$x_0^A (x_1^B \cdot x_1^B + x_2^B \cdot x_2^B + x_3^B \cdot x_3^B + \dots). \quad (6)$$

For the polynomials depicted in Fig. 1(b-l), the monomial $x_0^A \cdot x_j^B \cdot x_j^B$ is represented as a gray circle on site j . These self-interactions are usually considered undesirable due to their arguable unphysicality and effect on the conditioning of the regression problem [18, 19]. While this is a concern for the SROPE formalism, these basis functions may be otherwise advantageous. Notably, $\chi_1^{A0} \chi_1^{B1} \chi_1^{B1}$, which is represented by Fig. 1(b), directly encodes the second power of the composition in the first neighbor shell. In a CE, this term can only be partially replicated through the average of many different clusters.

Despite these differences, SROPEs and CEs compara-

bly describe certain first-principles datasets [15], such as the AgPd configurations from Ref. [4]. However, significant discrepancies emerge in other systems. As two examples, Fig. 3 examines the (a) fcc AlNi and (b) bcc NbNi datasets provided by Ref. [20]. To construct representative CEs, clusters are enumerated by iteratively increasing the maximum pair distance for each polynomial order, with legends noting converged values. As detailed in App. C, varying degrees of ℓ_1 regularization are applied to control the numbers of basis functions with nonzero coefficients. To assess the trade-off between complexity and accuracy, models are evaluated on a 20% holdout test set. For both systems, the CE slowly converges to an accuracy comparable to Ref. [20] (though reported as an average of three crystal structures).

Although the SROPE formalism most naturally involves a single cutoff defining the local chemical environment, similar optimizations can be motivated as combining a series of expansions with decreasing order over longer ranges, which is particularly justifiable if more distant SROPs are smaller. For both systems, SROPEs provide somewhat improved accuracy, though models remain quite large. In order to eliminate discrepancies due to truncation differences, the clusters contained within the SROPE cutoffs are used to construct equivalent CEs. Significant differences persist, which are attributed to more efficient parameterization of nonlinear composition dependence, perhaps due to terms such as $\chi_1^{A0} \chi_1^{B1} \chi_1^{B1}$.

Notably, SROPEs compare to the best “machine learning” interatomic potentials (MLIPs) examined by Refs. [20, 22], which would include similar constructions. While these results may not have been intended as careful comparisons, the present findings support the interpretation that MLIPs can describe the ordering energies of certain systems more accurately than conventional CEs. Consistent with this picture, Ref. [23] found no advantage for MLIPs in AgPd, where SROPEs performed similarly to CEs [15], presumably due to linear composition-dependence in the alloying energies of these chemically similar elements.

Of course, if energies directly depend on nonlocal composition ($\bar{x}^A$), expansions should explicitly consider the property. In the SROPE approach, this simply amounts to Taylor expanding $U(\mathbf{x}_i, \bar{x}^A, \dots, \bar{x}^M)$ instead of $U(\mathbf{x}_i)$. As composition is symmetrically invariant, $\bar{x}^A$ is the basis vector of a one-dimensional IIS, with symmetry corresponding to the identity representation of any $\mathcal{G}_i$. Eq. (5) equivalently expresses composition-dependent expansions, if β , γ , and δ include the invariant one-dimensional spaces formed by $\bar{x}^B$, $\bar{x}^C$, and $\bar{x}^D$. The resulting expression is referred to as a composition and short-range order parameter expansion (CSROPE), which treats the two phenomena on equal footing.

Fig. 3 includes fourth-order CSROPEs considering only (a) first- or (b) second-NN chemical environments in addition to long-range composition. Remarkably, these models achieve the accuracy of fully optimized SROPEs with only a few neighbors, one lower polynomial order,

and a fraction of the number of basis functions. However, despite the reductions in complexity, maximum accuracies are not significantly improved. This implies that optimized SROPEs are ultimately capturing the same information as CSROPEs, which may represent a limit of the dataset. In other words, most of the hundreds of basis functions in complex SROPEs seem to be inefficiently parametrizing nonlinear functions of composition, if more effectively than CEs.

When discussing model complexity, it should be noted that that ℓ_1 regularization provides moderately, but far from ideally sparse solutions. As for CEs [4, 24], ℓ_0 approaches that directly penalize nonzero basis functions produce significantly simpler SROPEs, which are both computationally cheaper and more physically interpretable. Though identifying optimally sparse solutions is challenging even for small CSROPEs, stepwise regression with replacement (SRR) [21] was found to provide excellent approximations [15], which are the final curves in Fig. 3. Effectively optimal models required only 16 basis functions for fcc AlNi and 17 for bcc NbNi. Accuracy within 1 meV/atom could be obtained with only 10 and 13, respectively.

Ultimately, CSROPEs provide a simple alternative to composition-dependent CEs such as the variable-basis cluster expansion (VBCE) [8, 10, 11]. This formalism’s foundation as a discrete wavelet transform could provide unique advantages, though model truncation still requires significant ad hoc judgments for the selection of clusters and the expression of interaction coefficients as polynomials in composition. One difference is that the VBCE basis considers site occupancies relative to the overall composition, such that basis functions become zero in random alloys, which are exclusively described by a polynomial in composition. While self-interaction terms prevent an exactly equivalent procedure in SROPEs, similar adjustments minimally affect SROPEs [15]. I.e., basis functions that remain nonzero in the random state are included in the description of the random alloy. In practice, previous demonstrations of the VBCE [8, 11] led to similar improvements in accuracy and complexity as found for CSROPEs. However, the VBCE’s rigorous derivation suggests rather restrictive philosophical prescriptions, namely that expansions only consider fixed lattices [11].

In contrast, CSROPEs simply express local properties as Taylor expansions in long-range chemistry and short-range chemical environments. By incorporating composition into geometrically compact polynomials, fourth-order expansions successfully described example systems using only first- or second-NNs. This straightforward truncation procedure requires very little optimization. Most expansion terms can be subsequently discarded through sparse regression—the CSROPEs fit using SRR required an order of magnitude fewer basis functions than conventionally constructed CEs. With compatible techniques such as basis compression [25, 26], minimalist SROPEs could particularly aid the modeling of composi-

tionally complex systems. The flexibility that allowed for the inclusion of composition can also agnostically incorporate additional couplings such as with magnetic moments. This approach has been used in ACE [27, 28], though adding degrees of freedom should be far simpler with lattice-based models. Indeed, expanding crystal properties in various order parameters represents a century-old physics technique.

ACKNOWLEDGMENTS

J.L. Kaufman, G.C. Moore, and M. Asta are thanked for helpful conversations. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 and was partly supported by the LLNL Laboratory Directed Research and Development (LDRD) program under project tracking code 22-SI-007.

Appendix A: SROPs

SROPs are identified using a procedure similar to those described in Refs. [29, 30]. For an N -dimensional chemical subspace, SROPs are generated from the eigenvectors of an arbitrary, full-rank, real-valued, $N \times N$ matrix, $\mathbf{P}$, that satisfies

$$\mathbf{P} = \mathbf{S}(g)^{-1} \mathbf{P} \mathbf{S}(g) \quad (\text{A1})$$

for all g in $\mathcal{G}_i$. A matrix $\mathbf{P}$ satisfying Eq. (A1) can be obtained by symmetrizing an initially random, symmetric, full-rank, real-valued, $N \times N$ matrix, $\mathbf{P}_0$, as $\mathbf{P} = \sum_g \mathbf{S}^{-1}(g) \mathbf{P}_0 \mathbf{S}(g)$, such that the application of further symmetry operations only permutes terms in the sum.

The eigensystem of $\mathbf{P}$ is defined as $\mathbf{P} \mathbf{e}_\kappa^\alpha = \lambda^\alpha \mathbf{e}_\kappa^\alpha$, where κ indexes the degenerate eigenvectors associated with eigenvalue λ^α . Substituting Eq. (A1) then yields

$$\mathbf{P} \mathbf{S}(g) \mathbf{e}_\kappa^\alpha = \lambda^\alpha \mathbf{S}(g) \mathbf{e}_\kappa^\alpha, \quad (\text{A2})$$

which establishes a linear relationship between $\mathbf{e}_\kappa^\alpha$ and $\mathbf{S}(g) \mathbf{e}_\kappa^\alpha$ for each λ^α . Specifically, the constraints of the eigenequation require that

$$\mathbf{S}(g) \mathbf{e}_\kappa^\alpha = \sum_{\kappa'} D_{\kappa\kappa'}^\alpha(g) \mathbf{e}_{\kappa'}^\alpha \quad (\text{A3})$$

for some constants $D_{\kappa\kappa'}^\alpha(g)$. That is, for each λ^α , all $\mathbf{S}(g)$ transform eigenvectors $\mathbf{e}_\kappa^\alpha$ into sums of other $\mathbf{e}_{\kappa'}^\alpha$, which form an IIS. Note that $\mathbf{e}_\kappa^\alpha$, and hence χ_κ^α , are arbitrary up to an IIS rotation; Table I provides manually symmetrized values.

The transformation of Eq. (A3) is described by matrices $\mathbf{D}^\alpha(g)$, which are irreducible representations of $\mathcal{G}_i$, as can be confirmed from character tables. $\mathbf{D}^\alpha(g)$ identified in this manner are used for constructing invariant polynomials.

Appendix B: Invariant polynomials

To generate invariant polynomials, products between the SROPs of IISs β and γ are decomposed onto the basis vectors of IIS α according to the Clebsch-Gordan coefficients for $\mathcal{G}_i$, which are written as $\mathcal{C}_{\kappa,\lambda\mu}^{\alpha,\beta\gamma}$. For each combination of α , β , and γ , coefficients are identified using the procedure described in Ref. [31], which applies the relation

$$\mathcal{C}_{\kappa,\lambda\mu}^{\alpha,\beta\gamma} \mathcal{C}_{\kappa',\lambda'\mu'}^{\alpha,\beta\gamma} = \frac{d^\alpha}{|\mathcal{G}|} \sum_g \mathbf{D}_{\kappa\kappa'}^\alpha(g)^* \mathbf{D}_{\lambda\lambda'}^\beta(g) \mathbf{D}_{\mu\mu'}^\gamma(g), \quad (\text{B1})$$

summing over the $|\mathcal{G}|$ operations in $\mathcal{G}$. Some κ', λ', μ' are chosen to establish a positive reference coefficient,

$$\mathcal{C}_{\kappa',\lambda'\mu'}^{\alpha,\beta\gamma} = \sqrt{\frac{d^\alpha}{|\mathcal{G}|}} \sum_g \mathbf{D}_{\kappa'\kappa'}^\alpha(g)^* \mathbf{D}_{\lambda'\lambda'}^\beta(g) \mathbf{D}_{\mu'\mu'}^\gamma(g), \quad (\text{B2})$$

which sets the phase of the remaining terms. These values are subsequently obtained from Eq. (B1)

Representing the SROP basis with $\hat{\chi}$,

$$\hat{\chi}_\kappa^\alpha = \sum_{\lambda\mu} \mathcal{C}_{\kappa,\lambda\mu}^{\alpha,\beta\gamma} \hat{\chi}_\lambda^\beta \hat{\chi}_\mu^\gamma. \quad (\text{B3})$$

Using invariant $\hat{\chi}_1^0$ as a basis function of the identity representation, the invariant components of products are identified using $\mathcal{C}_{1,\lambda\mu}^{0,\beta\gamma}$. These coefficients are numerically calculated from the irreducible matrix representations of $\mathcal{G}_i$ associated with each IIS. For d^α -dimensional IIS α matrices are written as $\mathbf{D}^\alpha(g)$ for operation g and indexed as $\mathbf{D}_{\kappa\kappa'}^\alpha$ for basis vectors χ_κ^α .

Higher-order invariants are obtained by recursively applying Eq. (B3) and selecting invariants through various intermediate couplings. For example, third-order polynomials are identified using an intermediate basis of $\hat{\chi}_\lambda^{\beta'}$ as

$$\hat{\chi}_1^0 = \sum_{\lambda\lambda'} \mathcal{C}_{1,\lambda\lambda'}^{0,\beta\beta'} \hat{\chi}_\lambda^\beta \sum_{\mu\nu} \mathcal{C}_{\lambda'\mu\nu}^{\beta',\gamma\delta} \hat{\chi}_\mu^\gamma \hat{\chi}_\nu^\delta, \quad (\text{B4})$$

where the choice of β' can lead to multiple polynomials for a given combination of β , γ , and δ . To avoid redundancy, β' is limited to unique irreducible representations. Still, this procedure can lead to linear dependencies among fourth- and higher-order polynomials [18, 32]. Though regularization techniques ameliorate overcompleteness, SROPs may benefit from more efficient basis constructions.

Appendix C: Model comparison

All models are fit using linear regression, which first requires the identification of possible features. Clusters were enumerated by constructing successively larger CEs, one polynomial order at a time. Starting with pairs, the

maximum distance between two clusters sites was incremented until ten-fold cross-validation error was judged to be converged. Then, clusters of the next-largest order were gradually added. SROPEs were similarly optimized by incrementing maximum neighbor distances for successive polynomial orders. This naive procedure is likely less than optimal, but is not unrepresentative. For fcc AlNi, the CE, SROPE, equivalent CE, and CSROPE could respectively contain 207, 1010, 304, and 60 basis functions. Equivalent numbers for bcc NbNi are 317, 3610, 1162,

and 108.

Models were trained on 80% of the datasets, including all structures with 6 or fewer atoms. RMSE error is reported on a 20% hold-out set of remaining 7 and 8 atom structures. The ℓ_1 regularization parameter was varied from 10^{-1} to 10^{-6} (for normalized input data) using GLMNet [33]. SRR was performed with forward regression initialization using `CompressedSensing.jl` [34]. CEs were performed with ICET [4] and figures were made using Makie [35].

-
- [1] J. W. D. Connolly and A. R. Williams, Density-functional theory applied to phase transformations in transition-metal alloys, *Phys. Rev. B* **27**, 5169 (1983).
 - [2] J. M. Sanchez, F. Ducastelle, and D. Gratias, Generalized cluster description of multicomponent systems, *Phys. A* **128**, 334 (1984).
 - [3] A. van de Walle and G. Ceder, Automating first-principles phase diagram calculations, *J. Phase Equilib.* **23**, 348 (2002).
 - [4] M. Ångqvist, W. A. Muñoz, J. M. Rahm, E. Fransson, C. Durniak, P. Rozyczko, T. H. Rod, and P. Erhart, ICET – A Python Library for Constructing and Sampling Alloy Cluster Expansions, *Adv. Theory Simul.* **2**, 1900015 (2019).
 - [5] J. H. Chang, D. Kleiven, M. Melander, J. Akola, J. M. Garcia-Lastra, and T. Vegge, CLEASE: a versatile and user-friendly implementation of cluster expansion method, *J. Phys. Condens. Matter* **31**, 325901 (2019).
 - [6] L. Barroso-Luque, J. H. Yang, F. Xie, T. Chen, R. L. Kam, Z. Jadidi, P. Zhong, and G. Ceder, smol: A Python package for cluster expansions and beyond, *J. Open Source Softw.* **7**, 4504 (2022).
 - [7] B. Puchala, J. C. Thomas, A. R. Natarajan, J. G. Goiri, S. S. Behara, J. L. Kaufman, and A. Van der Ven, CASM — A software package for first-principles based study of multicomponent crystalline solids, *Comput. Mater. Sci.* **217**, 111897 (2023).
 - [8] J. M. Sanchez, Cluster expansion and the configurational theory of alloys, *Phys. Rev. B* **81**, 224202 (2010).
 - [9] M. Asta, G. Ceder, and D. de Fontaine, Comment on “Nucleated and continuous ordering in Cu_3Au ”, *Phys. Rev. Lett.* **66**, 1798 (1991).
 - [10] J. M. Sanchez, Cluster expansions and the configurational energy of alloys, *Phys. Rev. B* **48**, 14013 (1993).
 - [11] J. M. Sanchez, Foundations and Practical Implementations of the Cluster Expansion, *J. Phase Equilib. Diffus.* **38**, 238 (2017).
 - [12] Y. A. Izyumov and V. N. Syromyatnikov, *Phase Transitions and Crystal Symmetry* (Kluwer, Dordrecht, 1990).
 - [13] R. Drautz, Atomic cluster expansion for accurate and transferable interatomic potentials, *Phys. Rev. B* **99**, 014104 (2019).
 - [14] A. R. Natarajan and A. Van der Ven, Machine-learning the configurational energy of multicomponent crystalline solids, *npj Comput. Mater.* **4**, 56 (2018).
 - [15] See Supplemental Material for additional figures examining species centering, an AgPd dataset, sparse regularization methods, composition normalization, and random alloys. The Supplemental Material also contains Refs. [16, 17]
 - [16] M. E. Tipping, Sparse bayesian learning and the relevance vector machine, *J. Mach. Learn. Res.* **1**, 211 (2000).
 - [17] L. Barroso-Luque and F. Xie, sparse-lm: Sparse linear regression models in Python, *J. Open Source Softw.* **8**, 5867 (2023).
 - [18] G. Dussan, M. Bachmayr, G. Csányi, R. Drautz, S. Etter, C. van der Oord, and C. Ortner, Atomic cluster expansion: Completeness, efficiency and stability, *J. Comput. Phys.* **454**, 110946 (2022).
 - [19] C. H. Ho, T. S. Gutleb, and C. Ortner, [arXiv:2401.01550](https://arxiv.org/abs/2401.01550) (2024)
 - [20] C. Nyshadham, M. Rupp, B. Bekker, A. V. Shapeev, T. Mueller, C. W. Rosenbrock, G. Csányi, D. W. Wingate, and G. L. W. Hart, Machine-learned multi-system surrogate models for materials prediction, *npj Comput. Mater.* **5**, 51 (2019).
 - [21] S. Ament and C. Gomes, *On the Optimality of Backward Regression: Sparse Recovery and Subset Selection*, in *International Conference on Acoustics, Speech, and Signal Processing* (IEEE, 2021).
 - [22] A. Bochkarev, Y. Lysogorskiy, S. Menon, M. Qamar, M. Mrovec, and R. Drautz, Efficient parametrization of the atomic cluster expansion, *Phys. Rev. Mater.* **6**, 013804 (2022).
 - [23] C. W. Rosenbrock, K. Gubaev, A. V. Shapeev, L. B. Pártay, N. Bernstein, G. Csányi, and G. L. W. Hart, Machine-learned interatomic potentials for alloys and alloy phase diagrams, *npj Comput. Mater.* **7**, 24 (2021).
 - [24] P. Zhong, T. Chen, L. Barroso-Luque, F. Xie, and G. Ceder, An ℓ_2 -norm regularized regression model for construction of robust cluster expansions in multicomponent systems, *Phys. Rev. B* **106**, 024203 (2022).
 - [25] N. Lopanitsyna, G. Fraux, M. A. Springer, S. de, and M. Ceriotti, Modeling high-entropy transition metal alloys with alchemical compression, *Phys. Rev. Mater.* **7**, 045802 (2023).
 - [26] J. P. Darby, D. P. Kovács, I. Batatia, M. A. Caro, G. L. W. Hart, C. Ortner, and G. Csányi, Tensor-Reduced Atomic Density Representations, *Phys. Rev. Lett.* **131**, 028001 (2023).
 - [27] R. Drautz, Atomic cluster expansion of scalar, vectorial, and tensorial properties including magnetism and charge transfer, *Phys. Rev. B* **102**, 024104 (2020).
 - [28] M. Rinaldi, M. Mrovec, A. Bochkarev, Y. Lysogorskiy, and R. Drautz, Non-collinear magnetic atomic cluster expansion for iron, *npj Comput. Mater.* **10**, 12 (2024).
 - [29] A. R. Natarajan, J. C. Thomas, B. Puchala, and A. Van der Ven, Symmetry-adapted order parameters and free

- energies for solids undergoing order-disorder phase transitions, *Phys. Rev. B* **96**, 134204 (2017).
- [30] F. Walsh, A. R. Natarajan, and A. Van der Ven, Order parameters for antiferromagnetic structures: A first-principles study of iridium manganese, *Phys. Rev. Mater.* **6**, 044402 (2022).
 - [31] K. Rykhinskaya and S. Fritzsche, Generation of Clebsch–Gordan coefficients for the point and double groups, *Comput. Phys. Commun.* **174**, 903 (2006).
 - [32] J. M. Goff, C. Sievers, M. A. Wood, and A. P. Thompson, Permutation-adapted complete and independent basis for atomic cluster expansion descriptors, *J. Comput. Phys.* **510**, 113073 (2024).
 - [33] J. Friedman, T. Hastie, and R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent, *J. Stat. Softw.* **33**, 3507 (2010).
 - [34] S. E. Ament and C. P. Gomes, *Sparse Bayesian Learning via Stepwise Regression*, in *Proceedings of the 38th International Conference on Machine Learning*, edited by M. Meila and T. Zhang (PMLR, 2021).
 - [35] S. Danisch and J. Krumbiegel, Makie.jl: Flexible high-performance data visualization for Julia, *J. Open Source Softw.* **6**, 3349 (2021).