









Cluster expansions of multicomponent ionic materials: Formalism and methodology

Luis Barroso-Luque ^{*}, Peichen Zhong , Julia H. Yang , Fengyu Xie , Tina Chen , Bin Ouyang , and Gerbrand Ceder [†]
Department of Materials Science and Engineering, University of California, Berkeley, California 94720, USA
and Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

 (Received 8 July 2022; revised 5 September 2022; accepted 8 September 2022; published 12 October 2022)

The cluster expansion (CE) method has seen continuous and increasing use in the study of configuration-dependent properties of crystalline materials. The original development of the CE method along with the underlying mathematical formalism and assumptions was focused on the study of metallic alloys. Since then the methodology has been actively and successfully used in the study of ionic materials as well. In this work, we present a cohesive reformulation of the mathematical formalism underlying the CE method based on a synthesis of its original formulation, several additions and extensions that have been proposed since, and a revised representation of its constituent mathematical objects. We then proceed to describe some of the formal implications of using the methodology for charge-neutral configurations in ionic systems. In particular, we discuss the reduction of the size of configuration spaces and the resulting linear dependencies that arise among correlation functions that span the larger unconstrained configuration space. Additionally, we explore the effects of long-range electrostatic interactions. We also demonstrate how the previously proposed use of a point electrostatic term successfully accounts for the majority of the longer-range electrostatic interactions, and leaves the cluster expansion terms to capture mostly short-range interactions. Finally, we present and discuss a variety of recently developed methodologies, including training structure selection, oxidation state assignment, structure mapping, and regression algorithms, that are necessary to address these formal mathematical notions for a practical implementation of the CE method in the study of multicomponent ionic materials.

DOI: [10.1103/PhysRevB.106.144202](https://doi.org/10.1103/PhysRevB.106.144202)

I. INTRODUCTION

The cluster expansion (CE) method is used to represent a coarse graining of materials properties of multicomponent crystals based on the possible configurations of species over the sites of an underlying disordered crystal structure. The CE method coupled with Monte Carlo (MC) analysis has become a standard computational tool used for calculating thermodynamic properties of crystalline materials with site disorder. By representing the formation energy in terms of atomic configurations, thermodynamic properties can then be calculated by sampling finite temperature states in MC simulations [1–3]. The CE method was originally proposed for the study of metallic alloys [4], subsequently extended to multisublattice systems [2,5], and eventually introduced in the study of oxides and ionic materials [6–10].

Since the focus in the original development of the CE method formalism was on metallic alloys, several of the nuances found in ionic systems (such as maintaining charge neutrality and describing long-range interactions) were

not initially considered. The CE method for simple ionic systems—such as CaO-MgO, which has binary configurational degrees of freedom for isovalent cations only [11]—has been shown to work well. However, there is increasing interest in applying these methods to more complex systems. Typical examples of such complex ionic systems include the recently discovered class of disordered rocksalts (DRX) [12,13], partially disordered spinels (PDSs) [14,15], and high-entropy perovskites [16], among many other promising multicomponent ceramic materials [17,18]. Specifically, DRX and PDSs are an appealing set of potential cathode materials whose systems generally involve a large number of components, redox mechanisms that require explicit treatment of cations with multiple oxidation states, and even anion disorder [19]. For these complex ionic materials, the CE and related *lattice* methods are of utmost importance to computationally probe relevant physics, such as ion percolation [20,21], short-range order [22–25], and phase diagrams [19,26]. The inherently large number of components that make these materials so promising also implies the exploration of configuration spaces with substantially larger dimension than that of binary or ternary alloys and layered oxides.

In theory, the CE method can be used to represent functions over a configuration space of arbitrary dimension. In practice, several complexities arise when fitting CE models for materials with high-dimensional configuration spaces. The most significant complexity comprises the increasing gap between the polynomial growth in the number of expansion basis functions within fixed radial cutoffs, and the number

^{*}lblueque@berkeley.edu

[†]gceder@berkeley.edu

and supercell size of training structures that are needed for successful training. Furthermore, first-principles calculations used for training cluster expansions become more involved as the complexity in the physical interactions among components increases. More complex calculations can result in additional difficulties such as insufficient training structure sampling or poorly conditioned and/or unstable feature matrices. To adequately fit cluster expansions for complex ionic materials, these practical issues can be addressed at the statistical learning level by using appropriate sampling strategies [1,27,28], regularized regression models [28,29], and possibly hierarchical and/or grouping algorithms [30–32] in order to obtain accurate and sparse CE models over high-dimensional configuration spaces.

Moreover, learning CE models of complex multicomponent ionic materials introduces an additional set of more fundamental obstacles that are not present in the case of metallic alloys. The most important difference arises from the restriction to a configuration space of only charge-neutral configurations, which implies satisfying composition constraints that often introduce linear dependencies between expansion functions. Additionally, one needs to properly account for long-range electrostatic interactions, which has been addressed to some extent by considering only structures with low electrostatic energy [6]. Lastly, correct oxidation state assignment of transition metal species with multiple oxidation states is necessary to correctly capture relevant physics in complex ionic systems [32].

In this work, we reformulate the mathematical formalism of the CE method and introduce formal notions—which have been largely absent in literature—that arise from its use in ionic materials. We then outline relevant regression models and auxiliary methodology to effectively address such implications in ionic systems and build sparse and accurate cluster expansion models of multicomponent ionic materials. The assortment of methodology presented is adapted and assessed within the context of the configuration spaces of ionic materials, and includes established methods, recently published methods by the same authors [31,32], and a handful of extensions. The presentation is done with particular focus on the use of the CE method to examine the high-dimensional configuration spaces associated with material research of complex ionic materials such as DRX and PDSs.

The paper is organized as follows: In Sec. II we cover the general formalism of the CE method and its specific implications for configuration spaces of ionic materials. Specifically, Sec. II A introduces the mathematical formalism used to represent configuration spaces and construct basis functions to represent crystal-symmetry-invariant functions of configuration. In Sec. II B 1 we describe the formal implications of charge-neutrality constraints in configuration spaces. Additionally, we discuss the practical incorporation of long-range electrostatic interactions into CE models. We then proceed to discuss and describe auxiliary methods necessary to fit cluster expansions of ionic materials using *ab initio* data in Sec. III. In Sec. III A we describe methodology that we have found useful for oxidation state assignment of systems that include species with multiple oxidation states, structure mapping methods to effectively include relaxed structures in training, and training structure sampling methods for improved accuracy and

stability of resulting CE models. We also briefly examine the implications of having configurations that are inaccessible by first-principles calculations due to mechanical and electronic instabilities. Finally, in Sec. III B we provide an overview of linear regression models with a particular focus on regression models with structured sparsity. We have found structured-sparsity regression models particularly useful when fitting CE models of complex ionic materials.

Throughout Sec. III we provide various results illustrating relevant methodology using for the most part a $\text{LiMnO}_2\text{-Li}_2\text{TiO}_3\text{-LiF}$ (LMTOF) disordered rocksalt material as an example. The LMTOF rocksalt system comprises a binary face-centered-cubic (fcc) anion lattice with O^{2-} and F^- disorder, and a fcc cation lattice with Li^+ , Mn^{3+} , and Ti^{4+} disorder. The ternary cation and binary anion disorder renders the configuration space large enough to exhibit cluster expansion complexities but still manageable to allow an efficient exploration of the methodology covered.

II. THE CLUSTER EXPANSION METHOD

A. Mathematical formalism

We start by giving an exposition of the mathematical formalism of the cluster expansion method. In its full generality, the CE method can be used to represent any scalar, vector, or tensor material property as a function of configuration [33–35]. We limit our exposition to only scalar properties—specifically formation energies of materials systems.

The specific configuration of a crystalline system can be captured by encoding the occupancy of each site in a multicomponent crystal structure using an *occupancy string*,

$$\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3, \dots), \quad (1)$$

where each element of the occupancy string $\sigma_i \in \Omega_i$ is an element of a *site space* Ω_i associated with a particular crystallographic site. In the present case, Ω_i represents the set of allowed species on the i th site [36]. In a CE, the site spaces associated with each site are usually not unique, and only a few distinct site spaces are needed to represent the configuration space of a real material; for example, ionic materials will commonly have a site space for cation sites and a site space for anion sites.

More formally, the occupancy string is an element of the product space of all included site spaces in the fully *disordered* structure. The product of site spaces in its most general form is given by

$$\boldsymbol{\sigma} \in \Omega_1 \times \Omega_2 \times \Omega_3 \times \dots = \boldsymbol{\Omega}. \quad (2)$$

The expression in Eq. (2) of the configuration space is universal. Equation (2) includes single lattice structures, such as simple alloys where all site spaces $\Omega_i = \Omega$ are the same, as well as ionic structures where more than one unique site space must be used to represent the configuration space. All sites with the same associated site spaces are said to belong to the same *sublattice* [37]. Specifically, the CE for an ionic system with more than a single sublattice was referred to as a coupled cluster expansion in its original development [7]. This scenario can be explicitly expressed by grouping equal site spaces (i.e., those corresponding to distinct anion and cation sublattices). For example, for an ionic system with one

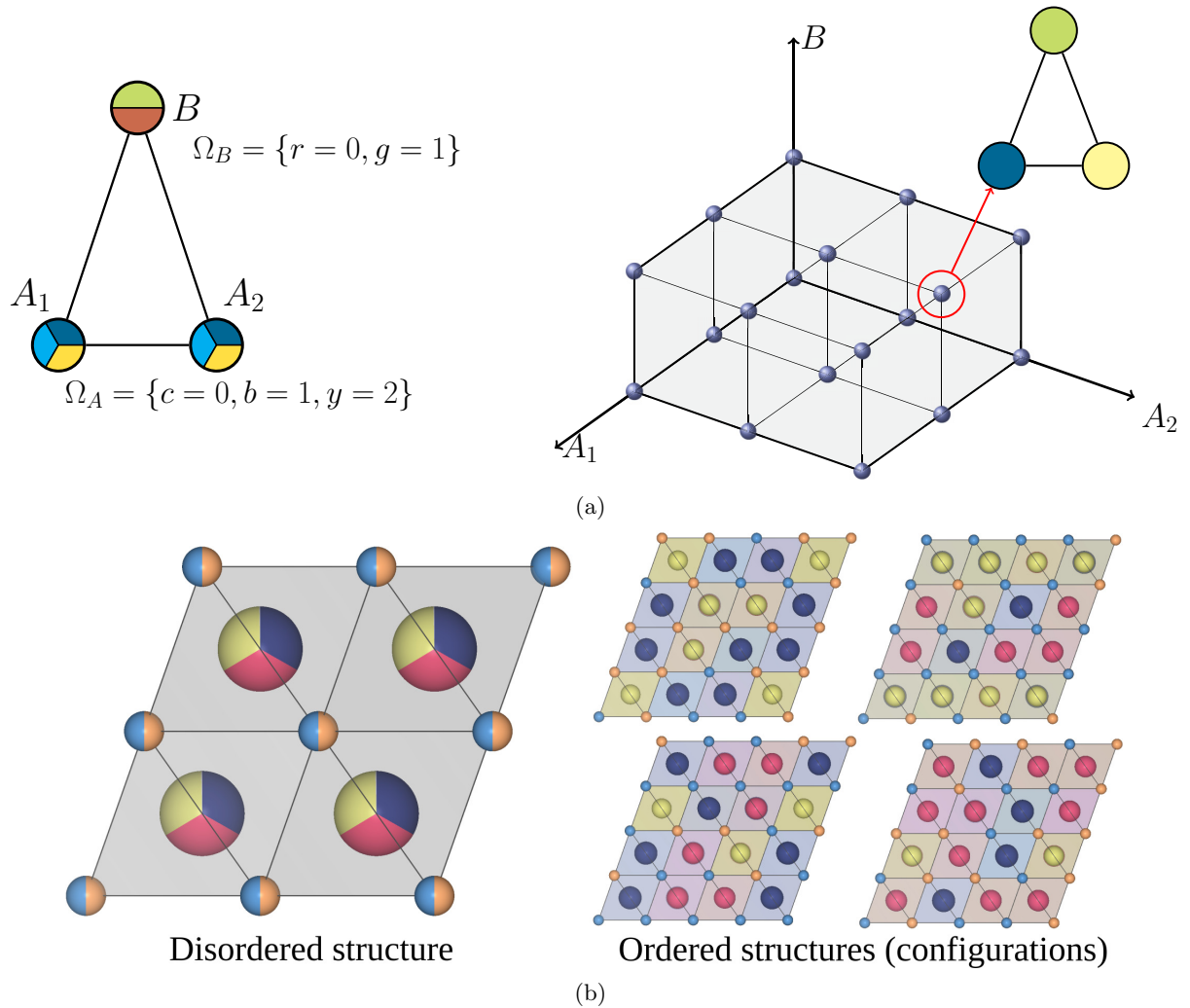


FIG. 1. (a) Illustration of the configuration space as a hypergrid for the triangular figure shown on the left. The figure has two ternary sites, A_1 and A_2 , where the allowed species are represented by the colors cyan (c), blue (b), and yellow (y), and one binary site B with allowed species red (r) and green (g). The vertex of the hypergrid corresponding to a specific configuration, $\sigma = (\text{blue } b = 1, \text{yellow } y = 2, \text{green } g = 1)$, is pointed out as an example of a point in the (A_1, A_2, B) configuration space. (b) An example for the configuration space of a two-dimensional (2D) rocksalt material represented by the underlying disordered structure on the left. A few sample configurations over a supercell of 32 sites. The hypergrid for this supercell would be in 32 dimensions and has two or three vertices per dimension depending on what type of site (cation or anion) the dimension corresponds to.

anion and one cation sublattice, the configuration space can be expressed as

$$\begin{aligned} \Omega &= (\times_{i \in A} \Omega_i) \times (\times_{i \in C} \Omega_i) \\ &= \Omega_A^{N_A} \times \Omega_C^{N_C} = \Omega_A \times \Omega_C, \end{aligned} \quad (3)$$

where A and C denote the sites in the anion and cation sublattices, respectively, and N_A and N_C refer to the number of anion and cation sites in their respective sublattices.

The configuration space in Eq. (2) can be formally represented as a *hypergrid*. Figure 1(a) shows the configuration space for a hypothetical finite three-site structure with two ternary sites (A_1 , A_2) and one binary site (B). The configuration space is depicted as the representative *disordered structure* and the corresponding three-dimensional configuration grid. Increasing the number of allowed species at a site translates to adding vertices to the hypergrid along the

corresponding dimension. And adding more sites to the structure increases the dimension of the hypergrid, since the dimension of the configuration space is equal to the number of sites in the given system. For a structure with N sites the configuration space corresponds to a hypergrid in N dimensions. Figure 1(b) shows the configuration space for a rocksalt system as its underlying disordered crystal structure. Figure 1(b) also shows four configurations for a supercell of 32 sites, which could be represented as particular vertices in a 32-dimensional hypergrid.

The CE method in its most general form is used to represent functions $H(\sigma)$ over the configuration space (hypergrid) Ω in Eq. (2),

$$H : \Omega \rightarrow \mathbb{R}. \quad (4)$$

The codomain of H in Eq. (4) does not need to be \mathbb{R} . In fact, as mentioned previously, CE models have been

used to represent vector and tensor properties of materials [33–35,38,39].

Furthermore, since the CE method is used to study crystalline materials, the method by construction ensures that the functions represented are invariant to symmetry operations of the underlying disordered structure,

$$H(T_\pi(\sigma)) = H(\sigma), \quad (5)$$

where T_π is a symmetry operation of the underlying disordered structure, which is formally represented as permutations of the elements of the configuration string σ .

In the following sections, we present the formalism and details involved in constructing the basis functions necessary to represent the functions in Eq. (4) that satisfy the necessary symmetry invariance.

1. Functions over single site spaces and configuration spaces

It has been shown in the development of the CE method [4] and is generally known from discrete harmonic analysis [40] that a basis for the function space over a product space can be obtained by taking the tensor product of basis functions over the single (site) spaces included in the product space.

Any linearly independent set of functions $\{\phi_0, \dots, \phi_{n-1}\}$ of size equal to the dimension of the corresponding site space, $n = |\Omega_i|$, constitutes a basis for the space of functions over a single site space Ω_i [40]. As such, any site function $f(\sigma_i)$ can be expressed as

$$f(\sigma_i) = \sum_{j=0}^{n-1} a_j \phi_j(\sigma_i), \quad (6)$$

where a_j are scalar expansion coefficients.

The three most common site basis sets used in the CE method are the following: (i) *polynomial* [4],

$$\phi_j(\sigma_i) = \begin{cases} \sum_{k=0}^{j/2} c_k \sigma_i^{2k} & \text{if } j \text{ is even} \\ \sum_{k=0}^{(j-1)/2} c_k \sigma_i^{2k+1} & \text{if } j \text{ is odd,} \end{cases} \quad (7)$$

where the coefficients c_k are chosen so that the basis is orthonormal; (ii) *trigonometric or sinusoidal* [2],

$$\phi_j(\sigma_i) = \begin{cases} 1 & \text{if } j = 0 \\ -\cos\left(\frac{\pi(j+1)\sigma_i}{n_i}\right) & \text{if } j \text{ is odd} \\ -\sin\left(\frac{\pi j \sigma_i}{n_i}\right) & \text{if } j \text{ is even,} \end{cases} \quad (8)$$

where n_i is the number of allowed species on the i th site; and (iii) *indicator or occupation* [41],

$$\phi_j(\sigma_i) = \begin{cases} 1 & \text{if } j = 0 \\ \mathbf{1}_{\sigma_j}(\sigma_i) & \text{if } j > 0, \end{cases} \quad (9)$$

where $\mathbf{1}_{\sigma_j}(\sigma_i)$ are singleton indicator functions, $\mathbf{1}_{\sigma_j}(\sigma_i) = 1$ if $\sigma_i = \sigma_j$ and 0 otherwise.

The CE method as originally developed [4] involves an orthonormal basis with respect to the following inner product:

$$\langle F, G \rangle = \frac{1}{|\Omega|} \sum_{\sigma \in \Omega} F(\sigma) G(\sigma). \quad (10)$$

In order to obtain an orthonormal cluster basis, it suffices to make the sets of site basis functions orthonormal over their

corresponding site space Ω_i [4,40] under an associated inner product,

$$\langle f, g \rangle = \frac{1}{|\Omega_i|} \sum_{\sigma_i \in \Omega_i} f(\sigma_i) g(\sigma_i). \quad (11)$$

We note that, from the given examples of site basis sets, only the polynomial basis is orthonormal. The trigonometric basis set is only orthogonal (not properly normalized) for site spaces with more than two allowed species. The indicator basis is not orthogonal whatsoever. Though orthonormality is not necessary, orthonormal basis sets have convenient mathematical and theoretical properties that were discussed in the original development of the CE method [4].

A basis for the function space over the product space Ω is obtained by taking the tensor product of single site bases, which in the present case is simply done by taking all possible N -fold products of site-basis functions,

$$\Phi_\alpha(\sigma) = \prod_{i=1}^N \phi_{\alpha_i}(\sigma_i), \quad (12)$$

where $\alpha = (\alpha_1, \alpha_2, \alpha_3, \dots)$ is a multi-index where each entry $\alpha_i \in \{0, \dots, n_i - 1\}$ labels the corresponding site basis function ϕ_{α_i} for each site i with a total of n_i allowed species, and N is the number of sites.

By including the constant site basis function ($\phi_0 \equiv 1$) in all site basis sets, the tensor product basis $\{\Phi_\alpha\}$ will have a *clusterlike* structure [4,42], such that the products in Eq. (12) reduce to terms that include only nonconstant site functions acting over a given cluster of sites. We call these basis sets a *cluster basis*. By using the multi-index α the clusters of sites are also conveniently indexed by the sites in the support of the multi-index (the support is defined as $\text{supp}(\alpha) = \{i; \alpha_i \neq 0\}$). The process of constructing a cluster basis for the space of functions over the configuration space depicted in Fig. 1(a) is shown schematically in Fig. 2.

Using a *cluster basis* $\{\Phi_\alpha\}$ any function over configuration space Ω can be expanded accordingly as

$$F(\sigma) = \sum_{\alpha} a_\alpha \Phi_\alpha(\sigma), \quad (13)$$

where the sum runs over all possible multi-indices α in the Cartesian product of the sets of values each element $\alpha_i \in \{0, \dots, |\Omega_i| - 1\}$ can take (i.e., the number of distinct site basis functions for each site). Written succinctly, $\alpha \in \{0, \dots, |\Omega_1| - 1\} \times \{0, \dots, |\Omega_2| - 1\} \times \dots$. This Cartesian product has an exponentially growing number of terms. When fitting a CE in practice, the summation in Eq. (13) is truncated by only considering a small set of cluster basis functions with small support (commonly $\text{supp}(\alpha) \leq 5$ is more than sufficient) and that operates over clusters of sites within small radial cutoffs in the underlying structure.

2. Symmetrically invariant functions over configuration spaces

Symmetry operations of the disordered crystal structure are applied to configuration strings as the corresponding permutation of site variables. Any such permutation must leave the function value unchanged. In order to obtain symmetry invariance, a basis of *correlation functions* is constructed from

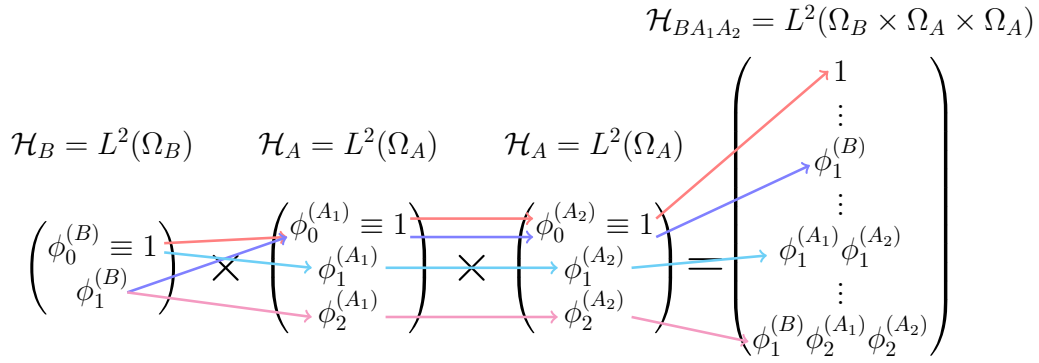


FIG. 2. Schematic illustrating the construction of a product basis from a set of site basis functions for functions over the configuration space illustrated in Fig. 1(a). The construction of subsets of product basis functions from the corresponding site basis functions are depicted with the colored arrows. The total number of basis functions for the product (configuration space) in this example is 18. The site spaces \mathcal{H}_B and \mathcal{H}_A and the product space $\mathcal{H}_{BA_1A_2}$ are L^2 Hilbert spaces over their respective domains.

a cluster basis set by applying the Reynolds operator [43,44], i.e., performing a *symmetry average*. Correlation functions are explicitly expressed by

$$\Theta_\beta(\sigma) = \langle \Phi(\sigma) \rangle_\beta = \frac{1}{N_\sigma m_\beta} \sum_{\alpha \in \beta} \Phi_\alpha(\sigma), \quad (14)$$

where β represents orbits of symmetrically equivalent multi-indices—obtained from permutations of the multi-index elements corresponding to the underlying crystallographic symmetry. N_σ is the normalization size of configuration σ , expressed in terms of the chosen normalization unit. A natural unit for size normalization is a crystallographic primitive cell. However, any unit cell choice is valid as long as it is used consistently. The constant m_β is the multiplicity of the orbit β per normalization unit.

A set of *correlation functions* represents a basis over a symmetrically invariant function subspace of the whole function space over configurations. Therefore, any symmetrically invariant function of a crystal's configuration can be expanded accordingly:

$$F(\sigma) = \sum_{\beta} N_\sigma m_\beta J_\beta \Theta_\beta(\sigma), \quad (15)$$

where J_β are expansion coefficients also known as *effective cluster interactions* (ECIs).

For notation convenience, one can group the values of all correlation vector functions for a specific configuration σ into a *correlation vector* expressed as

$$\mathbf{\Pi}(\sigma) = [\Theta_0 \equiv 1, \Theta_{\beta_1}(\sigma), \Theta_{\beta_2}(\sigma), \dots], \quad (16)$$

such that the expression for the cluster expansion in Eq. (15) can be written as the vector dot product between $\mathbf{\Pi}(\sigma)$ and \mathbf{J} ,

$$F(\sigma) = \mathbf{\Pi}(\sigma) \cdot \mathbf{J}, \quad (17)$$

where we have implicitly accounted for the multiplicities of each term m_β in the vector of expansion coefficients \mathbf{J} .

For a finite or truncated cluster expansion, with only correlations acting over a predefined and finite set of clusters, we will write

$$F(\sigma) = \mathbf{\Pi}_\sigma^T \mathbf{J}, \quad (18)$$

where $\mathbf{\Pi}_\sigma$ is a truncated correlation vector for occupancy σ ; it is implied that the coefficient vector is finite and of dimension matching that of $\mathbf{\Pi}_\sigma$.

B. Cluster expansions of ionic materials

In this section we introduce considerations for constructing CE models specifically for ionic materials. A handful of important issues arise when using the CE method to fit properties of ionic materials due mostly to the fact that species (ions) are charged. The first issue involves long-range electrostatic interactions, which has been addressed to some extent in the literature [6,9,45]. The second issue, which to our knowledge has not been formally addressed, relates to the composition constraints that configurations of an ionic material must meet in order to satisfy charge neutrality. Long-range electrostatics and charge composition constraints considered together produce another set of complications that includes species oxidation state assignment, mapping highly relaxed structures onto the fixed CE structure, and dealing with cluster configurations whose energy is inaccessible in density functional theory (DFT) calculations.

1. Charge-neutrality constraints

Charge-neutrality constraints on composition can be expressed as a sum of the oxidation states associated with the values of the elements in a configuration string in the following manner:

$$\sum_{\sigma_i \in \sigma} z(\sigma_i) = 0, \quad (19)$$

where $z(\sigma_i)$ represents a mapping from the species represented by the occupation variable σ_i to its corresponding oxidation state.

Although Eq. (19) is straightforward, it has important repercussions in both the formalism of the CE and in its practical application to ionic material systems with heterovalent ions. Composition constraints formally change the domain of valid configurations, such that the function space over the configurations of a heterovalent ionic material system is not the same as the product configuration space $\mathbf{\Omega}$ introduced in Eq. (2). More specifically, the configuration space $\hat{\mathbf{\Omega}}$ of

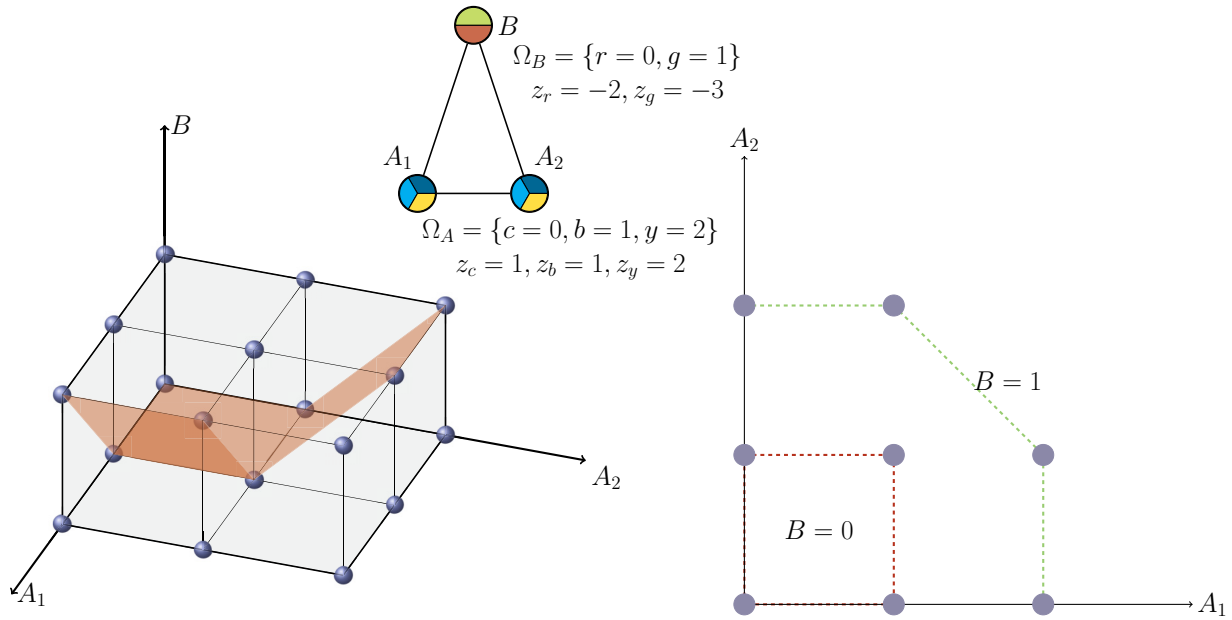


FIG. 3. Illustration of the configuration space as a slice of the hypergrid for the triangular figure shown on the top. The figure has two ternary sites A_1 and A_2 and one binary site B , and the labeled ternary sites and binary site have positive and negative oxidation states, respectively. The charge-neutral slice of the original unconstrained three-dimensional (3D) grid is shown as the grid points intersected by the orange planes. The 2D figure depicts the constrained space where the occupation of the binary site is implicit given the occupations of the two ternary sites based on charge-neutrality constraints.

an ionic material system is a set of *slices* of the product configuration space—or equivalently a set of *slices of the configuration hypergrid*, given as

$$\hat{\Omega} = \left\{ \sigma \in \Omega : \sum_{\sigma_i \in \sigma} z(\sigma_i) = 0 \right\}. \quad (20)$$

By construction $|\hat{\Omega}| \leq |\Omega|$. The constrained space $\hat{\Omega}$ is equal to the unconstrained space only for cases where all species associated with each sublattice in the system are isovalent; thus every point in the full configuration space is charge neutral. Furthermore, since the total number of functions in a product basis over Ω is precisely $|\Omega|$, a set of product basis functions is *overcomplete* for the function space over the constrained configuration space $\hat{\Omega}$ of a heterovalent ionic system. Consequentially, orthonormal or orthogonal CE basis sets have no such orthogonality properties when restricted to $\hat{\Omega}$.

Figure 3 shows an example of the configuration space with composition constraints for the previously introduced three-site system from Fig. 1(a). The composition constraints considered in the example reduce the total number of configurations in the unconstrained configuration space from 18 configurations to only 8 charge-neutral configurations. Additionally, since the configuration can be expressed by explicitly specifying the occupation of two out of the three sites, the configuration grid can be represented in a lower dimension as shown. These observations extend to higher-dimensional configuration spaces such that, when charge-neutrality constraints are considered, the dimensionality of the constrained space is effectively reduced, and the total number of configurations is in most cases substantially reduced.

The *overcompleteness* of the CE for ionic materials does not formally prevent its use, since the set still spans the functions over charge-neutral configurations. However, since it is overcomplete, the set of all correlation functions is not linearly independent. The result is that the use of an overcomplete set to express functions over a configuration space with composition constraints $\hat{\Omega}$ introduces linear dependencies between correlation functions. This implies that—in contrast to systems without any active composition constraints such as metallic alloys—material properties of ionic materials as a function of their configuration do not have a unique cluster expansion for a given set of CE correlation functions.

The simplest case can be illustrated by considering the most trivial linear relation that arises among the constant and single-site cluster functions only,

$$\sum_k \rho_k \Theta_k(\sigma) + \rho_\emptyset = 0, \quad (21)$$

where the sum runs over all orbits k with single-site clusters only, and the constants ρ_k can be obtained from the composition constraints in Eq. (19), the respective multiplicities associated with each point function, and the particular choice of site basis set used.

To further illustrate these linear dependencies, consider the constant and set of single-site correlation functions based on site indicator functions for the system in Fig. 3. The resulting linear constraint is

$$\langle \mathbf{1}_r(\sigma) \rangle - \langle \mathbf{1}_t(\sigma) \rangle - \langle \mathbf{1}_b(\sigma) \rangle - 1 = 0. \quad (22)$$

The example constraint in Eq. (22) results in linear dependencies that give rise to infinitely many expressions for the same CE. This implies that, once a set of ECI coefficients is obtained for a particular CE, the ECI can be transformed

according to

$$\begin{aligned} J'_\emptyset &= J_\emptyset - x, \\ J'_r &= J_r + x, \\ J'_t &= J_t - x, \\ J'_b &= J_b - x, \end{aligned}$$

for any scalar x . The CE with transformed ECI represents exactly the same function as the one with the original set of ECI.

Additional linear dependencies exist among higher-order correlation functions as well. Referring again to the finite example in Fig. 3, the following linear relationships exist among pair functions:

$$\begin{aligned} \langle \mathbf{1}_r(\boldsymbol{\sigma}) \mathbf{1}_t(\boldsymbol{\sigma}) \rangle - \langle \mathbf{1}_t(\boldsymbol{\sigma}) \mathbf{1}_t(\boldsymbol{\sigma}) \rangle - \frac{1}{2} \langle \mathbf{1}_t(\boldsymbol{\sigma}) \mathbf{1}_b(\boldsymbol{\sigma}) \rangle &= 0, \\ \langle \mathbf{1}_r(\boldsymbol{\sigma}) \mathbf{1}_b(\boldsymbol{\sigma}) \rangle - \langle \mathbf{1}_b(\boldsymbol{\sigma}) \mathbf{1}_b(\boldsymbol{\sigma}) \rangle - \frac{1}{2} \langle \mathbf{1}_t(\boldsymbol{\sigma}) \mathbf{1}_b(\boldsymbol{\sigma}) \rangle &= 0. \end{aligned}$$

In order to remove the resulting linear dependencies, one could in theory remove $|\boldsymbol{\Omega}| - |\hat{\boldsymbol{\Omega}}|$ functions to obtain a linearly independent set. However, for real bulk ionic systems, obtaining analytical expressions for the linear relationships among higher-order correlation functions may be too lengthy of a task, let alone constructing an orthogonal basis set which is far from trivial. Furthermore, it is not clear that removing all linear dependencies is even necessary. The cluster basis still spans the constrained configuration space, and the CE method has been successfully used as is to fit properties of ionic materials [8–10,19,21,46–50]. Indeed, the deleterious effects of linear dependencies on ECI can be managed with appropriate sampling strategies and choices of regularization during fitting, as we describe in Sec. III.

2. Long-range electrostatic interactions

The physical nature of long-range electrostatic interactions complicates a critical underlying premise of the CE method under which truncation is justified by rapid decay of correlations with respect to physical distance between sites. For example, in the case of a system with only Coulomb interactions on a rigid lattice, the terms of the CE can be easily solved for analytically [6]. For the specific case of a binary system with sites with either positive charge q_+ or negative charge q_- , an expansion using a polynomial site basis will have pair correlation terms with ECI given by

$$J_{ij} = \frac{\kappa(q_+ - q_-)^2}{4r_{ij}}. \quad (23)$$

The ECI for pair terms decays slowly as the underlying Coulomb potential $\sim r^{-1}$. This slow decay in theory requires longer pair clusters to be included in a CE to correctly capture the long-range electrostatic interactions.

It was demonstrated that when only structures with electrostatic energy below a prescribed energy cutoff are considered for the simple binary system with only $+q$ and $-q$ charged species, a CE with rapidly converging ECI can be obtained [6]. Furthermore, such CE was shown to have low prediction error for out-of-sample structures below the prescribed energy cutoff [6]. This can be attributed to the locally neutral environments associated with low-electrostatic-energy structures [6].

For more complex ionic systems, such as those with heterovalent species and/or cases including the effects of structural relaxations, considering only low-electrostatic-energy configurations and simply truncating the CE is usually not sufficient to ensure accurate and sufficiently sparse CE models with only short-range terms [45]. Even CE models with acceptable cross-validation (CV) scores may result in erroneous MC sampling—such as states with unphysical charge segregation—for large supercells when long-range interactions are not correctly accounted for.

A very effective way to handle systems with strong electrostatic interactions has been proposed and tested empirically [45,51,52]. By including an electrostatic term along with the CE Hamiltonian, a sparse and accurate CE model can be constructed much more reliably, and MC sampling is improved by more accurately computing long-range electrostatics even in large supercell sizes that were absent in the training set. To do so, the CE and electrostatic interaction Hamiltonian is expressed as the following mixture model:

$$H(\boldsymbol{\sigma}) = \sum_{\beta} m_{\beta} J_{\beta} \Theta_{\beta}(\boldsymbol{\sigma}) + \frac{1}{\epsilon_r} E_C(\boldsymbol{\sigma}), \quad (24)$$

where E_C represents the point electrostatic energy for a Coulomb potential, which can be computed efficiently and to high accuracy using the Ewald summation method [53] or the fast multipole method [54]. The constant ϵ_r , which can be interpreted as an effective dielectric constant, should be fitted simultaneously by including the electrostatic term directly as a feature in the regression problem.

To illustrate the shortcomings of a CE in capturing electrostatics in heterovalent systems and the improvements obtained when including an explicit electrostatic term, we carried out several CE fits of a Coulomb electrostatic potential, as well as a sum of a Coulomb and a Buckingham pair potential. Both potentials were computed for a system with heterovalent ($+1, +3$ cation and $-1, -2$ anion) charges in a rocksalt structure. Further details, parameters, and results of the calculations are given in the Supplemental Material [55]. The expansions used to fit the Coulomb potential included correlation functions only (i.e., no explicit electrostatic term). An expansion with correlation functions only and one with correlation functions and an electrostatic term were used to fit the Buckingham-Coulomb potential. All fits were carried out including the constant term, all point correlations, and various sets of pair correlations with increasing pair distance. Fits were also done with three different training sets: one with structures only up to 16 sites, another with structures up to 36 sites, and the last one with structures up to 64 sites. In all cases, an out-of-sample set of structures with up to the same number of sites used in training was kept for validation. An additional set with structures up to 144 sites was used to test the accuracy of extrapolated predictions to larger superstructures. For all the cases a total of 50 fits randomly shuffling training and validation structures were carried out.

Figure 4 shows curves for the resulting prediction accuracy metrics with their corresponding standard deviation for the previously described fits. Based on Figs. 4(a) and 4(b), we see that including longer-range pairs in the CE models without explicit electrostatics monotonically improves prediction

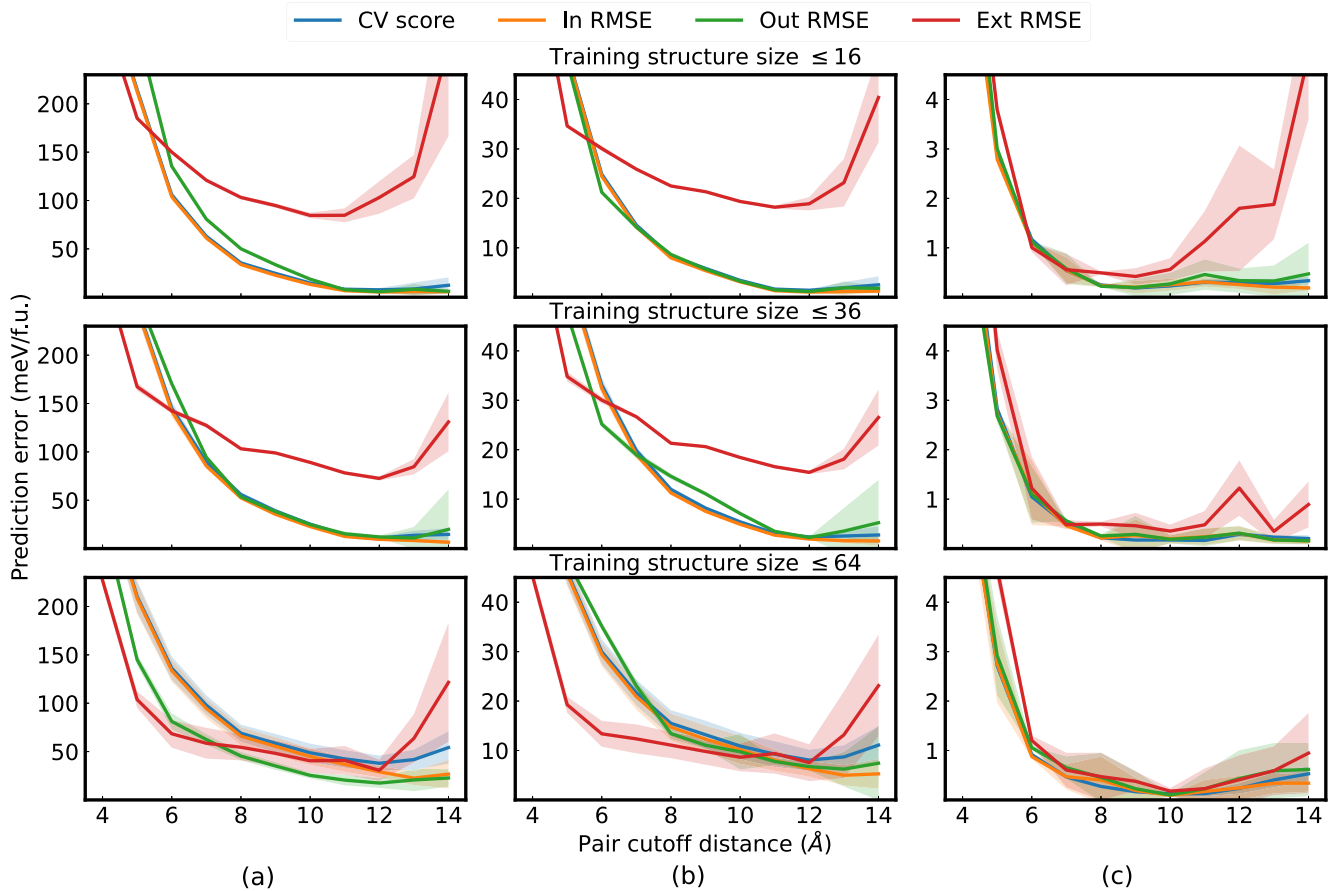


FIG. 4. Prediction accuracy metrics for CE fits of empirical pair potentials of heterovalent (+1, +3 cation) and (-1, -2 anion) charges in a rocksalt structure. The metrics shown are root-mean-square error (RMSE) cross-validation score (CV), in-sample RMSE (in), and out-of-sample RMSE (out) with supercells with the same number of sites as the listed training structure size, and extrapolation RMSE to larger supercell sizes up to 144 sites (ext). Shaded areas denote \pm one standard deviation for 50 different fits. (a) Accuracy metrics for CE fits of a Coulomb potential only. (b) Accuracy metrics for CE fits of a Buckingham-Coulomb potential. (c) Accuracy metrics for a CE and electrostatic model fits of a Buckingham-Coulomb potential.

accuracy for samples of similar sized supercells. However, the extrapolation prediction of longer-period superstructures is severely compromised. The extrapolation prediction accuracy can only be reduced by adding pairs with distances up to those sampled in the training set structures. Including pairs with longer distances that are not present in the training set ruins the extrapolation accuracy.

Figure 4 also shows that including larger-period superstructures in training improves fits, as previously suggested in similar work [45]. However, doing so, such that the resulting CE converges to an acceptable level of accuracy, requires very large (and in many cases prohibitively large) data sets that must include large supercell structures. Furthermore, when fitting CE models of ionic systems with more complex physical interactions in addition to long-range electrostatics, these issues can become worse such that fitting a reliable cluster expansion that captures both short- and long-range interactions is not straightforward.

In contrast, Fig. 4(c) shows fit metrics for the same Buckingham-Coulomb potential as Fig. 4(b), but for a fit using the CE with an explicit point electrostatic term computed with the Ewald summation method. The results show that the

addition of the point electrostatic term in the CE substantially improves the resulting accuracy. Furthermore, by setting the cutoff for CE terms relatively shorter (≤ 8 Å), the resulting fit is substantially improved and has high prediction accuracy even for longer-period superstructures. These results are also consistent with previous results computed for a similar point charge system with a spinel structure [45].

In addition, Fig. 5 shows prediction accuracy metrics and the fitted value for the effective dielectric constant in terms of the regularization hyperparameter. As the error in the fit converges, the value of the fitted dielectric constant approaches the true value used in the Buckingham-Coulomb potential, meaning that the electrostatic interactions are exactly captured by the electrostatic term, and the CE needs only to capture the short-range Buckingham interactions.

In the case of this simple additive potential, the convergence of the dielectric is only illustrative and the use of regularization is actually not necessary since the fit converges at the lowest values of the regularization hyperparameter. In fact ordinary least squares (OLS) can be used to correctly fit the Buckingham-Coulomb potential, since the short-ranged Buckingham interactions can be almost exactly captured by

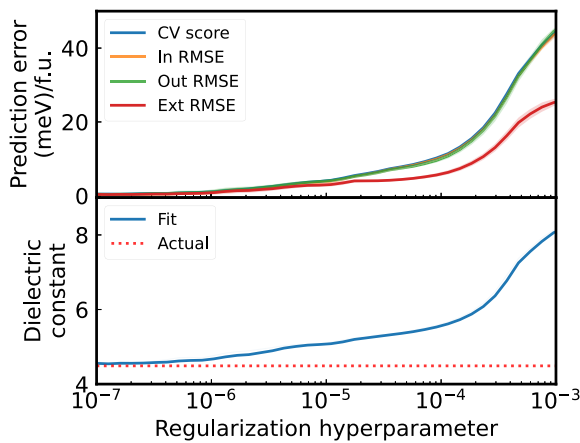


FIG. 5. Prediction accuracy and value of fitted effective dielectric constant vs Lasso regularization hyperparameter for a fit of a Buckingham-Coulomb potential using a CE and electrostatic model.

short-range correlation functions, and the electrostatic energy is exactly captured by the Ewald summation. Table I lists the fitted dielectric values and accuracy metrics for the same CE with point electrostatic term using OLS, Ridge regression, and the least absolute shrinkage and selection operator (Lasso).

The results in Figs. 4 and 5 show convincing evidence that including a point electrostatic term in a CE effectively captures the long-range point electrostatics and allows the CE to represent short-range interactions. While we cannot expect to recover a true dielectric constant when using first-principles calculations of a real material, it can be considered an *effective* dielectric constant for the model. Furthermore, the addition of the point electrostatic term has been shown to substantially improve the stability and performance of a CE fit using DFT energies, particularly for prediction values of longer-period superstructures [45,52].

III. FITTING CLUSTER EXPANSIONS TO *AB INITIO* DATA

In theory, a cluster expansion as expressed in Eq. (15) or Eq. (17) can exactly represent any function of configuration if all expansion terms—which is infinitely many for bulk systems—are included. In practice, a CE is truncated and fit using a training set of representative structures and energies calculated using first-principles methods such as DFT. The general concept of using a least-squares regression to fit a CE was first proposed as the *structure inversion* method [66]. More recently a variety of linear models with different forms of regularization have been proposed in the literature [28–30,42,67,68]. The general form of a regularized

regression optimization problem is

$$\mathbf{J}^* = \underset{\mathbf{J}}{\operatorname{argmin}} \|\Pi\mathbf{J} - \mathbf{E}\|_2^2 + \rho(\mathbf{J}), \quad (25)$$

where $\Pi \in \mathbb{R}^{m \times d}$ is a *correlation matrix* (feature matrix) where the rows are truncated correlation vectors $\Pi_\sigma \in \mathbb{R}^d$ for m training structures. $\mathbf{E} \in \mathbb{R}^m$ is a vector of calculated energies for each of the m training structures. $\mathbf{J}, \mathbf{J}^* \in \mathbb{R}^d$ are vectors of the expansion coefficients (ECI times their orbit multiplicities). The function ρ is a regularization term, which usually involves a norm or pseudonorm of the coefficients \mathbf{J} .

The overall process necessary to obtain a converged, sparse, and accurate CE model for a complex ionic material usually requires an iterative procedure. Figure 6 shows a general workflow diagram of the steps necessary to successfully fit a CE. Obtaining an adequate feature matrix Π for a CE of a real system, and in particular for a complex ionic system such as a DRX or PDS multicomponent oxides, requires a sequence of nuanced preparation steps that are still the subject of active study. Additionally, the choice of the regularization is of critical importance such that recovered ECI should in principle follow predefined priors, sparsity patterns, and/or hierarchical relations.

In this section we will first briefly describe the training data preparation and preprocessing necessary to obtain an appropriate training set (Π, \mathbf{E}) . In particular, we will briefly introduce structure sampling methods geared to obtain well-conditioned feature matrices. We also touch on methodology to effectively assign oxidation states in ionic systems using DFT magnetic moment results. We additionally describe structure matching methodology to account for large structure relaxations that commonly occur in ionic materials systems that include both oxidation states and vacancies. We also discuss the effects and offer practical solutions for handling systems with physically *inaccessible* configurations, such as those that undergo substantial relaxation and can no longer be mapped to the underlying disordered structure.

Lastly, we provide an exposition of penalized regression algorithms that yield sparse solutions and can be used to successfully fit CE models of complex ionic materials with a large number of active components. We particularly emphasize those that yield models with *structured sparsity* by including group regularization and/or hierarchical constraints. We have found that a structured sparsity regression paradigm yields more robust, accurate, and sparse models compared to standard Lasso-based fits.

The methodology discussed here is not meant to be exhaustive or conclusive. More so, it represents methodology that can be further optimized, but that we have found particularly effective in dealing with the aforementioned challenges that

TABLE I. Fitted dielectric constant and accuracy metrics in meV/f.u. using ordinary least squares (OLS), and Ridge and Lasso regression. The exact dielectric value in the model is 4.5.

Regression	Fitted dielectric	CV score	In RMSE	Out RMSE	Extrapolation RMSE
OLS	4.494	NA	0.235	0.242	0.365
Lasso	4.543	0.372	0.297	0.347	0.343
Ridge	4.502	0.314	0.263	0.272	0.337

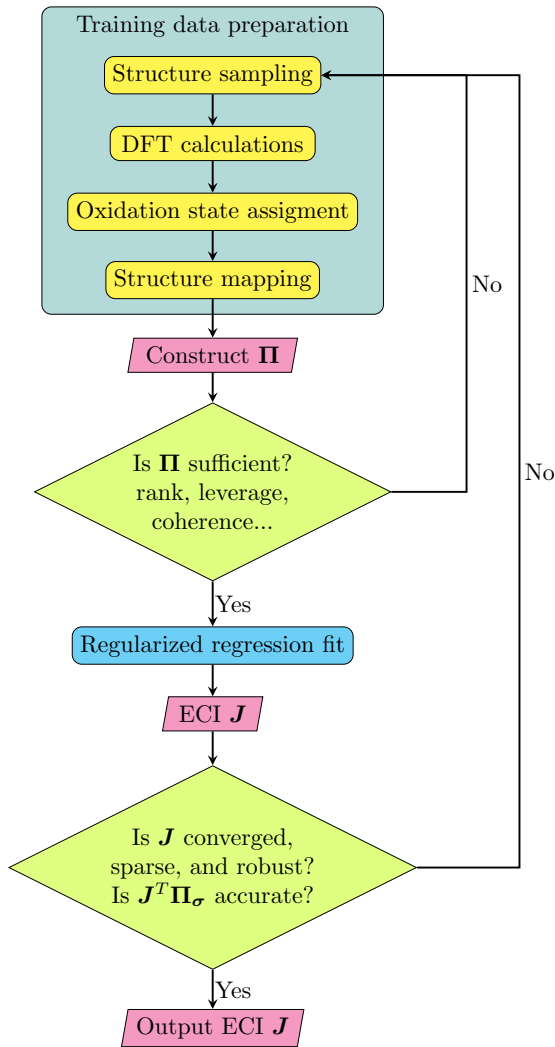


FIG. 6. General workflow diagram depicting the necessary steps required to generate and prepare training data and successfully fit a converged, sparse, and accurate CE of a complex ionic material.

occur when fitting CE models over high-dimensional configuration spaces, particularly in high-component ionic materials.

A. Training data generation and preprocessing

1. Training structure sampling

Sampling of representative training structures is a critical step to obtain useful CE models. Ideally structure sampling can cover all of the relevant areas of configuration space—areas such that CE predictions are interpolating rather than extrapolating. However, completely covering all relevant areas of very large configuration spaces is usually not possible.

The vast majority of all possible configurations tend to be concentrated at particular correlation function values [69]. These have been previously named majority structures. And structures far from those correlation values have been named minority structures [69]. The values of orthogonal correlation functions concentrate at the origin for uniformly sampled configurations in systems without any composition constraints. This has been well known in studies of metal alloys [69].

However, this is not the case in ionic systems because charge-neutrality constraints reduce the number of allowed points in configuration space as detailed in Sec. II B 1. Instead, correlation functions in ionic systems, or more generally systems with composition constraints, will concentrate at different values based on the particular set of constraints.

Figure 7(a) shows the number of structures in the LMTOF system at two correlation values for a set of charge-neutral and a set of unconstrained (including charged) uniformly random sampled structures. The vast majority of structures are concentrated around particular correlation values [yellow pixels in Fig. 7(a)]. Additional correlation sample values for a sinusoid and indicator basis as well as sampling details are described in the Supplemental Material [55]. These distributions of correlation function values can differ substantially between the case of unconstrained and constrained configuration spaces. The highly biased distribution of correlation functions in ionic systems, which results in higher coherence or similarity between correlation functions, should be considered when using structure sampling mechanisms that have been developed considering unconstrained configuration spaces only [1,27,28,67,69,70].

Structure sampling approaches generally depend on the relationship between the number of structures, m , and the number of correlation functions, d , that will be used in fitting a CE model. Based on the relationship between m and d (i.e., the shape of the correlation matrix $\mathbf{\Pi}$), the linear system in Eq. (25) can be categorized as an overdetermined problem ($m > d$) or an underdetermined one ($m < d$). For an ionic material, the full linear system is always underdetermined, but, based on the cluster cutoffs used to truncate the expansion, the resulting linear system can be made overdetermined. Structure sampling methods and their mathematical rationalization differ accordingly based on the relationship between m and d .

Most theoretical properties as well as the practical stability of regression depend on the correlation matrix $\mathbf{\Pi}$ being full rank, $\text{rank}(\mathbf{\Pi}) = \min\{m, d\}$. In other words the rank is equal to the number of columns for the overdetermined case (when $m > d$), and it is equal to the number of rows in the underdetermined case (when $m < d$). We briefly discuss the two situations and how they pertain to structure sampling of ionic materials. Figure 8 shows two flow-charts depicting established training data sampling processes used to fit a CE using an overdetermined and underdetermined (compressive sensing) linear system respectively.

For the overdetermined case, a full-rank matrix is one in which the sampled values for each correlation function (i.e., feature vectors) are linearly independent. For any finite set of samples there is likely to be a combination of intrinsic linear dependencies (those introduced in Sec. II B 1) and insufficient sampling that contribute to rank deficiencies in $\mathbf{\Pi}$. Rank deficiency can be further aggravated by configurations with energies that are inaccessible to first-principles calculations, which we address in more detail in Sec. III A 4. Furthermore, based on the aforementioned effects of charge-neutrality constraints, obtaining a full-rank overdetermined feature matrix in ionic systems is technically never possible (unless, as previously mentioned, correlation functions that give rise to intrinsic linear dependencies are removed). Appropriate sampling should seek to minimize the former effects and improve the overall rank of the correlation matrix.

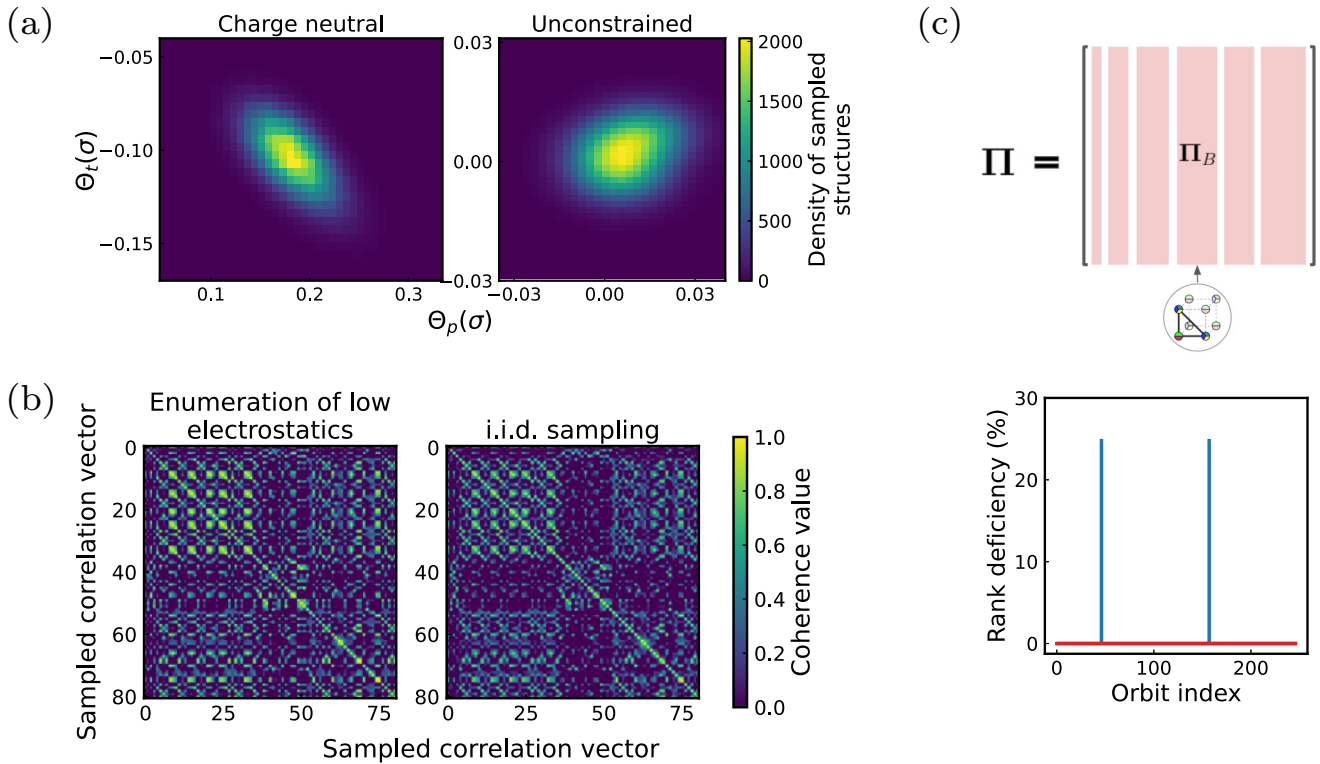
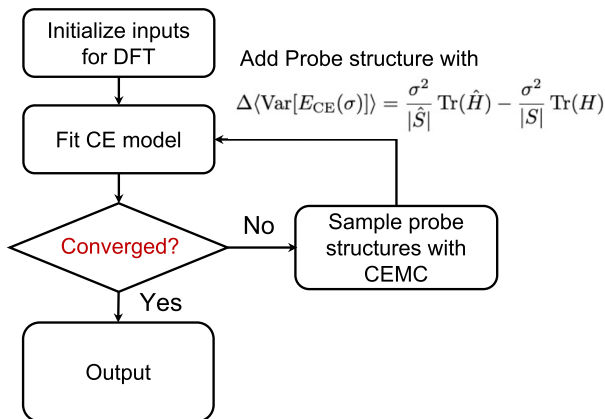


FIG. 7. (a) Histograms of uniformly random sampled structures in terms of a pair Θ_p and triplet Θ_t sinusoid basis correlation values for charge-neutral configurations only and unconstrained (any possible) configurations. (b) Gram matrices (coherence) for random sampled structures and Gaussian sampled sinusoid site basis correlation vectors for a LMTOF system using the following cutoffs: 7 Å for pairs, 5 Å for triplets, and 5 Å for quadruplet clusters (only a subset of the total 994 correlation functions are shown for better visualization). (c) An illustration of orbit submatrices making up a correlation matrix. Orbit submatrices correspond to all correlation functions that act over the same set of symmetrically equivalent clusters, as depicted by the schematic triplet cluster below. Orbit submatrix rank deficiency for a set of sampled correlation vectors for the LMTOF rocksalt system.

In overdetermined cases, even though $m > d$, the rank(Π) can be smaller than d . Under such circumstances, the rank(Π) can be increased by adding more structures to cover a wider

range of correlation values and/or by including additional correlation functions that introduce new linearly independent features. In simple systems this can minimize the rank defi-

Sampling for Overdetermined System



Sampling for Compressive Sensing

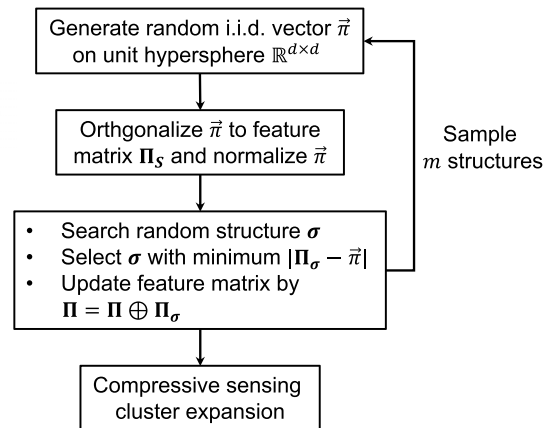


FIG. 8. (a) Sampling procedure for overdetermined problems, including initialization of inputs for DFT calculations, fit of the CE model, convergence checks, and addition of probe (additional) structures [27]. The probe structures are selected by maximizing the reduction of leverage score (uncertainty) between previous set S and new set \hat{S} . (b) Sampling procedure for the compressive sensing cluster expansion. In such a procedure, structures are selected by selecting correlation vectors Π_σ that most closely align with uniformly random vectors over the hypersphere $\vec{\pi}$ [28].

ciency up to only the trivial linear dependency between point functions from the constraint of charge neutrality, but for more complex high-dimensional systems this procedure may not be tenable based on the large number of structures that would be required to appropriately sample the fast-growing charge-constrained configuration space.

Nevertheless, overdetermined penalized linear regression, in particular variants of the Lasso (ℓ_1 -norm regularization), with rank-deficient matrices still yield valid solutions with which useful CEs can be constructed. As explained previously, the solutions will be degenerate (i.e., certain linear transformations of the estimated ECI will represent the exact same CE) [71,72], but this degeneracy is not by itself a practical point of concern. Instead the focus of structure sampling should be on improving the predictions and variances for a fitted CE for any acceptable estimates of ECI.

To simplify our analysis of prediction variance, we assume that a fitted CE model is fitted with an overdetermined, full-rank correlation matrix and captures the real target energy as follows:

$$\mathbf{E}(\boldsymbol{\sigma}) = \boldsymbol{\Pi}_\sigma^T \mathbf{J} + \boldsymbol{\varepsilon}, \quad (26)$$

where $\mathbf{E}(\boldsymbol{\sigma})$ is the real energy, and $\boldsymbol{\varepsilon}$ is a random error with *heteroskedastic* uncorrelated variances, $\text{cov}(\boldsymbol{\varepsilon}) = s^2 \mathbf{I}$.

Under the assumptions above, the variance of the predicted energy by a CE fitted with least-squares regression can be expressed as [1,27,29]

$$\text{Var}[E_{\text{CE}}(\boldsymbol{\sigma})] = s^2 \boldsymbol{\Pi}_\sigma^T (\boldsymbol{\Pi}^T \boldsymbol{\Pi})^{-1} \boldsymbol{\Pi}_\sigma, \quad (27)$$

where s^2 represents the variance from intrinsic noise in the DFT calculations for a given population of structures, and $\boldsymbol{\Pi}_\sigma$ is the truncated correlation vector for the particular occupancy $\boldsymbol{\sigma}$ used in prediction. The expression above can be adjusted for penalized regression models under a Bayesian interpretation [29]. However, for the purpose of our current explanation, Eq. (27) is sufficient.

According to Eq. (27), the average variance for predicted energies is given as

$$\begin{aligned} \langle \text{Var}[E_{\text{CE}}(\boldsymbol{\sigma})] \rangle &= \frac{\sigma^2}{|S|} \sum_{\boldsymbol{\sigma} \in S} \boldsymbol{\Pi}_\sigma^T (\boldsymbol{\Pi}^T \boldsymbol{\Pi})^{-1} \boldsymbol{\Pi}_\sigma \\ &= \frac{\sigma^2}{|S|} \text{trace}(\mathbf{H}), \end{aligned} \quad (28)$$

where S is the number of training structures. $\mathbf{H} = \boldsymbol{\Pi}^T (\boldsymbol{\Pi}^T \boldsymbol{\Pi})^{-1} \boldsymbol{\Pi}$ is the so-called hat matrix [73], and its diagonal elements H_{ii} are the predicted variances for a particular structure, which are also known in the statistics literature as *leverage scores*. The leverage score ranks the uncertainty of the corresponding probe occupancy $\boldsymbol{\sigma}$ into high-leverage or low-leverage points according to regression diagnostics [74]. A handful of methods for structure sampling have been proposed that seek to minimize the average leverage score, or equivalently maximize the reduction in average predicted variance, for each additional structure included [1,27,29]. These methods can lead to improved robustness and accuracy in CE fits of ionic systems.

For the underdetermined linear regression case ($m < d$), obtaining a full-rank correlation matrix is much more straightforward. An underdetermined system has full rank when all

correlation vectors (rows of $\boldsymbol{\Pi}$ are linearly independent), as opposed to linearly independent correlation functions. In such a case, maximizing the rank($\boldsymbol{\Pi}$) $\leq m$ instead requires obtaining m structures with linearly independent correlation vectors.

Since there are more unknowns than samples, sampling and regression for an underdetermined CE system is suitably addressed within the framework of compressive sensing (CS). As revealed by previous studies, a CS approach to cluster expansions can result in accurate and sparse solutions of ECIs using a relatively small amount of DFT measurements compared to the number of correlation functions ($m \ll d$) [28,75]. However, the necessary structure sampling for classical CS that maximizes the probability of accurate coefficient recovery has strict requirements based on the coherence—a measure of the degree of similarity—among the sampled correlation functions [67,76].

It is well known that certain probabilistic sampling methods will yield feature matrices that satisfy these requirements with high probability [28,76]. Sampling methods resulting in correlation matrices appropriate for CS have been proposed in the context of the CE method for metallic alloys. Specifically, correlation matrices appropriate for CS can be obtained by sampling correlation vectors that are random and independent and identically distributed (i.i.d.) over the unit hypersphere [28,67].

However, charge-neutrality constraints and strong electrostatic interactions complicate such random sampling in ionic systems. As an illustration of this, two normalized Gram matrices $\mathbf{G} = \boldsymbol{\Pi}^T \boldsymbol{\Pi}$ for the first 200 correlation functions from a set of 994 are shown in Fig. 7(b) for sinusoid basis correlation functions of a LMTOF system. The left-hand matrix corresponds to low-electrostatic-energy enumeration for cells up to 64 atoms, and the right-hand matrix corresponds to structures with correlations as close as possible to i.i.d. random vectors on the unit hypersphere. The i.i.d. sampling was done according to the method from Nelson *et al.* [67] from a pool of 1251 structures of supercell sizes up to 144 sites. The elements G_{ij} of Gram matrices are the dot product of sampled correlation function values i and j , which measure the level of coherence between correlation functions i and j . High coherence or similarity between sampled correlation functions is visualized as the off-diagonal yellow pixels. From left to right, although a slight decrease of the coherence values between sampled correlation vectors is successfully obtained, the maximal coherence, which is often taken as the coherence value for the full matrix, remains unchanged, and the coherence is likely too high to reliably use CS recovery of ECIs. The comparison indicates that generating structures to obtain correlation matrices that approximate i.i.d. random matrices may not be an effective way to minimize the coherence for classical CS. The full Gram matrices for all 994 correlation functions are shown in the Supplemental Material [55].

Nonetheless, in a recent study we have found that for underdetermined systems in CEs of ionic materials, the over-complete nature of the correlation basis can be leveraged under a newer variant of CS that relies on redundant expansion terms [77]. This form of CS with redundancy can be used to fit sparse and accurate CEs even with highly coherent sampling [75].

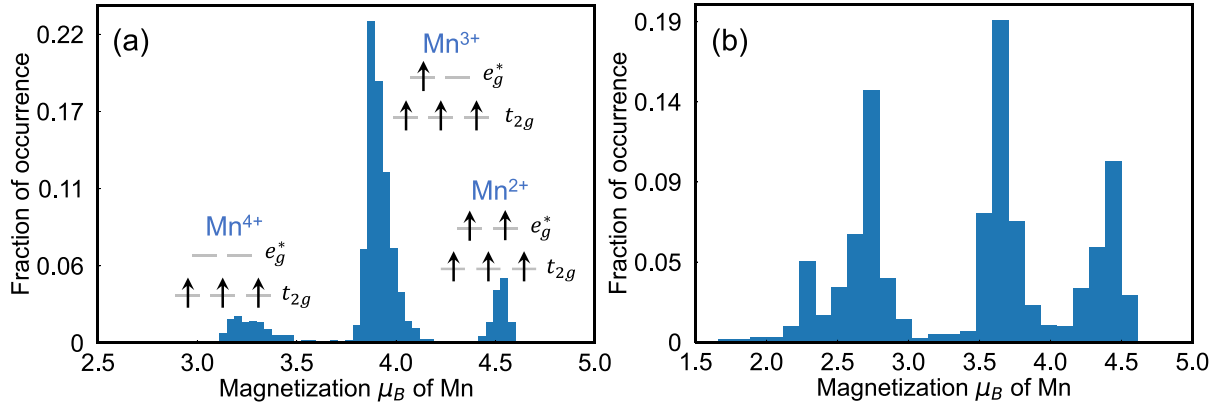


FIG. 9. (a) The magnetization distribution of Mn calculated with GGA+U in the system of $\text{Li}_{1.2}\text{Mn}_{0.6}\text{Nb}_{0.2}\text{O}_{2.0}$. The valence of each Mn atom is determined by the onsite Bohr magnetization μ_B . From the histogram, we can manually estimate the boundary for $\text{Mn}^{4+/3+}$ and $\text{Mn}^{3+/2+}$ classification to be $3.6\mu_B$ and $4.2\mu_B$. (b) The magnetization distribution of Mn calculated using the strongly constrained and appropriately normed (SCAN) density functional in the system of Li-Mn-O-F; a more continuous distribution is observed. The boundary for $\text{Mn}^{4+/3+}$ and $\text{Mn}^{3+/2+}$ classification is $3.22\mu_B$ and $4.08\mu_B$, determined by Bayesian optimization via Gaussian processes.

Though it is hard to obtain a full-rank feature matrix for overdetermined systems, or a low-coherency matrix for compressive sensing in an overdetermined system, it is still feasible to obtain accurate and well-converged CE models by also relying on appropriate use of *structured sparsity* regularization. By classifying correlation functions into groups based on the unlabeled clusters over which they operate, the correlation matrix can be analyzed in terms of *orbit submatrices*. An orbit submatrix of a given correlation matrix is made up of all the column vectors that correspond to the same orbit of unlabeled clusters (i.e., the same orbit of geometric figures). Figure 7(c) shows a schematic illustration of a correlation matrix and its orbit submatrices. For such regularization, structure sampling should strive to keep the orbit submatrices of the training correlation matrix Π full rank or as close to full rank as possible. Without full-rank (or near-full-rank) orbit submatrices grouped regularized regression may result in poorly conditioned problems and nonunique solutions. In cases where this is unavoidable, group-level and within-group regularized regression, such as using the sparse Group Lasso or Ridged Group Lasso, can be used to help avoid degenerate solutions [32,78,79]. Figure 7(c) also shows the orbit rank degeneracy, defined as one minus the ratio between the submatrix orbit rank and the total number of correlation functions in the orbit, for a set of structures of the LMTOF system. In this example only 3 of 248 orbits show a small amount of rank deficiency ($\leq 25\%$), which is sufficient to obtain accurate fits with grouped regularization as detailed in Sec. III B. We address this structured sparsity paradigm and methodology that enables it in more detail in Sec. III B.

2. Oxidation state assignment

In ionic materials containing heterovalent transition metals, it is necessary to assign formal valence to ions, since the same ion can behave differently when it has a different valence. For instance, according to crystal field theory, valence electron d filling of the transition metal–oxygen states is one factor controlling whether a transition metal ion prefers

tetrahedral or octahedral coordination. Furthermore, size and charge effects can cause metal ions to have different kinds of short-range order [22]. This thermodynamic preference arising from different formal valence necessitates treating ions with heterovalent oxidation states as different species.

However, in determining the formal valence of an ion, the DFT charge density on a metal cannot be directly used as it is invariant to the valence state due to hybridization with the anion [80]. Instead, we can rely on the magnetic moment for a given metal site to assign a formal charge and can either use the sum of s , p , and d local orbital contributions or the individual d -orbital contribution to assign this charge state. This local contribution can be obtained by integrating the spin-up minus spin-down magnetic moment around each atom.

Figure 9(a) presents a histogram of the magnetic moments on the ions in structures with composition $\text{Li}_{1.2}\text{Mn}_{0.6}\text{Nb}_{0.2}\text{O}_{2.0}$, by taking the sum of s -, p -, and d -orbital contributions. In this example, the values $\approx 3.6\mu_B$ (differentiates Mn^{4+} from Mn^{3+}) and $\approx 4.2\mu_B$ (Mn^{3+} from Mn^{2+}) have enough separation in the magnetic moments to clearly delineate oxidation states.

In other cases, the separation of oxidation states is not as obvious. For example, the histogram of the Mn d -orbital magnetic moment in the Li-Mn-O-F system [32] is shown in Fig. 9(b). It is not straightforward to define cutoff values to classify the different Mn oxidation states. In this case, one can use black-box optimization approaches (such as Bayesian optimization via Gaussian processes [81]) to assign oxidation states that are optimally consistent with a maximal number of charge-neutral structures.

More specifically, the loss function for Bayesian oxidation state assignment can be formulated as the sum of the absolute value of each structure's charge, taken over all structures in a DFT computed data set. The loss function depends on a black-box function f , which is the mapping function between any local magnetic moment for a metal to its formal valence. The exact form of f is neither known nor differentiable, but it depends solely on the magnetic moment upper cutoff for

TABLE II. Magnetic moments for Mn in three configurations of $\text{Li}_7\text{Mn}_7\text{O}_{12}\text{F}_2$ calculated with DFT-SCAN [82], and sorted into their oxidation states as determined by Bayesian optimization. The d -orbital magnetic moments and energy above hull (eV/atom) are listed.

Configuration	Mn^{2+}	Mn^{3+}	Mn^{4+}	Energy above hull (eV/atom)
A	4.207, 4.26, 4.31	3.602, 3.629, 4.017	2.916	0.133
B	4.169, 4.208, 4.264,	3.615, 3.65, 3.982	3.217	0.137
C	4.169, 4.278, 4.33, 4.366	4.07	2.703, 2.974	0.157

each different metal species of interest. For the data set used in Fig. 9(b) the function is $f(c_1, c_2, c_3)$, where c_1 , c_2 , and c_3 are three upper magnetic moment cutoffs for Mn^{2+} , Mn^{3+} , and Mn^{4+} . After black-box optimization, the upper cutoffs, corresponding to a minimal loss of structures with nonzero total charge for a given DFT data set, can be used to assign the formal valence for any structure.

Table II additionally shows three configurations of $\text{Li}_7\text{Mn}_7\text{O}_{12}\text{F}_2$, with oxidation states assigned using the recently published Bayesian optimized solution [32]. The cutoffs are $3.228\mu_B$ (differentiating Mn^{4+} from Mn^{3+}) and $4.0815\mu_B$ (differentiating Mn^{3+} from Mn^{2+}).

Configurations A and B both have three Mn^{2+} , three Mn^{3+} , and one Mn^{4+} . It is less straightforward to determine where the Mn^{3+} and Mn^{2+} cutoff lies for configuration A because $4.017\mu_B$ is closer to the magnetic moments assigned to Mn^{2+} atoms ($4.207\mu_B$, $4.26\mu_B$, $4.31\mu_B$) than to the moments assigned to Mn^{3+} atoms ($3.602\mu_B$, $3.629\mu_B$). Using Bayesian optimization circumvents this complication.

Interestingly, within configuration B the magnetic moments are more clearly separated, as the ranges of magnetic moments for Mn^{2+} and Mn^{3+} are notably less than that for configuration A, but this is not associated with a lower energy since configuration B is 4 meV/atom higher in energy. Configuration C has an entirely different set of charge orderings (four Mn^{2+} , one Mn^{3+} , and two Mn^{4+}) which can be recognized and assigned by the algorithm.

This optimization approach to assign charge states was successfully used in other chemical systems, including Li-Mn^{2+/3+/4+}-Ti-O [83] and Li-V^{4+/5+}-O [84], further supporting how Bayesian optimization can find nontrivial solutions for charge-state assignments onto magnetic moments and increase efficiency of using DFT-calculated configurations to train ionic CE.

3. Structure mapping

In practice, DFT calculations performed to obtain a set of training structures for a CE involve calculations for structures that have different supercell sizes and shapes. In many available packages [3], initial structures of the *ab initio* calculations are generated from the cluster expansion, the occupancy strings are obtained from the cluster-expansion-generated initial structures, and the energies (or other properties) are obtained from relaxation of the ionic and electronic structure. However, doing so requires that the relaxed structure still corresponds to the occupancy string from which the initial unrelaxed structure was obtained. In many cases encountered in ionic systems, ions relax too far away from their initial site, such that reassigning them to sites corresponding to the unrelaxed initial structure is infeasible. This is especially no-

ticeable in systems containing vacancies, which allow atoms to relax towards the vacant lattice site. This is also common in structures with large electrostatic or repulsive interactions because the strong interactions often force ions to maximize the distance between the interacting ions.

In cases where the structure relaxation is significant, converting the relaxed structure itself (after *ab initio* relaxation) to an occupancy string is a more appropriate way to capture the configurational energy landscape. A practical implementation of this requires a mapping between sites of the underlying disordered crystal structure and the training structure that has been relaxed by first-principles calculations. We call the ordered structure that has been appropriately mapped to the rigid lattice the *refined* structure. This mapping can then be used to construct the corresponding configuration strings σ for the relaxed structures. A schematic illustration of the relationship between the initial, relaxed, and the refined structures is shown in Fig. 10.

Formally, the procedure of structure mapping for purposes of the CE method can be stated as follows. We represent the *disordered structure* (that represents the domain of the CE) using a set of lattice vectors $L_U = [\vec{l}_1 \vec{l}_2 \vec{l}_3]$ and a set of fractional

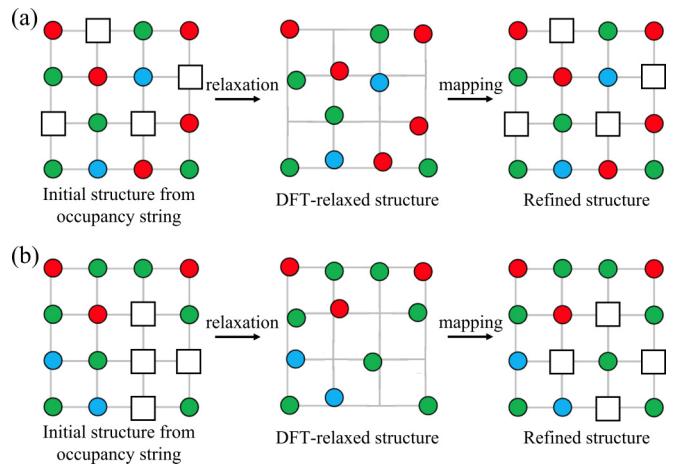


FIG. 10. Schematics of an input structure corresponding to an occupancy string σ , the resulting relaxed (DFT-calculated) structure, and a *refined* structure. The refined structure is represented by the sites of the relaxed structure mapped to the locations of the sites of the rigid disordered structure underlying the CE. The different colors represent multiple species on the lattice. The empty boxes are explicit representations of vacancies (which in the CE are treated as a species). (a) An example case where the refined structure effectively maps back to the initial structure and occupancy string. (b) An example case where the refined structure does not correspond to the initial structure or occupancy string due to substantial relaxation.

coordinates $P_U = \{\vec{p}_i, \dots, \vec{p}_{N_U} \mid \vec{p}_i \in [0, 1]^3\}$ for N_U sites. To each site we assign a site space Ω_i . Similarly, for a given ordered structure, we label the set of fractional coordinates P_Q for N_Q sites, and refer to the corresponding set of lattice vectors as L_Q . Each site in the ordered structure is occupied by a specific species σ_i . A supercell of the disordered primitive cell must be obtained to enable a one-to-one mapping between the sites of an ordered structure and the sites of the corresponding supercell of the disordered structure. We write the lattice vectors of this supercell L_{UQ} . The number of atomic sites, or equivalently the set of fractional coordinates P_{UQ} of the disordered supercell, must be the same as in the ordered structure $|P_{UQ}| = |P_Q|$. Having obtained the disordered supercell L_{PQ} , a map between the sites P_Q of an ordered structure and the sites P_{UQ} of the appropriate disordered supercell is represented by the following bijection:

$$A : P_Q \rightarrow P_{UQ} \text{ such that } \sigma_i \in \Omega_{A(i)} \forall i \in \{1, \dots, N_Q\}. \quad (29)$$

The map A can be practically established within reasonable tolerances for structural deformations of the lattice L_Q . In practice, performing these two steps (finding the disordered structure supercell, and finding the map between sites of the ordered structure and the disordered supercell) requires a crystallographic structure matching algorithm, such as the StructureMatcher in the pymatgen library [85]. A handful of other effective algorithms for crystallographic matching are freely available [86–88].

However, most approaches treat the inputs of allowed tolerances for all sites on equivalent grounds. For many ionic systems, and in particular those including vacancies, cations tend to undergo larger displacement than anions during DFT relaxation. Usually the anion sublattice undergoes less distortion and, as a result, can be more easily mapped with the predefined primitive cell. This practical observation can be revealed by comparing the drift force in DFT outputs for cations and anions, respectively. As a result structure mapping methods may fail for many ordered structures that may still have well-defined structure mappings A . One method for correcting this during mapping involves first performing a search over varying lattices to map the relaxed anions to the fixed anion sublattice sites within a fractional tolerance. Subsequently, cation centers within anion polyhedra (based on the relaxed anion-to-anion lattice site mapping) can be used to map the cation sublattice sites [32].

Effective structure mapping methods allow practical calculations of the minimum or relaxed energy landscape in terms of atomic configuration. However, it is well known—and has been numerically quantified—that the extent of structural relaxations affects the number of correlation functions required to obtain a robust and well-converged CE [89]. Rigorous quantification of strain and a corresponding metric for structure mapping may prove very useful to further establish a formal understanding of the effects of structural relaxations in the CE method. The majority of available crystallographic matching algorithms lack a rigorous quantification of the strains and symmetry breaking involved. This has only been recently addressed in a newly proposed matching algorithm [88], where cost functions for lattice strain and

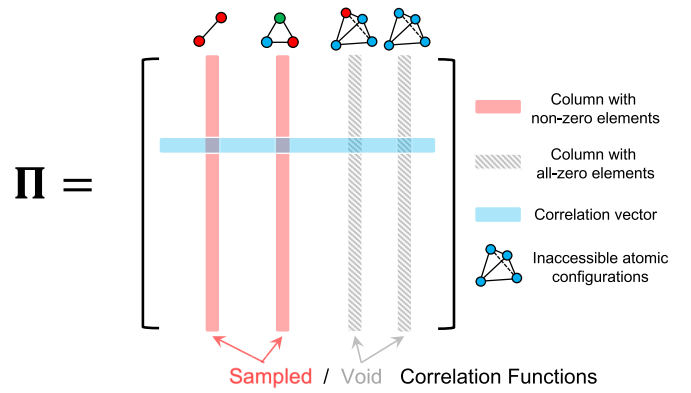


FIG. 11. Illustration of feature matrix Π with inaccessible (non-sampled) configurations using an indicator basis. The red columns represent the correlation functions that are covered by DFT calculations, while the gray (shaded) columns represent the inaccessible atomic configurations (e.g., the blue sites are occupied by high-valence transition metals such as Nb^{5+} and Mo^{6+} , which have strong repulsion in one tetrahedron and cannot be well evaluated via DFT). The blue row represents the correlation vector of one specific structure.

atomic displacement are constructed for scale-invariant geometric distortions and symmetry-breaking distortions.

4. Physically inaccessible configurations

When fitting a CE model of a complex ionic material, there will usually exist configurations that cannot be reached due to convergence issues in DFT calculations. There are two main categories of configurations that can be inaccessible to DFT: geometrical inaccessibility and charge-valence inaccessibility.

Geometrical inaccessibility occurs when the DFT-relaxed structures drift far from their original lattice sites and cannot be correctly mapped. Although Sec. III A 3 addresses some ways to find mappings when the cations relax substantially, large anion drift can make the mapping impossible. Consider, for example, anion drift that destroys the fcc anion framework of a rocksalt. Although the initial configuration may have been in the rocksalt configuration space, the resulting relaxed structure no longer is. This becomes a very notable problem when considering configurations with a large number of vacancies.

Charge-valence inaccessibility happens when the DFT-relaxed configuration can be appropriately mapped back to a lattice model with oxidation-assigned ionic species; however, charge transfer prevents specific oxidation states for particular configurations of the predefined lattice model. This happens mostly in transition metal oxides when the valence of the transition metal cannot be well assigned and results in non-charge-balanced configurations. This can also be the result of internal charge transfer in configurations with very high electrostatic energy.

The efficiency of structure sampling is thus reduced depending on how many physically inaccessible states occur in the sampled training configurations. For example, as shown in Fig. 11, the blue sites in the cluster figures are occupied by high-valence transition metal (such as Nb^{5+} and Mo^{6+}), which have strong repulsion in a single tetrahedron. Such fea-

tures cannot be appropriately computed by DFT calculations. The effect on sampling is most clear when using an indicator basis, since this will result in a void correlation function in the feature matrix $\mathbf{\Pi}$. The void correlation function manifests itself as a column with all elements equal to zero. This happens since no information has been obtained for those particular configurations, such that this correlation function is rendered uninformative and should be removed prior to fitting. For CE models with orthogonal correlation functions, the effect manifests itself more subtly. In the orthogonal case, inaccessible states will manifest as linear dependencies or equivalently rank deficiency of the corresponding orbit submatrix.

Such inaccessibility can further induce configuration sampling problems in Monte Carlo simulations. This occurs because the CE model, fitted as described above, has no information regarding the ECIs associated with the inaccessible high-energy configurations. Consider the case in which a configuration with one or more inaccessible features lies close in configuration space to a low-energy configuration (i.e., a few MC steps away). The configuration with inaccessible features may be accepted since its energy will be incorrectly predicted. The end result is that unfavorable configurations can be incorrectly sampled in MC and will distort ensemble statistics and computed thermodynamic properties.

To resolve this issue, one should include as many configurations to reduce the number of undersampled correlation functions. However, since inaccessible states are in principle caused by DFT instability, undersampled correlation functions may remain. We suggest two approaches that are useful to deal with the remaining inaccessible sampling issues. First, the ECI can be regularized with more importance given to those corresponding to lower degree clusters (such as pairwise interactions). This can be achieved by using hierarchy constraints or groupwise regularization as detailed in Secs. III B 3 and III B 2, respectively. These fitting strategies are effective when the configuration energy can be well depicted by correlations of clusters with small support; therefore, void or undersampled correlation functions for clusters with larger support will contribute minimally to the total energy.

If the resulting CE model still underpredicts the energy of configurations that are likely to be high energy, rejection of these configurations can be easily achieved in MC. The rejection can be done by including a cluster indicator function of the orbit β associated with such inaccessible atomic configurations. The probability evaluated in Monte Carlo simulation that guarantees the rejection of inaccessible configurations is

$$p \propto \exp\left(-\frac{1}{k_B T} \left(E_{\text{CE}} + \sum_{\beta \in \text{void}} M \cdot \mathbf{1}_\beta\right)\right), \quad (30)$$

where E_{CE} is the CE energy evaluated with actual ECIs, M is a large positive number, and $\mathbf{1}_\beta$ is the indicating function of orbit β . Since the cluster indicator function will only be nonzero when the specific inaccessible cluster configuration is present, all other configurations that do not include such configuration will not be affected. However, this approach requires practitioners to explicitly detect the inaccessible configurations in the first place.

B. Linear regression models

Although the use of regression for estimating ECI in the original structure inversion method [66] was based on ordinary least squares, currently some form of regularized regression is almost always used and necessary in practice. Regularized models can be derived and/or interpreted under a Bayesian framework, such that the choice of regularization function is based on the assumed prior distribution of ECI [29,90]. Apart from a possible Bayesian motivation, there are three main reasons motivating the use of regularization in linear regression and particularly for learning CE models.

The first reason, as the name *regularization* suggests, is for improving the stability of the solution to small disturbances of the sampled correlation functions. Regularization prevents the linear system from being close to singular. Due to sampling complications previously discussed, correlation matrices can be poorly conditioned. Regularization directly improves the condition number such that more numerically stable solutions can be obtained [90].

The second purpose of regularization is that of *shrinkage*, which entails forcing solutions to have small norm. The motivation for seeking regularization and shrinkage can be succinctly summed up in the *bias-variance trade-off*, where introducing a regularization term will increase the model bias—or its flexibility to represent the training data—but lower the model variance and as a result yield more stable model coefficients [90]. The bias-variance trade-off usually leads to better out-of-sample prediction accuracy at the cost of lowering in-sample prediction accuracy; in other words, it prevents overfitting.

The third reason motivating regularization involves *feature selection*. The use of specific sparsity-inducing norms for regularization results in feature selection by shrinking coefficients for less important features to zero. This allows to fit sparser and simpler CE models in one shot. Although there exist other methods for feature selection that do not rely on regularization [91,92], feature selection by regularization is now overwhelmingly used over other methods for fitting CE models [1,28,29,93].

We provide an overview of the different types of regularized linear regression models of the form given in Eq. (25), which is reproduced below for readability:

$$\mathbf{J}^* = \underset{\mathbf{J}}{\operatorname{argmin}} \|\mathbf{\Pi}\mathbf{J} - \mathbf{E}\|_2^2 + \rho(\mathbf{J}).$$

In the vector of expansion coefficients, \mathbf{J} , $\mathbf{J}^* \in \mathbb{R}^d$, the multiplicities for the actual ECI are usually treated implicitly as $\mathbf{J} = (J_0, m_{\beta_1} J_{\beta_1}, \dots, m_{\beta_{d-1}} J_{\beta_{d-1}})$. However, the ECI can be fitted directly by accounting for the multiplicities in the feature matrix instead. The function ρ is a regularization term, which usually involves a norm or pseudonorm of the coefficients \mathbf{J} . The coefficient for the data offset, or formally the empty cluster ECI, is commonly not penalized in the regularization [29,90]. Additional constraints can be added to the optimization problem in Eq. (25), such as cluster hierarchy constraints [30,94] or constraints to preserve certain configurations as ground states [94].

The focus of the remainder of this section will be mainly on the choice of regularization function $\rho(\mathbf{J})$ and the use of linear

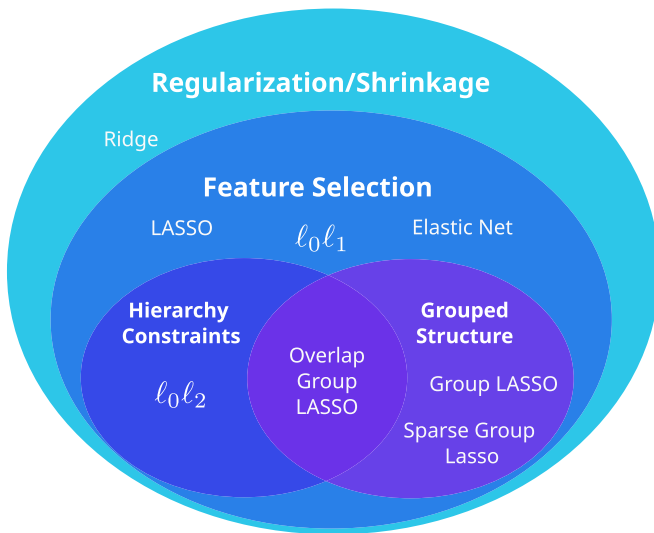


FIG. 12. Venn diagram summarizing mathematical properties of solutions of regularized regression models used to fit cluster expansions.

constraints. We provide context and motivation for their use based on physical arguments and the mathematical structure of the resulting regression problem. We refer the readers to the references noted herein for further details on mathematical formalism and algorithmic implementations for the various regularization functions described.

Figure 12 shows a Venn diagram with several regularized regression models. All regression models involve regularization and shrinkage solely by the use of norm regularization. The vast majority of regression models used for fitting CE models also involve feature selection. Lastly, a subset of regression models additionally result in solutions with structured sparsity, such that additional heuristics and/or model assumptions can be guaranteed.

We highlight methods that produce results with *structured sparsity*, where additional structure in the resulting coefficients can be obtained by using *group* norms and/or by including linear constraints in the regression optimization problem. We describe previously proposed methods and extensions for structured sparsity based on hierarchical relations between fitted coefficients [30,68]. In addition, we explore methodology to obtain solutions where coefficients are grouped by the orbits of site clusters over which the corresponding correlation functions act on. We emphasize the use of structured-sparsity fits, since we have found it to yield sparser and more accurate CE models for complex ionic systems compared to models with unstructured sparsity.

1. Sparsity-inducing regularization norms

In order to understand the properties of solutions for regression solutions with regularization norms, it is important to consider the geometry of each norm, especially the convexity and singular points of their norm balls, i.e., their level sets $||J||_p = k$ for any constant k . The solution geometry for regularized regression and the norm balls in \mathbb{R}^3 for the

regularization norms that will be described here are shown in Fig. 13.

Shrinkage can be understood as arising from the monotonically increasing nature of the regularizing norm. This means that the level sets for any of the norm balls depicted are physically larger for increasing values of the norm (with the exception of the ℓ_0 pseudonorm). Hence the norm penalization will tend to drive solutions closer to the origin compared to the ordinary least squares solution. Shrinkage and regularization can be obtained with any norm even if its norm ball is smooth everywhere. For example, the ℓ_2 norm used shown in Fig. 13 used for regularization and shrinkage in Ridge regression has a norm ball that is smooth everywhere.

On the other hand, feature selection from regularization can only be obtained by regularizing with nonsmooth norms. Feature selection occurs only when the elliptical level sets (isosurfaces) of the least-squares objective in Eq. (25) impinge on singular points (sharp edges or vertices) of the norm ball for the regularization term used. Solutions for problems using nonsmooth norms will appear with very high probability at a singular point [95]. This behavior yields sparse solutions precisely because many elements of the solution vector are exactly zero at those singular points. This solution geometry is shown in \mathbb{R}^3 for the case of the Group Lasso in Fig. 13(b), where sections of the isosurfaces of the least-squares objective are shown in different colors. Additional norms with different feature selection properties are also shown in Fig. 13(a), where one can observe that sharp edges and vertices occur at axes and/or planes spanned by the axes.

To further understand feature selection into a regression problem it is useful to introduce the ℓ_0 pseudonorm shown in Fig. 13, which can be formally defined by the limiting procedure [96]

$$||J||_0 = \lim_{p \rightarrow 0} ||J||_p^p = |\{i, : J_i \neq 0\}|.$$

The ℓ_0 norm essentially counts the number of nonzero coefficients in a vector [97] and so is a direct measurement of *sparsity*. However, the regression problem with ℓ_0 regularization is nonconvex, which is a direct result of the nonconvexity of the ℓ_0 norm ball shown in Fig. 13(b). As a result, obtaining solutions constitutes a complex combinatorial search and is in general an NP-hard problem [98].

The most common approach to obtain an approximate solution to ℓ_0 regularized regression is to solve the corresponding *convex relaxation* of the problem by replacing the ℓ_0 norm with an ℓ_1 norm [99]. Such a feature selection can be thought of as a convex relaxation of ℓ_0 selection. Linear regression using an ℓ_1 norm for regularization is known as the least absolute shrinkage and selection operator (Lasso) [100]. The Lasso has become a popular and efficient method for fitting sparse cluster expansions [28,67,75]. However, the Lasso has some notable limitations that include its lack of strict convexity and selection irregularity. In addition, the Lasso can have reduced prediction performance (compared to the Ridge) in cases with highly correlated features [101]. We have found that these issues can be practically addressed, and more robust and often sparser solutions can be obtained by using structured-sparsity-based regression.

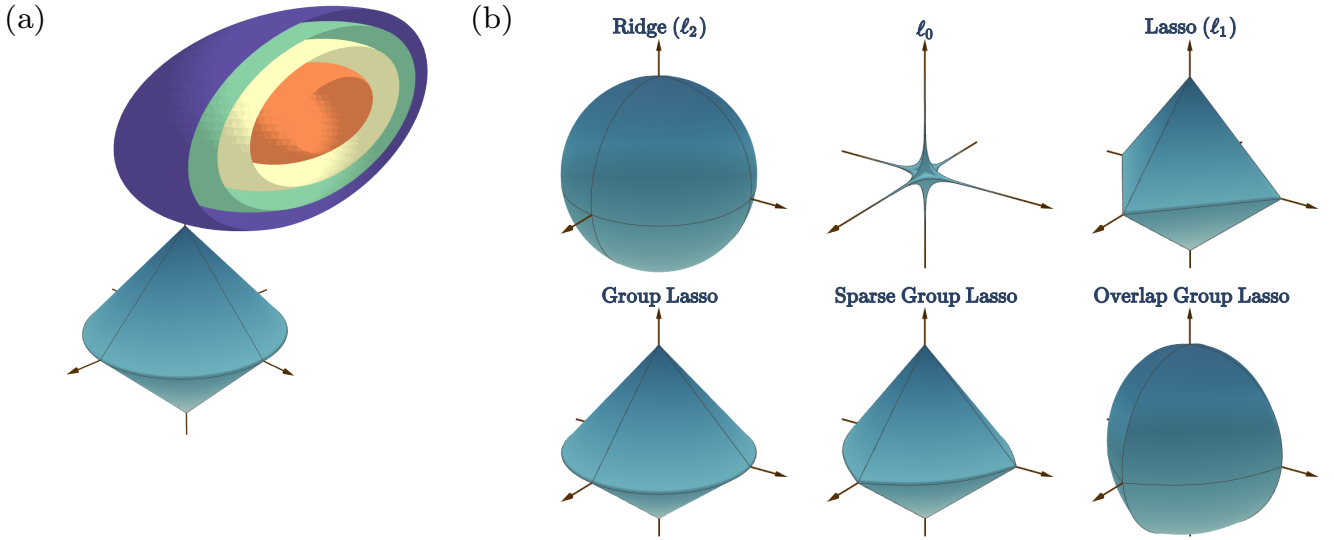


FIG. 13. (a) Sparse regression solution geometry for vector $\mathbf{J} \in \mathbb{R}^3$ using Group Lasso. The ellipsoids are isosurfaces of the least-squares solution error. The solution to the regularized regression problem in Eq. (25) in this figure is the point on the vertical axis where an ellipsoid isosurface contacts the unit norm ball. In the case shown, the solution will only include one nonzero element. (b) Unit norm balls for the solution vector \mathbf{J} corresponding to each of the regularization models described. Feature selection occurs when the OLS problem level sets contact singular points of the corresponding norm ball. The figure for the ℓ_0 pseudonorm is actually for a small value of ℓ_p , $p \rightarrow 0$ and not exactly zero. When the value of $p = 0$ exactly, the surface becomes six singular points only at values of ± 1 along each of the axes.

Several other regression models and algorithms such as the elastic net, stepwise regression, genetic algorithms [91], least angle regression, and automatic relevance determination [93] have been used to successfully fit CE models. The majority of these models, however, have only been developed for feature selection in overdetermined systems. Furthermore, we have found that Group Lasso variants and $\ell_0\ell_2$ -norm regression models that result in structured sparsity are more reliable and yield robust, sparse, and accurate CE models of complex ionic materials. After describing these structured-sparsity models, we illustrate the performance of these regression algorithms using the LMTOF disordered rocksalt system.

2. Orbit group sparsity

Selecting ECI based on grouping correlation functions by the orbits of site clusters over which they operate is a judicious form of structured sparsity for CE models. For any underlying disordered structure with three or more allowed species per site, the CE basis will have more than one correlation function acting over any orbit of symmetrically equivalent site clusters. This is illustrated schematically for a template rocksalt system in Figs. 14(a) and 14(b). Figure 14(a) shows a graphical representation for a triplet correlation function. The color labels represent function indices (the nonzero entries of a multi-index α). Figure 14(b) shows schematics for all symmetrically distinct correlation functions that operate over the orbit of site clusters represented by the colored sites. This corresponds to all the symmetrically distinct labelings of site functions over the sites of the cluster shown. This underlying structure of a multicomponent CE can be used to motivate feature selection by grouping ECI by orbits of site clusters and regularizing all functions that act on the same input variables together. This so-called orbit group regularization is shown

schematically in Fig. 14(c), where the circled figures represent groups \mathbf{g} of correlation functions. This approach to structure sparsity is mathematically and physically motivated to regularize over group correlation functions that represent a single multiple-body term in the expansion.

Orbit group sparsity can be achieved using Group Lasso regression [102]. The Group Lasso regularization problem is

$$\mathbf{J}^* = \underset{\mathbf{J}}{\operatorname{argmin}} \|\Pi\mathbf{J} - \mathbf{E}\|_2^2 + \lambda \sum_{\mathbf{g} \in G} \sqrt{|\mathbf{g}|} \|\mathbf{J}_{\mathbf{g}}\|_2, \quad (31)$$

where G is a set of groups of ECI indices \mathbf{g} (grouped by orbits of site clusters as previously described). $\mathbf{J}_{\mathbf{g}} \in \mathbb{R}^{|\mathbf{g}|}$ is a vector of only the ECI in group \mathbf{g} . The scaling $\sqrt{|\mathbf{g}|}$ is commonly used to consider all groups equally regardless of size; however, other weighting schemes can be used [72].

The Group Lasso behaves similarly to the Lasso, but feature selection occurs in groups, such that all coefficients in a group are zero or all are nonzero. This can be visualized considering the corresponding norm ball in Fig. 13, where the two grouped variables—which would correspond to the ECI of correlation functions in the same orbit group—have a continuous (circular) locus of singular points. In order for the Group Lasso to have unique solutions, each group must be full column rank [78]. As previously discussed, this represents an additional metric to consider during structure selection.

A further extension of the Group Lasso that allows for in-group sparsity, called the Sparse Group Lasso [79, 103], can yield results with improved sparsity and provide within-group regularization. In Sparse Group Lasso, an ℓ_1 norm over all coefficients is added to the ℓ_2 norm over groups as a convex

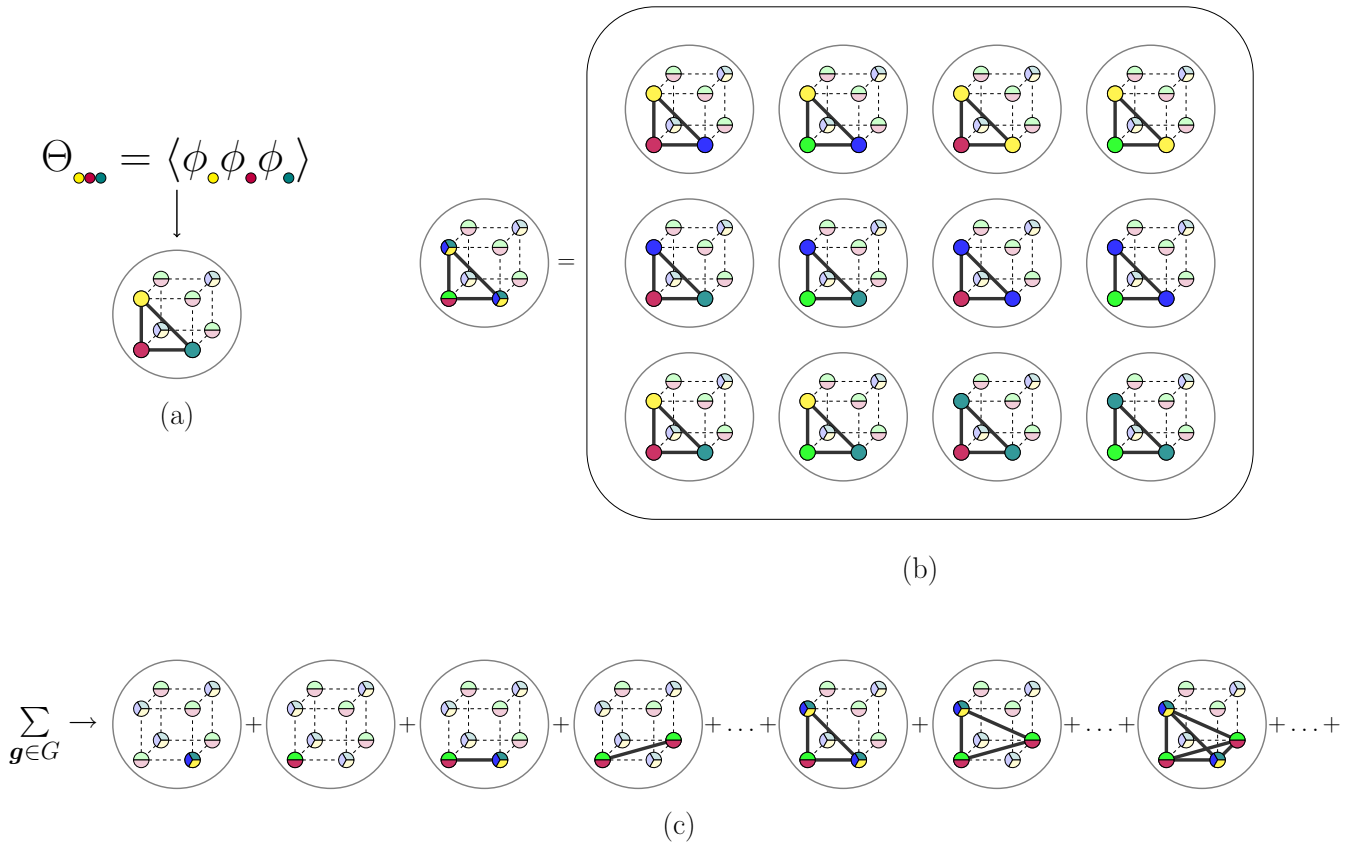


FIG. 14. Schematic illustrations of correlation functions, orbit groups of correlation functions, and structured sparsity by orbit groups for a template disordered rocksalt structure. The site coloring in the images represents nonconstant site functions. In the illustration there are two types of site spaces, one with four allowed species (three nonconstant site functions) and another with three allowed species (two nonconstant site functions). (a) Schematic of a triplet correlation function. (b) Illustration of an orbit group of triplet correlation functions. (c) Illustration of orbit group regularization by grouping correlation functions that act over the same orbits of site clusters. \mathbf{g} labels the group of correlation functions. Each circled figure in the sum represents a different group of correlation functions, analogous to the one shown in (b). G is the set of all orbit groups considered in the expansion.

combination to the regularization term:

$$\mathbf{J}^* = \underset{\mathbf{J}}{\operatorname{argmin}} \|\mathbf{\Pi}\mathbf{J} - \mathbf{E}\|_2^2 + (1 - \alpha)\lambda \sum_{\mathbf{g} \in G} \sqrt{|\mathbf{g}|} \|\mathbf{J}_{\mathbf{g}}\|_2 + \alpha\lambda \|\mathbf{J}\|_1. \quad (32)$$

Intuitively, the Sparse Group Lasso combines both the regular Lasso and the Group Lasso, as seen in both the curved edges on a group plane and the sharp vertices on all axes in the respective norm ball in Fig. 13(b). Within-group sparsity can be particularly useful for large complex cluster expansion models where charge-neutrality constraints and inaccessible configurations give rise to *suborbit* matrix rank deficiency. For cases of complex systems where orbit rank deficiency is difficult or impossible to address with sampling alone, the ℓ_1 penalization at the individual correlation level yields an additional level of regularization that improves the conditioning of the overall regression problem [32]. ℓ_2 penalization within groups has also been proposed for this reason [78].

Furthermore, the Sparse Group Lasso may yield even sparser solutions for similar levels of accuracy compared to the Group Lasso, and as a result models can have

better MC sampling performance. Furthermore, using iteratively reweighted versions (also known as adaptive versions) [104–106] of generalized Group Lasso regression models can result in substantially sparser CE models for the same level of accuracy. Further details of adaptive regression and improved sparsity in fits can be found in the Supplemental Material [55]. In Sec. III B 5 we show how the adaptive Group Lasso and especially the adaptive Sparse Group Lasso result in models that have high accuracy and surpassed levels of sparsity compared to Lasso-only solutions.

3. Hierarchically constrained sparsity

Another compelling form of structured sparsity involves establishing hierarchical relations between correlation functions. This can be used to enforce physically motivated heuristics, such as the inclusion of correlation functions over larger clusters only if correlation functions over all subclusters are included [1,30,31,68,107]. We have investigated two different forms of hierarchically constrained sparsity that prove to be quite effective for fitting CE models of complex ionic materials.

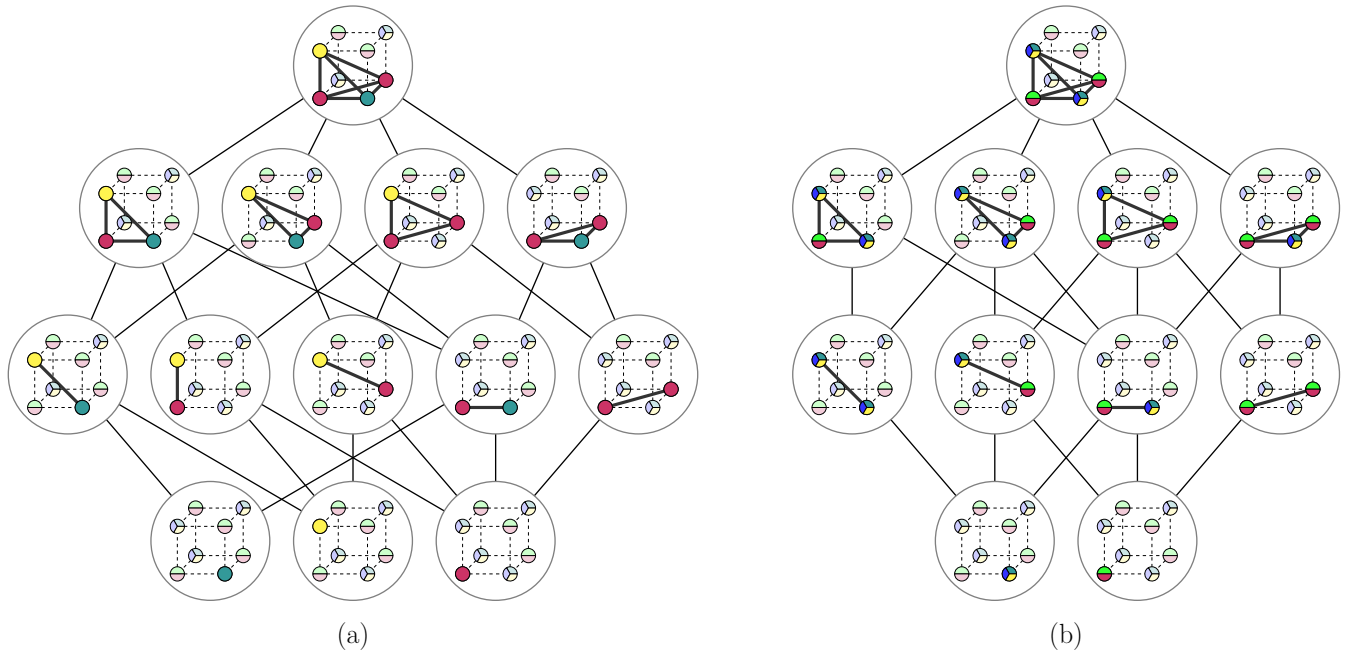


FIG. 15. Schematic illustrations of hierarchically constrained sparsity for a template rocksalt structure. The site coloring in the images represents nonconstant site functions. In the illustration there are two types of site spaces, one with four allowed species (three nonconstant site functions) and another with three allowed species (two nonconstant site functions). (a) Hierarchical relations for a specific quadruplet correlation function and all its possible factors. (b) Hierarchical relations between groups of correlation functions acting over the same orbits of quadruplet clusters and all correlation function groups acting over the orbits of subclusters of the quadruplet cluster.

The first involves imposing hierarchical constraints between higher-degree correlation functions and their lower degree factors. Higher-degree correlation functions are only allowed to have nonzero coefficients if all of their lower-degree factors do so too. This form of structured sparsity is shown schematically in Fig. 15(a), where the constraints between correlation functions and their factors are represented by edges connecting them. This form of hierarchically constrained sparse regression has been recently applied for fitting CE models of ternary alloys and disordered ionic materials [30,31,68].

In the alternative form of hierarchy constraints, correlation functions are grouped by their associated orbits of site clusters (as described for orbit group regularization in Sec. III B 2). In this case, the hierarchical constraints are between the groups of correlation functions. A group of correlation functions acting over the same orbit of clusters can only have nonzero coefficients if all the groups of correlation functions that act over the orbits of all subclusters have nonzero coefficients as well. This form of structured sparsity is essentially the combination of hierarchical constraints and the orbit group structure previously introduced. A representation of this hierarchy structure is shown in Fig. 15(b) as a graph representing the hierarchical relations between orbit groups of correlation functions. To the best of our knowledge, this regression model has not been previously used for fitting CE models [108]. We show in Sec. III B 5 that this structure-sparsity form yields solutions that are competitively accurate and better aligned with physical heuristics at the cost of lower sparsity due to the more restrictive hierarchical constraints that are imposed.

Hierarchically constrained sparsity can be obtained by using an extension of the Group Lasso that allows for overlap-

ping groups—known as the *Overlap Group Lasso* [109]—and casting the problem in terms of auxiliary variables [30]. Additionally, using a regression algorithm with a convex combination of ℓ_1 and ℓ_0 , linear constraints can be added as a way to obtain weakly hierarchically constrained sparse solutions [68]. A related variant using $\ell_2\ell_0$ regularization allows strict enforcement of hierarchical constraints while still resulting in suitably sparse solutions [31]. The $\ell_0\ell_2$ -regularized regression problem is

$$\mathbf{J}^* = \underset{\mathbf{J}}{\operatorname{argmin}} \|\Pi\mathbf{J} - \mathbf{E}\|_2^2 + \alpha\lambda\|\mathbf{J}\|_0 + (1 - \alpha)\lambda\|\mathbf{J}\|_2, \quad (33)$$

where the regularization hyperparameter $\alpha \in [0, 1]$ is constrained to lie between zero and one.

The ℓ_0 -regularized regression problem in Eq. (33) is an NP-hard problem, but suitable near-optimal solutions can be found for moderately sized CE models (up to 500 ECI) using *mixed-integer quadratic programming* (MIQP). This transformation of the regression problem into MIQP is also what allows the introduction of *hierarchical* constraints as linear constraints on auxiliary slack variables. The problem in Eq. (33) transformed to MIQP takes the following form:

$$\begin{aligned} \min_{\mathbf{J}} \quad & \mathbf{J}^T \Pi^T \Pi \mathbf{J} - 2\mathbf{E}_S^T \Pi_S \mathbf{J} + \alpha\lambda \sum z_\beta + (1 - \alpha)\lambda \mathbf{J}^T \mathbf{J} \\ \text{such that} \quad & Mz_\beta \geq J_\beta \\ & Mz_\beta \geq -J_\beta \\ & z_\beta \in \{0, 1\}, \end{aligned} \quad (34)$$

where z_β is a slack variable that describes whether a correlation function Θ_β is active ($z_\beta \neq 0$) or inactive ($z_\beta = 0$).

The introduction of cluster hierarchy constraints corresponding to those depicted in Fig. 15(a) to the ℓ_0 norm with MIQP is straightforward. Since $z_\beta \in \{0, 1\}$, such hierarchical constraints can be expressed in terms of the slack variables z_β as

$$z_\beta \leq z_\gamma, \quad \forall \alpha(\beta) \subset \alpha(\gamma), \quad (35)$$

where β and γ are function cluster orbits, and the notation $\alpha(\beta) \subset \alpha(\gamma)$ indicates that for any function cluster $\alpha(\beta) \in \beta$, there exists a function cluster $\alpha(\gamma) \in \gamma$ such that $\alpha(\beta)$ is a subcluster of $\alpha(\gamma)$. A function cluster with multi-index $\alpha(\beta)$ is a subcluster of another function cluster with multi-index $\alpha(\gamma)$ if all the nonzero entries of $\alpha(\beta)$ are contained in $\alpha(\gamma)$.

Fitting CE models using the MIQP paradigm that satisfy correlation function hierarchical constraints has been shown to result in faster-converging, robust, and more physically accurate models for some disordered materials [31,68]. Applying these types of hierarchical constraints using the Overlap Group Lasso has also been shown to yield robust and accurate CE models for refractory ternary alloys [30]. We corroborate these results in Sec. III B 5 for small- and medium-sized models (<500 correlation functions). However, for larger models, obtaining near-optimal solutions to ℓ_0 -regularized problems may become too computationally intensive even with state-of-the-art integer solvers [31].

4. Hyperparameter selection

The regularized regression models introduced have at least one hyperparameter associated with the regularization term, and models that mix more than one norm, such as the $\ell_2\ell_0$ or Sparse Group Lasso, have two hyperparameters. Selecting appropriate hyperparameters is critically important since the hyperparameters control the importance given to a regularization term and consequently the amount of shrinkage and/or feature selection. As a result, the hyperparameters strongly affect the resulting prediction accuracy. The standard way to determine these hyperparameters is by using cross-validation (CV) optimization.

Determining a hyperparameter value using CV optimization involves minimizing a CV score, most commonly the root-mean-square error (RMSE), with respect to the relevant hyperparameter. CV involves splitting the available training data randomly into k sets of equal size. Subsequently k fits are computed using the data from all combinations involving $k - 1$ sets. The CV score is the average RMSE for all k fits computed with respect to the k th set that was not included in each fit. When using k sets, the procedure is called k -fold CV, and the most commonly used values of k are 1, 5, and 10. For $k = 1$ the procedure is known as leave-one-out CV (LOOCV), and its use in learning CE models has been extensively discussed [1,70].

Choosing the number of folds for CV is a choice left to the practitioner, and there are no hard rules on which and when to choose a particular value of k . Nevertheless, we can say that in general smaller values of k tend to produce models with lower bias (and higher variance) and may have a tendency to exhibit overfitting. Larger values of k will show lower variance but higher bias which may affect overall model performance, particularly considering the fact that training data for CEs is

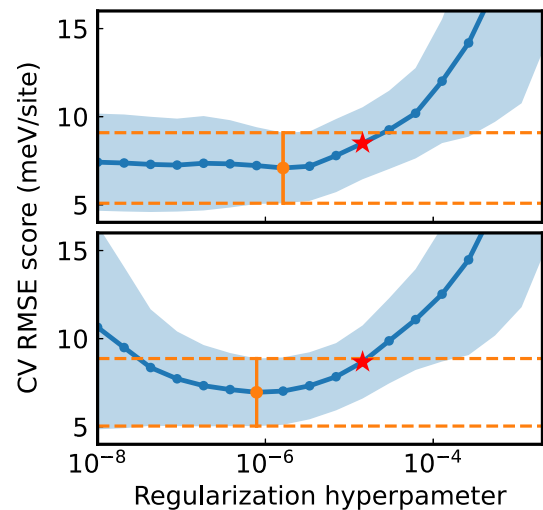


FIG. 16. CV score regularization paths for Sparse Group Lasso fits of an LMTOF rocksalt system. The top plot shows the path for a fit using pairs up to 7 Å, triplets up to 4.2 Å, and quadruplets up to 4.2 Å. The bottom plot shows the path for a fit using pairs up to 7 Å, triplets up to 5.6 Å, and quadruplets up to 5.6 Å.

expensive and often scant. A good recommended compromise for the number of folds is 5 or 10 [92].

In addition, we emphasize what is known as the *one-standard-error rule* [92], because it is particularly applicable to CE fitting. The one-standard-error rule states that when choosing a hyperparameter with feature selection (sparsity) as one of the goals, it is recommended to choose the largest value of the hyperparameter for which the CV RMSE is within one standard deviation of the minimum CV RMSE [92] and results in better sparsity. The reason behind this is that the hyperparameter value that minimizes CV error optimizes for prediction accuracy but not for feature selection, and a sufficient reduction in model complexity may be well worth the cost in terms of a slightly larger CV RMSE.

Figure 16 shows the regularization paths for two fits of the LMTOF system using Sparse Group Lasso regression with different sets of cutoffs. The mean CV score is shown in blue, and the standard deviation is shaded. The minimum CV score is marked in yellow, and the corresponding standard deviation region is marked with dashed lines. According to the one-standard-error rule, the models that should be chosen are marked with a red star. Although the one-standard-error rule by itself should not be taken as a definitive rule for CE model selection, it serves as a general guidance for practitioners to select models that are both accurate and parsimonious, rather than solely optimizing CV score at the cost of sparsity. This is particularly important to keep in mind for the common *CV-plateau* scenario, which is present in the top plot of Fig. 16. In systems exhibiting a CV plateau the CV minimum can often occur at hyperparameter values far into the plateau region, and as a result using the hyperparameters for the CV minimum results in models with severely compromised sparsity and only marginal improvements in CV score compared to those obtained following the one-standard-error rule.

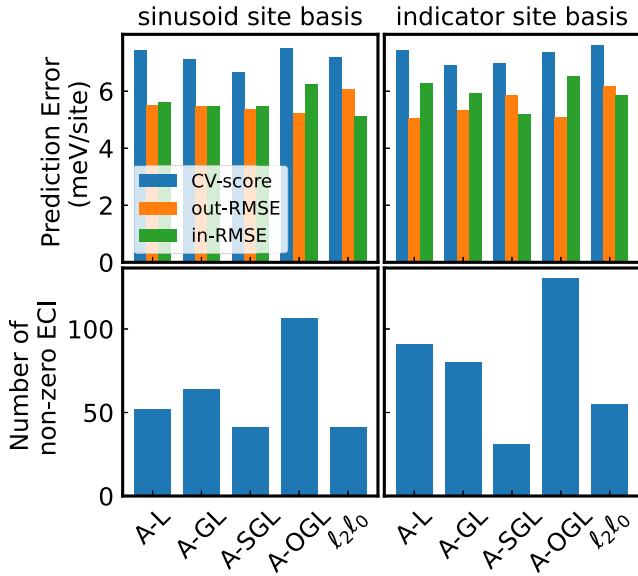


FIG. 17. Fitted LMTOF CE accuracy metrics and resulting model sparsity using Lasso and structured-sparsity-based regression algorithms: A-, adaptive variants; L, Lasso; GL, Group Lasso; SGL, Sparse Group Lasso; OGL, Overlap Group Lasso. All fits shown were done using correlation functions for cluster size cutoffs of 7, 4.2, and 4.2 Å for pair, triplet, and quadruplet clusters, respectively, using a primitive cell of the rocksalt structure with lattice parameter $a = 3$ Å.

5. Fits of a LMTOF system

CE fits of the LMTOF disordered rocksalt system were computed using standard Lasso and the structured sparsity-based regression models previously introduced. All fits include an explicit electrostatic term as expressed in Eq. (23), which is computed using the Ewald summation method. We compare the resulting model prediction accuracy, sparsity, and ECI structure of the various fits using a training set of DFT calculated energies for 983 structures with supercells up to 72 atoms. An additional test set of 247 structures of supercell sizes 128 and 132 atoms is used for validation. Additional details of the DFT training structure calculations and CE fitting are reported in the Supplemental Material [55].

Figure 17 shows prediction accuracy metrics for fits using each regression model with cluster size cutoffs of 7, 4.2, and 4.2 Å, for pair, triplet, and quadruplet clusters, respectively. Fits were carried out using a sinusoid site basis and an indicator site basis. Hyperparameter tuning curves for the various regression models are reported in the Supplemental Material [55]. Fits with different cutoffs which show similar trends as those shown in Fig. 17 were also computed and the results are reported in the Supplemental Material [55]. From the results in Fig. 17 we see that all regression models yield similar levels of predictive accuracy. However, although all regression models achieve some degree of feature selection (the total number of correlation functions in the truncated correlation matrix was 143), Sparse Group Lasso and l_2l_0 regression are the most effective in reducing the total number of features required to achieve similar levels of accuracy. This is further evidenced in the Supplemental Material [55] by the additional

fits we calculated for different sets of cluster cutoffs. We make note specifically that Overlap Group Lasso has the worst performance in feature selection due to the restrictive hierarchical constraints imposed, as described in Sec. III B 3.

Figure 18 shows the resulting sets of fitted ECI for each regression model using sinusoid and indicator site basis sets. Although we will not attempt to make statements regarding the *interpretability* of the fitted ECI values, there are some notable observations and trends regarding Fig. 18. First, fits with an indicator site basis tend to result in higher overall ECI magnitudes regardless of the regression model used. This can be attributed to the lack of orthogonality at the site-basis level and, as a result, the higher coherence values of sampled correlation values. However, structured sparsity models on average result in lower-magnitude ECI compared to the Lasso. Furthermore, although the solutions obtained with these regression models are not unique, the different models tend to identify a few apparently important correlation functions, in particular short-range pair correlations and some larger-diameter triplet correlations. Lastly, hierarchy-based regularization, and in particular the orbit-level hierarchy implemented with the Overlap Group Lasso, results in ECIs that much better align with physical intuition and heuristics (i.e., decay with physical distance and cluster size), albeit, for overlap group Lasso, this occurs at the cost of obtaining less sparse models as previously discussed.

All in all, the results from the fitted expansions for the LMTOF system shown in Figs. 17 and 18 and the accompanying results in the Supplemental Material [55] demonstrate how expansions with structured sparsity have similar or improved levels of accuracy as those from the Lasso, and additionally tend to have higher sparsity and trends in the resulting ECI that much better aligns with physical priors and heuristics.

IV. CONCLUSION

We have given a revised and extended presentation of the mathematical formalism of the CE [4] method integrated with the extension to multiple sublattice systems [7]. We have further described the formal implications of using the formalism for configuration spaces with charge-neutrality constraints, particularly those with heterovalent cations and anions. Charge-neutrality constraints give rise to the linear dependencies between correlation functions which essentially render the full set of CE correlation functions overcomplete for the space over charge-neutral configurations. We also showed how including an explicit point electrostatic term effectively captures long-range electrostatic interactions allowing correlation functions with short associated cluster diameters to capture short-range interactions more effectively [45,52].

In addition, we provided a cohesive overview of data preparation methodology and regression algorithms that are useful to successfully fit CEs of complex multicomponent ionic materials. In particular, we explicitly addressed some of the differences and nuances of applying structure sampling and structure mapping to complex ionic systems—which have been largely unaddressed in the literature. We also briefly described methods and issues arising from oxidation state assignment and physically inaccessible configurations

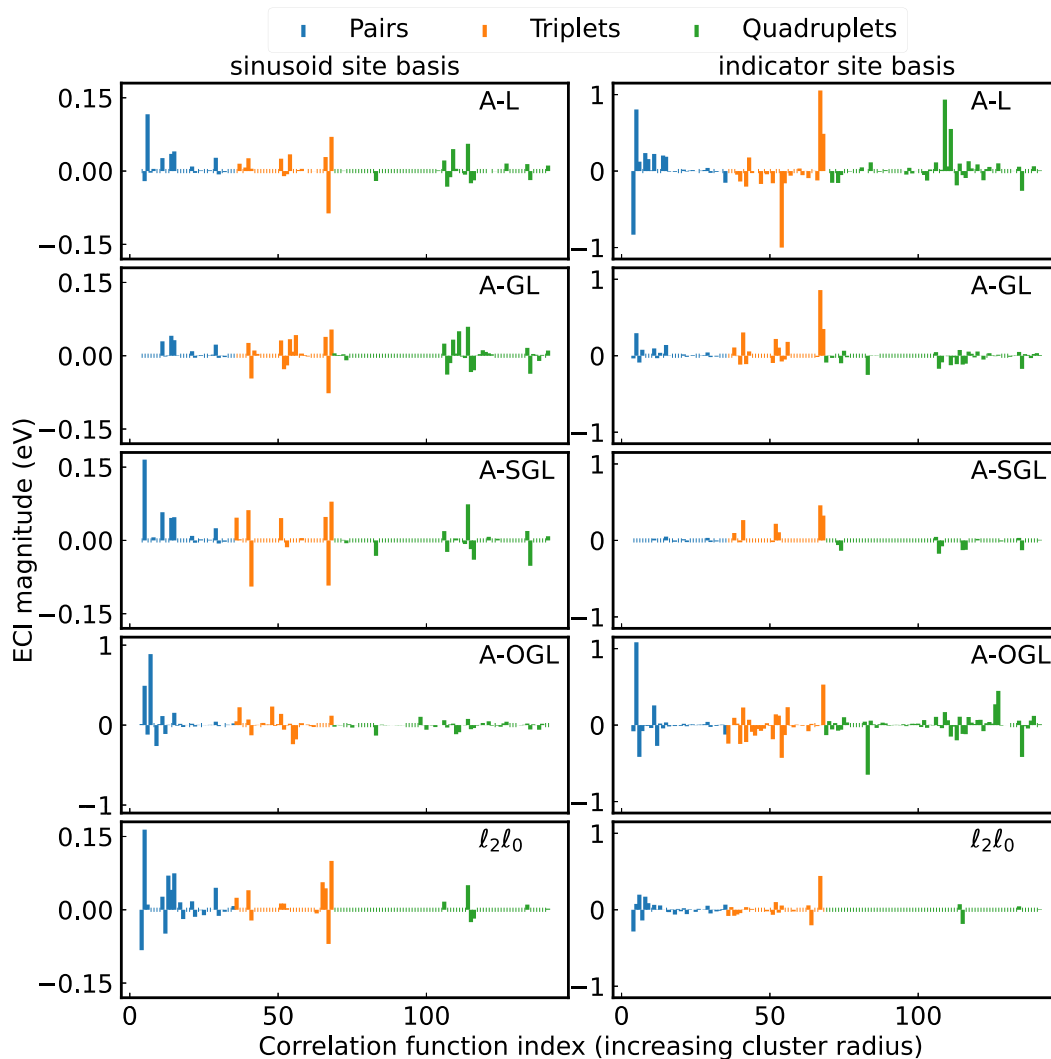


FIG. 18. Fitted LMTOF effective cluster interactions using adaptive Lasso and structured sparsity-based regression algorithms: A-, adaptive variants; L, Lasso; GL, Group Lasso; SGL, Sparse Group Lasso; OGL, Overlap Group Lasso. All fits shown were done using correlation functions for cluster size cutoffs of 7, 4.2, and 4.2 Å for pair, triplet, and quadruplet clusters, respectively, using a primitive cell of the rocksalt structure with lattice parameter $a = 3$ Å.

that commonly arise in ionic systems. Finally, we described regularized linear regression models and placed particular emphasis on models resulting in structured sparsity. We showed how structured sparsity is an effective way to address the theoretical and practical nuances that occur in complex ionic materials and results in more robust models that can also have higher sparsity compared to the commonly used Lasso model.

Overall, successful CE fits of complex ionic materials require careful evaluation of the discussed implications, such as long-range electrostatics and charge-neutrality constraints, as well as effective management of the effects that arise in applications, namely, linear dependencies, large structural and/or electronic relaxations, and inaccessible configurations. This requires selection of the appropriate sampling methods based on the end purpose of the CE, the regression model, and regularization that will be used in the fit. In particular, structured-sparsity regression models that allow the introduction of mathematically or physically motivated constraints result in more robust and sparse CE models than those fitted

with simpler algorithms such as the Lasso. We have discussed two recently proposed structured-sparsity paradigms and appropriate regression algorithms to implement them. For correlation and/or orbit-level hierarchy constraints the l_2l_0 [31] or the Overlap Group Lasso algorithms can be used. For smaller systems (up to ~ 500 ECIs), the former has been shown to yield fast-converging and physically accurate CE models [31]. However, due to its NP-hard nature, applying it to larger models is inefficient. In such cases the Overlap Group Lasso can be used to implement hierarchical constraints [30], which as a convex problem scales more favorably to larger problems. We have further described Group Lasso and Sparse Group Lasso implementation of orbit group regularization, where all ECIs for correlation functions that act over the same clusters of sites are penalized together. This form of structured sparsity has also been shown to give accurate and highly sparse CE models, which we illustrated in the current work, and has also been effectively used to fit one of the largest CE models to date [32].

This work is an exposition of the formalism of the cluster expansion with a focus on the nuances of its application to ionic systems. Similarly, the work provides an overview of state-of-the-art methodology for constructing CE models of complex multicomponent ionic materials. Even in the wake of the explosion of machine learning interatomic potentials, the CE method remains a critical tool for the study of atomic configuration phenomena due its simplicity and amenability to MC sampling. Further development of advanced structure sampling and fitting techniques in the CE method beyond those discussed here is imperative to permit its successful use in the computational study of multicomponent ionic materials.

The source used to construct all cluster expansions is implemented in the Statistical Mechanics on Lattices (SMOL) package [110]. Implementations of all regularized regression models used are available at [111]. An implementation of the Bayesian charge assignment can be accessed at [112].

ACKNOWLEDGMENTS

This work was primarily funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sci-

ences, Materials Sciences and Engineering Division under Contract No. DE-AC02-05-CH11231 (Materials Project program KC23MP). L.B.L. and T.C. also gratefully acknowledge support from the National Science Foundation Graduate Research Fellowship under Grants No. DGE 1752814 and No. DGE 1106400, respectively. 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-AC02-05CH11231 using NERSC award BES-ERCAP0020531.

G.C. and L.B.L. conceived and supervised the project. L.B.L. worked out the CE formalism and its implications for charge-neutral ionic systems. L.B.L. computed all empirical pair potential calculations and all CE fits. L.B.L., P.Z., and F.X. calculated the structure sampling examples. J.Y. and P.Z. did the magnetic charge assignment calculations. J.Y. developed and implemented the Bayesian oxidation state assignment algorithm. T.C. and J.Y. developed and implemented the split anion sublattice and cation polyhedra structure matching algorithm. B.O. did the LMTOF DFT calculations. L.B.L. drafted the manuscript. All authors discussed, reviewed, and edited the manuscript.

-
- [1] A. Walle and G. Ceder, Automating first-principles phase diagram calculations, *J. Phase Equilib.* **23**, 348 (2002).
- [2] A. van de Walle, Multicomponent multisublattice alloys, nonconfigurational entropy and other additions to the Alloy Theoretic Automated Toolkit, *Calphad* **33**, 266 (2009).
- [3] A. Van der Ven, J. Thomas, B. Puchala, and A. Natarajan, First-principles statistical mechanics of multicomponent crystals, *Annu. Rev. Mater. Res.* **48**, 27 (2018).
- [4] J. M. Sanchez, F. Ducastelle, and D. Gratias, Generalized cluster description of multicomponent systems, *Physica A* **128**, 334 (1984).
- [5] J. M. Sanchez, Cluster expansions and the configurational energy of alloys, *Phys. Rev. B* **48**, 14013 (1993).
- [6] G. Ceder, G. D. Garbulsky, and P. D. Tapesch, Convergent real-space cluster expansion for configurational disorder in ionic systems, *Phys. Rev. B* **51**, 11257 (1995).
- [7] P. D. Tapesch, G. D. Garbulsky, and G. Ceder, Model for Configurational Thermodynamics in Ionic Systems, *Phys. Rev. Lett.* **74**, 2272 (1995).
- [8] C. Wolverton and A. Zunger, Cation and vacancy ordering in Li_xCoO_2 , *Phys. Rev. B* **57**, 2242 (1998).
- [9] G. Ceder, A. F. Kohan, M. K. Aydinol, P. D. Tapesch, and A. van der Ven, Thermodynamics of oxides with substitutional disorder: A microscopic model and evaluation of important energy contributions, *J. Am. Ceram. Soc.* **81**, 517 (1998).
- [10] C. Wolverton and A. Zunger, First-Principles Prediction of Vacancy Order-Disorder and Intercalation Battery Voltages in Li_xCoO_2 , *Phys. Rev. Lett.* **81**, 606 (1998).
- [11] P. D. Tapesch, A. F. Kohan, G. D. Garbulsky, G. Ceder, C. Coley, H. T. Stokes, L. L. Boyer, M. J. Mehl, B. P. Burton, K. Cho, and J. Joannopoulos, A Model to compute phase diagrams in oxides with empirical or first-principles energy methods and application to the solubility limits in the CaOMgO system, *J. Am. Ceram. Soc.* **79**, 2033 (1996).
- [12] J. Lee, A. Urban, X. Li, D. Su, G. Hautier, and G. Ceder, Unlocking the potential of cation-disordered oxides for rechargeable lithium batteries, *Science* **343**, 519 (2014).
- [13] R. J. Clément, Z. Lun, and G. Ceder, Cation-disordered rock-salt transition metal oxides and oxyfluorides for high energy lithium-ion cathodes, *Energy Environ. Sci.* **13**, 345 (2020).
- [14] H. Ji, J. Wu, Z. Cai, J. Liu, D.-H. Kwon, H. Kim, A. Urban, J. K. Papp, E. Foley, Y. Tian, M. Balasubramanian, H. Kim, R. J. Clément, B. D. McCloskey, W. Yang, and G. Ceder, Ultrahigh power and energy density in partially ordered lithium-ion cathode materials, *Nat. Energy* **5**, 213 (2020).
- [15] M. M. Thackeray, E. Lee, B. Shi, and J. R. Croy, Review— from LiMn_2O_4 to partially-disordered $\text{Li}_2\text{MnNiO}_4$: The evolution of lithiated-spinel cathodes for li-ion batteries, *J. Electrochem. Soc.* **169**, 020535 (2022).
- [16] S. Jiang, T. Hu, J. Gild, N. Zhou, J. Nie, M. Qin, T. Harrington, K. Vecchio, and J. Luo, A new class of high-entropy perovskite oxides, *Scr. Mater.* **142**, 116 (2018).
- [17] Y. Ma, Y. Ma, Q. Wang, S. Schweidler, M. Botros, T. Fu, H. Hahn, T. Brezesinski, and B. Breitung, High-entropy energy materials: Challenges and new opportunities, *Energy Environ. Sci.* **14**, 2883 (2021).
- [18] C. Oses, C. Toher, and S. Curtarolo, High-entropy ceramics, *Nat. Rev. Mater.* **5**, 295 (2020).
- [19] W. D. Richards, S. T. Dacek, D. A. Kitchaev, and G. Ceder, Fluorination of lithium-excess transition metal oxide cathode materials, *Adv. Energy Mater.* **8**, 1701533 (2018).
- [20] A. Urban, J. Lee, and G. Ceder, The configurational space of rocksalt-type oxides for high-capacity lithium battery electrodes, *Adv. Energy Mater.* **4**, 1400478 (2014).

- [21] B. Ouyang, N. Artrith, Z. Lun, Z. Jadidi, D. A. Kitchaev, H. Ji, A. Urban, and G. Ceder, Effect of fluorination on lithium transport and short-range order in disordered-rocksalt-type lithium-ion battery cathodes, *Adv. Energy Mater.* **10**, 1903240 (2020).
- [22] H. Ji, A. Urban, D. A. Kitchaev, D.-H. Kwon, N. Artrith, C. Ophus, W. Huang, Z. Cai, T. Shi, J. C. Kim, H. Kim, and G. Ceder, Hidden structural and chemical order controls lithium transport in cation-disordered oxides for rechargeable batteries, *Nat. Commun.* **10**, 592 (2019).
- [23] Z. Lun, B. Ouyang, D.-H. Kwon, Y. Ha, E. E. Foley, T.-Y. Huang, Z. Cai, H. Kim, M. Balasubramanian, Y. Sun, J. Huang, Y. Tian, H. Kim, B. D. McCloskey, W. Yang, R. J. Clément, H. Ji, and G. Ceder, Cation-disordered rocksalt-type high-entropy cathodes for Li-ion batteries, *Nat. Mater.* **20**, 214 (2021).
- [24] R. J. Clément, D. Kitchaev, J. Lee, and G. Ceder, Short-range order and unusual modes of nickel redox in a fluorine-substituted disordered rocksalt oxide lithium-ion cathode, *Chem. Mater.* **30**, 6945 (2018).
- [25] P. Zhong, Z. Cai, Y. Zhang, R. Giovine, B. Ouyang, G. Zeng, Y. Chen, R. Clément, Z. Lun, and G. Ceder, Increasing capacity in disordered rocksalt cathodes by Mg doping, *Chem. Mater.* **32**, 10728 (2020).
- [26] D. A. Kitchaev, Z. Lun, W. D. Richards, H. Ji, R. J. Clément, M. Balasubramanian, D.-H. Kwon, K. Dai, J. K. Papp, T. Lei, B. D. McCloskey, W. Yang, J. Lee, and G. Ceder, Design principles for high transition metal capacity in disordered rocksalt Li-ion cathodes, *Energy Environ. Sci.* **11**, 2159 (2018).
- [27] A. Seko, Y. Koyama, and I. Tanaka, Cluster expansion method for multicomponent systems based on optimal selection of structures for density-functional theory calculations, *Phys. Rev. B* **80**, 165122 (2009).
- [28] L. J. Nelson, G. L. W. Hart, F. Zhou, and V. Ozoliņš, Compressive sensing as a paradigm for building physics models, *Phys. Rev. B* **87**, 035125 (2013).
- [29] T. Mueller and G. Ceder, Bayesian approach to cluster expansions, *Phys. Rev. B* **80**, 024103 (2009).
- [30] Z. Leong and T. L. Tan, Robust cluster expansion of multicomponent systems using structured sparsity, *Phys. Rev. B* **100**, 134108 (2019).
- [31] P. Zhong, T. Chen, L. Barroso-Luque, F. Xie, and G. Ceder, An $\ell_0\ell_2$ -norm regularized regression model for construction of robust cluster expansions in multicomponent systems, *Phys. Rev. B* **106**, 024203 (2022).
- [32] J. H. Yang, T. Chen, L. Barroso-Luque, Z. Jadidi, and G. Ceder, Approaches for handling high-dimensional cluster expansions of ionic systems, *npj Comput. Mater.* **8**, 133 (2022).
- [33] A. van de Walle, A complete representation of structure–property relationships in crystals, *Nat. Mater.* **7**, 455 (2008).
- [34] R. Drautz and M. Fähnle, Spin-cluster expansion: Parametrization of the general adiabatic magnetic energy surface with *ab initio* accuracy, *Phys. Rev. B* **69**, 104404 (2004).
- [35] R. Drautz, Atomic cluster expansion for accurate and transferable interatomic potentials, *Phys. Rev. B* **99**, 014104 (2019).
- [36] This can be generalized to other spaces; for example, Ω_i can be the vector space over \mathbb{R}^3 when representing magnetic spins or atomic positions [34,35].
- [37] Our use of the term *sublattice* is descriptivist and does not follow the rigorous use of the word in crystallography.
- [38] J. C. Thomas and A. Van der Ven, Finite-temperature properties of strongly anharmonic and mechanically unstable crystal phases from first principles, *Phys. Rev. B* **88**, 214111 (2013).
- [39] R. Drautz, Atomic cluster expansion of scalar, vectorial, and tensorial properties including magnetism and charge transfer, *Phys. Rev. B* **102**, 024104 (2020).
- [40] T. Ceccherini-Silberstein, F. Scarabotti, and F. Tolli, *Discrete Harmonic Analysis: Representations, Number Theory, Expanders, and the Fourier Transform*, Cambridge Studies in Advanced Mathematics (Cambridge University Press, Cambridge, UK, 2018).
- [41] X. Zhang and M. H. F. Sluiter, Cluster expansions for thermodynamics and kinetics of multicomponent alloys, *J. Phase Equilib. Diffus.* **37**, 44 (2016).
- [42] J. M. Sanchez, Foundations and practical implementations of the cluster expansion, *J. Phase Equilib. Diffus.* **38**, 238 (2017).
- [43] M. D. Ado Jorio and M. S. Dresselhaus, Basis functions, in *Group Theory* (Springer, Berlin, 2008), pp. 57–75.
- [44] J. C. Thomas, J. S. Bechtel, and A. Van der Ven, Hamiltonians and order parameters for crystals of orientable molecules, *Phys. Rev. B* **98**, 094105 (2018).
- [45] A. Seko and I. Tanaka, Cluster expansion of multicomponent ionic systems with controlled accuracy: Importance of long-range interactions in heterovalent ionic systems, *J. Phys.: Condens. Matter* **26**, 115403 (2014).
- [46] A. Seko, K. Yuge, F. Oba, A. Kuwabara, I. Tanaka, and T. Yamamoto, First-principles study of cation disordering in MgAl_2O_4 spinel with cluster expansion and Monte Carlo simulation, *Phys. Rev. B* **73**, 094116 (2006).
- [47] A. Seko, A. Togo, F. Oba, and I. Tanaka, Structure and Stability of a Homologous Series of Tin Oxides, *Phys. Rev. Lett.* **100**, 045702 (2008).
- [48] E. Lee, F. B. Prinz, and W. Cai, Enhancing ionic conductivity of bulk single-crystal yttria-stabilized zirconia by tailoring dopant distribution, *Phys. Rev. B* **83**, 052301 (2011).
- [49] E. Lee and K. A. Persson, Revealing the coupled cation interactions behind the electrochemical profile of $\text{Li}_x\text{Ni}_{0.5}\text{Mn}_{1.5}\text{O}_4$, *Energy Environ. Sci.* **5**, 6047 (2012).
- [50] D. Wang, L.-M. Liu, S.-J. Zhao, B.-H. Li, H. Liu, and X.-F. Lang, $\beta\text{-MnO}_2$ as a cathode material for lithium ion batteries from first principles calculations, *Phys. Chem. Chem. Phys.* **15**, 9075 (2013).
- [51] A. van de Walle and D. E. Ellis, First-Principles Thermodynamics of Coherent Interfaces in Samarium-Doped Ceria Nanoscale Superlattices, *Phys. Rev. Lett.* **98**, 266101 (2007).
- [52] W. D. Richards, Y. Wang, L. J. Miara, J. C. Kim, and G. Ceder, Design of $\text{Li}_{1+2x}\text{Zn}_{1-x}\text{PS}_4$, a new lithium ion conductor, *Energy Environ. Sci.* **9**, 3272 (2016).
- [53] A. Y. Toukmaji and J. A. Board, Ewald summation techniques in perspective: A survey, *Comput. Phys. Commun.* **95**, 73 (1996).
- [54] L. Greengard and V. Rokhlin, A fast algorithm for particle simulations, *J. Comput. Phys.* **135**, 280 (1997).
- [55] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevB.106.144202> for details of the parametrization of the Buckingham-Coulomb potential,

- density functional theory calculations, cluster expansion fits, and supplemental sampling results, which include Refs. [56–65].
- [56] A. Jain, G. Hautier, S. P. Ong, C. J. Moore, C. C. Fischer, K. A. Persson, and G. Ceder, Formation enthalpies by mixing GGA and GGA + U calculations, *Phys. Rev. B* **84**, 045115 (2011).
- [57] G. Kresse and J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, *Comput. Mater. Sci.* **6**, 15 (1996).
- [58] G. Kresse and D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, *Phys. Rev. B* **59**, 1758 (1999).
- [59] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized Gradient Approximation Made Simple, *Phys. Rev. Lett.* **77**, 3865 (1996).
- [60] M. Wang and G.-L. Tian, Adaptive group Lasso for high-dimensional generalized linear models, *Stat. Papers* **60**, 1469 (2019).
- [61] L. Wang, T. Maxisch, and G. Ceder, Oxidation energies of transition metal oxides within the GGA + U framework, *Phys. Rev. B* **73**, 195107 (2006).
- [62] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton, LAMMPS—a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales, *Comput. Phys. Commun.* **271**, 108171 (2022).
- [63] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization, *ACM Trans. Math. Softw.* **23**, 550 (1997).
- [64] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, A limited memory algorithm for bound constrained optimization, *SIAM J. Sci. Comput.* **16**, 1190 (1995).
- [65] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey *et al.*, SciPy 1.0: Fundamental algorithms for scientific computing in PYTHON, *Nat. Methods* **17**, 261 (2020).
- [66] 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).
- [67] L. J. Nelson, V. Ozoliņš, C. S. Reese, F. Zhou, and G. L. W. Hart, Cluster expansion made easy with Bayesian compressive sensing, *Phys. Rev. B* **88**, 155105 (2013).
- [68] W. Huang, A. Urban, P. Xiao, Z. Rong, H. Das, T. Chen, N. Artrith, A. Toumar, and G. Ceder, An L_0L_1 -norm compressive sensing paradigm for the construction of sparse predictive lattice models using mixed integer quadratic programming, [arXiv:1807.10753](https://arxiv.org/abs/1807.10753).
- [69] A. Seko and I. Tanaka, Grouping of structures for cluster expansion of multicomponent systems with controlled accuracy, *Phys. Rev. B* **83**, 224111 (2011).
- [70] T. Mueller and G. Ceder, Exact expressions for structure selection in cluster expansions, *Phys. Rev. B* **82**, 184107 (2010).
- [71] R. J. Tibshirani, The lasso problem and uniqueness, *Electron. J. Statist.* **7**, 1456 (2013).
- [72] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations* (Chapman and Hall/CRC, New York, 2015).
- [73] J. J. Faraway, *Linear Models with Python* (CRC Press, Boca Raton, FL, 2021).
- [74] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff, Fast approximation of matrix coherence and statistical leverage, *J. Mach. Learn. Res.* **13**, 3475 (2012).
- [75] L. Barroso-Luque, J. H. Yang, and G. Ceder, Sparse expansions of multicomponent oxide configuration energy using coherency and redundancy, *Phys. Rev. B* **104**, 224203 (2021).
- [76] E. J. Candes and M. B. Wakin, An introduction to compressive sampling, *IEEE Signal Process. Mag.* **25**, 21 (2008).
- [77] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall, Compressed sensing with coherent and redundant dictionaries, *Appl. Comput. Harmonic Anal.* **31**, 59 (2011).
- [78] N. Simon and R. Tibshirani, Standardization and the group lasso penalty, *Stat. Sinica* **22**, 983 (2012).
- [79] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, A sparse-group lasso, *J. Comput. Graphical Stat.* **22**, 231 (2013).
- [80] G. Ceder, Y.-M. Chiang, D. R. Sadoway, M. K. Aydinol, Y.-I. Jang, and B. Huang, Identification of cathode materials for lithium batteries guided by first-principles calculations, *Nature (London)* **392**, 694 (1998).
- [81] J. Snoek, H. Larochelle, and R. P. Adams, Practical Bayesian optimization of machine learning algorithms, *Adv. Neural Inf. Process. Syst.* **25** (2012).
- [82] J. Sun, A. Ruzsinszky, and J. P. Perdew, Strongly Constrained and Appropriately Normed Semilocal Density Functional, *Phys. Rev. Lett.* **115**, 036402 (2015).
- [83] T. Chen, J. H. Yang, L. Barroso-Luque, and G. Ceder, Removing the two-phase transition in spinel LiMn_2O_4 through cation disorder (unpublished).
- [84] Z. Jadidi, J. H. Yang, T. Chen, L. Barroso-Luque, and G. Ceder, Ab-initio study of short-range-ordering in vanadium-based disordered rocksalt structures (unpublished).
- [85] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Python Materials Genomics (pymatgen): A robust, open-source PYTHON library for materials analysis, *Comput. Mater. Sci.* **68**, 314 (2013).
- [86] C. Su, J. Lv, Q. Li, H. Wang, L. Zhang, Y. Wang, and Y. Ma, Construction of crystal structure prototype database: Methods and applications, *J. Phys.: Condens. Matter* **29**, 165901 (2017).
- [87] D. Hicks, C. Toher, D. C. Ford, F. Rose, C. D. Santo, O. Levy, M. J. Mehl, and S. Curtarolo, AFLOW-XtalFinder: A reliable choice to identify crystalline prototypes, *npj Comput. Mater.* **7**, 30 (2021).
- [88] J. C. Thomas, A. R. Natarajan, and A. Van der Ven, Comparing crystal structures with symmetry and geometry, *npj Comput. Mater.* **7**, 164 (2021).
- [89] A. H. Nguyen, C. W. Rosenbrock, C. S. Reese, and G. L. W. Hart, Robustness of the cluster expansion: Assessing the roles

- of relaxation and numerical error, *Phys. Rev. B* **96**, 014107 (2017).
- [90] T. Hastie, R. Tibshirani, and J. Friedman, Linear methods for classification, in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, edited by T. Hastie, R. Tibshirani, and J. Friedman (Springer, New York, 2009), pp. 101–137.
- [91] G. L. W. Hart, V. Blum, M. J. Walorski, and A. Zunger, Evolutionary approach for determining first-principles Hamiltonians, *Nat. Mater.* **4**, 391 (2005).
- [92] T. Hastie, R. Tibshirani, and J. Friedman, Model assessment and selection, in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, edited by T. Hastie, R. Tibshirani, and J. Friedman (Springer, New York, 2009), pp. 219–259.
- [93] 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).
- [94] W. Huang, A. Urban, Z. Rong, Z. Ding, C. Luo, and G. Ceder, Construction of ground-state preserving sparse lattice models for predictive materials simulations, *npj Comput. Mater.* **3**, 1 (2017).
- [95] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, Optimization with sparsity-inducing penalties, *Found. Trends Mach. Learn.* **4**, 1 (2012).
- [96] M. Elad, Uniqueness and uncertainty, in *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, edited by M. Elad (Springer, New York, 2010), pp. 17–33.
- [97] It is not formally a norm, since it does not satisfy all mathematical requirements for a norm; most notably it is not sensitive to scale: $\|k\mathbf{J}\|_0 = \|\mathbf{J}\|_0$ for any scalar k .
- [98] M. Elad, Pursuit algorithms—practice, in *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, edited by M. Elad (Springer, New York, 2010), pp. 35–54.
- [99] M. Elad, From exact to approximate solutions, in *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, edited by M. Elad (Springer, New York, 2010), pp. 79–109.
- [100] R. Tibshirani, Regression shrinkage and selection via the Lasso, *J. R. Stat. Soc. Ser. B Methodol.* **58**, 267 (1996).
- [101] H. Zou and T. Hastie, Regularization and variable selection via the elastic net, *J. R. Stat. Soc. B* **67**, 301 (2005).
- [102] M. Yuan and Y. Lin, Model selection and estimation in regression with grouped variables, *J. R. Stat. Soc. B* **68**, 49 (2006).
- [103] J. Friedman, T. Hastie, and R. Tibshirani, A note on the group lasso and a sparse group lasso, [arXiv:1001.0736](https://arxiv.org/abs/1001.0736).
- [104] H. Zou, The adaptive lasso and its oracle properties, *J. Am. Stat. Assoc.* **101**, 1418 (2006).
- [105] E. J. Candès, M. B. Wakin, and S. P. Boyd, Enhancing sparsity by reweighted ℓ_1 minimization, *J. Fourier Anal. Appl.* **14**, 877 (2008).
- [106] H. Wang and C. Leng, A note on adaptive group lasso, *Comput. Stat. Data Anal.* **52**, 5277 (2008).
- [107] N. A. Zarkevich and D. D. Johnson, Reliable First-Principles Alloy Thermodynamics via Truncated Cluster Expansions, *Phys. Rev. Lett.* **92**, 255702 (2004).
- [108] For binary cluster expansions there is no distinction between the two forms of hierarchical constraints described, since there is only one correlation function associated with each orbit of site clusters.
- [109] L. Jacob, G. Obozinski, and J.-P. Vert, Group lasso with overlap and graph lasso, in *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09* (ACM Press, Montreal, Quebec, Canada, 2009), pp. 1–8.
- [110] <https://github.com/CederGroupHub/smol>.
- [111] <https://github.com/CederGroupHub/sparse-lm>.
- [112] <https://github.com/juliayang/high-component-ce-tools>.