UC Berkeley
UC Berkeley Previously Published Works

Title

Sparse expansions of multicomponent oxide configuration energy using coherency and redundancy

Permalink

https://escholarship.org/uc/item/7s29k3z6

Authors

Barroso-Luque, Luis
Yang, Julia H
Ceder, Gerbrand

Publication Date

2021-12-01

DOI

10.1103/physrevb.104.224203

Sparse expansions of multicomponent oxide configuration energy using coherency & redundancy

Luis Barroso-Luque[*] and Julia H. Yang
Department of Materials Science and Engineering,
University of California Berkeley, Berkeley CA, USA

Gerbrand Ceder[†]
Department of Materials Science and Engineering,
University of California Berkeley, Berkeley CA, USA and
Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley CA, USA
(Dated: December 17, 2021)

Compressed sensing has become a widely accepted paradigm to construct high dimensional cluster expansion models used for statistical mechanical studies of atomic configuration in complex multicomponent crystalline materials. However, strict sampling requirements necessary to obtain minimal coherence measurements for compressed sensing to guarantee accurate estimation of model parameters are difficult and in some cases impossible to satisfy due to the inability of physical systems to access certain configurations. Nevertheless, the dependence of energy on atomic configuration can still be adequately learned without these strict requirements by using compressed sensing by way of coherent measurements using redundant function sets known as frames. We develop a particular frame constructed from the union of all occupancy-based cluster expansion basis sets. We illustrate how using this highly redundant frame yields sparse expansions of the configuration energy of complex oxide materials that are competitive and often surpass the prediction accuracy and sparsity of models obtained from standard cluster expansions.

I. INTRODUCTION

The cluster expansion method (CEM)[1–3] has become a standard tool in computational thermodynamics of multicomponent crystalline systems as an effective and efficient means to compute functions of atomic configuration. The CEM is used to represent a coarse-grained lattice model of materials properties as a function of the possible atomic configurations, i.e., the occupations of crystallographic sites. The most common use of the CEM involves the expansion of the formation energy of a particular multicomponent material system. A cluster expansion is fitted using formation energies for a set of representative structures calculated from first principles electronic structure methods, most often calculated with density functional theory (DFT). Well-converged cluster expansions fitted to DFT energies have been shown to give very accurate predictions of formation energy to within a few meV/atom.[3–5] A converged cluster expansion can subsequently be used in Monte Carlo simulations to compute thermodynamic properties, such as configuration entropy,[6, 7] phase diagrams,[8] ionic percolation,[9, 10] and short range order.[11, 12]

The CEM involves the use of a discrete harmonic expansion in terms of basis functions—known by practitioners as correlation functions—that act on groups of symmetrically related clusters of crystallographic sites. The correlation functions used in the CEM are constructed

by taking symmetry adapted averages of all possible N-*fold* products of site basis functions for all N sites in a crystal structure. Site basis functions operate over a domain associated with each site in a structure, which we refer to as a site space. The site functions must span the site space associated with each given site. In the CEM the site functions constitute a basis for the space of functions acting over the allowed occupations at a site, such that the number of site basis functions must be equal to the number of allowed species (i.e. the size of the site space). Additionally, in order for the resulting product basis to have a cluster-like framework, where basis functions operate on different clusters of sites, each set of site basis functions must include a constant function $\phi_0 \equiv 1$ as one of the functions used to span the corresponding occupations for each site.[3]

Estimating the correlation function coefficients, known as effective cluster interactions (ECI), for a specific expansion is most often done using linear regression.[3, 5, 13] The values of a predefined finite set of correlation functions—chosen within a spatial cutoff range and maximum cluster size—are computed for each training structure to form a truncated correlation matrix $\Pi_S = A\Pi$, where Π is the full system (infinite for bulk systems) correlation matrix, and A is a *sensing* matrix that selects the set of correlation functions to be considered in the fit, and also selects the values of those correlation functions for each structure included in the set of training structures. A column of Π_S is made up of the values of an included correlation function evaluated for each of the training structures, and are correspondingly feature vectors. A row of Π_S is called a correlation vector and its entries correspond to each of the included correla-

---
[*] lbluque@berkeley.edu
[†] gceder@berkeley.edu

tion functions evaluated for a particular structure in the training set.

Several linear regression algorithms for estimating ECI have been proposed and benchmarked in literature.[4, 14–18] Although most of the proposed algorithms can be used in either overdetermined and underdetermined linear systems, the formal analysis of solutions and mathematical guarantees on the accuracy of coefficients can be very different and are for the most part carried out for each type of linear system separately.[19, 20] In over-determined systems more training structures are included than the number of correlation functions used in the fit. The motivation for using an over-determined system comes from several studies showing decreasing cross validation errors with increasing number of training structures.[18, 21] In the case of underdetermined systems, more correlation functions are included in the model than structures used when fitting; which is usually motivated for the CEM as a way to avoid a near-sighted pre-selection of correlation functions. In the current work, we focus on the case of using underdetermined linear systems using ℓ_1 regularization as a convex relaxation of the ℓ_0 pseudo-norm, in order to fit sparse expansions of configuration energy. This approach can be formally cast and analyzed within the framework of Compressed Sensing (CS).[22, 23]

For relatively small dimensional systems, such as binary or ternary alloys, that allow accurate fits using only a small number of short range clusters of low degree (number of sites in cluster) such that the number of total correlation functions is manageable, practitioners have the option to choose between using an overdetermined system or an underdetermined system approach. However, for more complex high dimensional systems there will often be no such choice. Since the number of expansion terms that need to be considered for an acceptable fit quickly becomes too large, computing DFT energies for enough structures to obtain an overdetermined system becomes untenable. The reason for this is twofold. First of all, the number of terms in a truncated expansion grows polynomially with the number of allowed species at each site. Additionally, complex physical interactions may require longer range and/or higher degree clusters. For example, these situations frequently occur in the study of high entropy materials which has recently gained much attention in the design of metallic alloys,[24] and ceramic materials.[25] Although theoretically the CEM is well suited for the study of high entropy materials, the large high dimensional configuration spaces involved render an *over-determined* system approach prohibitive. For such scenarios, there is no choice but to work with underdetermined systems only, and hope that *betting on sparsity*[20] applies favorably to high dimensional cluster expansions for the increasing number of materials being studied that require fits using severely underdetermined systems.

Fortunately, classical compressed sensing[22] in the CEM has already been demonstrated to work well with some metallic alloy systems.[17, 26] Here we go beyond previous work and apply the theory of compressed sensing with coherency and redundancy to the realm of the CEM, and demonstrate how it applies in much larger configuration spaces for ionic systems with multiple substructures and oxidation state assignment. In Section II we introduce the main concepts of compressed sensing with coherence and redundancy and restate the main signal recovery error bound[23] adapted to the domain of cluster expansions. In Section III we lay out the details of cluster expansions using site indicator basis functions. This is subsequently used as a building block in the construction of a redundant function set, which we call the *generalized Potts frame*, which we describe in Section IV. Finally, we demonstrate how the generalized Potts frame yields sparse and accurate expansions by comparing resulting fits of three different fluorinated lithium-transition metal oxide materials with standard cluster expansions using site indicator functions and orthogonal trigonometric/sinusoid basis functions.[27]

II. COMPRESSED SENSING WITH COHERENCY & REDUNDANCY

The notion of classical compressed sensing (CS) that has been previously shown to yield accurate cluster expansions of metallic alloys,[17, 26] relies strictly on the concept of *incoherence*[22] in order to guarantee accurate recovery of the underlying coefficients/ECI. The need for incoherent measurements—which in the CEM correspond to correlation function values evaluated for selected training structures—is clear when the goal is to accurately recover the exact coefficients in the expansion of a function. This requirement can be made intuitive by thinking of *coherence* as a measurement of how similar correlation function samples (i.e., columns of Π_S) are to each other. High *incoherence* (low *coherence*), means that the set of correlation function samples used are more uniformly spread out in their span (they are closer to mutually orthogonal), and as a result the portion of the energy represented by each, can be easily distinguished. In the limit of zero coherence, the correlation matrix is orthogonal, and in principle the portion of the energy for each correlation function can be identified exactly. For the cases with high coherence, correlation function samples will be much more closely correlated and it is no longer possible to distinguish which correlation function a specific portion of the energy comes from.

Formally, the *coherence* of a measurement matrix M—which in the CEM corresponds the truncated correlation matrix Π_S—is defined as,[28]

$$\mu(M) = \max_{i<j} \frac{|\langle M_i, M_j \rangle|}{||M_i||_2 ||M_j||_2} \tag{1}$$

where M_i and M_j are columns of matrix M. The normalized definition given above bounds the coherence of

a measurement matrix to values between zero and one, $0 \leq \mu(M) \leq 1$.

In the case of the CEM, the measurement matrix $M = A\Pi$ is made up of the values of a relatively small set of truncated correlation vectors for each of the training structures for which DFT energies have been calculated, and the truncated correlation vectors only include functions associated with clusters within a physical predefined cutoff radius and cluster size.

The guarantees of classical compressed sensing depend on the coherence of the measurement matrix M being as close to zero as possible in order to maximize the probability of accurate reconstruction of coefficients/ECI. As suggested above, incoherent measurements improve the chances of accurate recovery of coefficients/ECI, specifically by requiring a lower number of necessary training structures for a given level of accuracy.[22] Furthermore, it is necessary for M to satisfy the *restricted isometry property* (RIP) with a small isometry constant δ.[22] In broad terms, the RIP is a condition which ensures that most of the possible sparse solutions for the linear system lie outside the nullspace of the measurement matrix M.[28]

It has been shown that matrices made up of random measurements, such as those composed of Gaussian vectors on the unit hypersphere, satisfy the RIP with overwhelming probability and lead to high levels of incoherence.[22] Random measurements have been previously reported to lead to effective training structure selection and resulting accurate underdetermined fits of CE's for binary metallic alloys.[17, 26] Nevertheless, selecting training structures such that the resultant measurement matrix has high incoherence and ideally satisfies the RIP with a small constant, becomes difficult and in some cases almost impossible for materials systems with more complex physics. This can be generally understood as a result of physically imposed sampling constraints. Usually the vast majority of all possible configurations will have high energies that are complicated if not impossible to compute with first principles methods. For example, systems with configurations that undergo large structural relaxations can no longer be mapped back to the fixed structure underlying the CE. As another example, certain cluster occupations in ionic systems are difficult to access when very high electrostatic repulsion exists. Finally, charge neutrality constraints in heterovalent ionic systems restricts the possible configurations that can be sampled. These phenomena complicate and in many cases prevent the possibility of obtaining appropriate correlation matrices with minimal coherence required for classical CS recovery, even with previously proposed methods, such as the aforementioned use of uniformly random vectors over the hypersphere.

Fortunately, there is a variant of compressed sensing that has been shown to give accurate reconstructions of sparse or compressible signals with only a small set of *coherent* measurements by including redundancy in these measurements.[23] We will show how this concept

can be used for learning sparse expansions of functions of configuration in multicomponent crystalline materials. The essence of CS with coherent and redundant measurements is not to recover the model coefficients as accurately as possible, but rather to recover an approximation for the actual function as accurately as possible by way of a redundant representation and possibly highly coherent measurements.[23] For systems constructed using measurements with high coherence and a redundant representation, solutions will always be highly degenerate, but again the focus is on the accuracy of predictions obtained using the recovered expansion and not the exact values of the fitted coefficients. However, we will still seek only sparse solutions such that the large degeneracy of solutions becomes manageable and the resulting expansion can be used efficiently for subsequent predictions and simulations such as in Monte Carlo studies. Figure 1 shows a schematic of the pertinent mathematical objects, the corresponding domains and their relationships as it pertains to CS. For our purposes we seek an accurate representation in the function domain made up from a small number of coefficients but without much regard to whether or not the obtained coefficients correspond to those of any underlying *exact* expansion.

In what follows we only give a brief overview of CS with coherent and redundant measurements adapted to our purposes. Further details and accompanying proofs can be found in the original publication.[23]

The goal of CS, is to reconstruct a sparse representation or approximation of a function by solving the following optimization problem.

$$\hat{H} = \underset{\widetilde{H}}{\mathrm{argmin}} ||\Pi^* \widetilde{H}||_1 \text{ subject to } ||A\widetilde{H} - E||_2 \leq \varepsilon \quad (2)$$

where H is the function sought—in our case a function of atomic configuration. Π is a linear operator mapping coefficients J to the function H as depicted in Figure 1. In the CEM, Π is the matrix of all correlation basis functions. A is a sensing matrix used in selecting the training data; and E is the measured value of the function H for the training configurations. The measurements are of the form $E = AH + z$, where z represents an additive noise with some upper bound $||z||_2 \leq \varepsilon$.

The main theorem of CS with coherency and redundancy gives the following error bounds for an s-sparse reconstruction of H,

$$||\hat{H} - H||_2 \leq C_0 \varepsilon + C_1 \frac{||\Pi^* H - (\Pi^* H)_s||_1}{\sqrt{s}} \quad (3)$$

where $\hat{H}$ is the s-sparse approximation to H. $(\Pi^* H)_s$ denotes a vector with the largest s nonzero coefficients of H and zeros elsewhere. C_0 and C_1 are constants that depend only on the sensing matrix A. The theorem holds if the measurement matrix satisfies a restricted isometry property (D-RIP) adapted to the union of the span of all sets of s columns of Π with isometry constant $\delta_{2d} < 0.08$.[23] The D-RIP prevents possible solutions
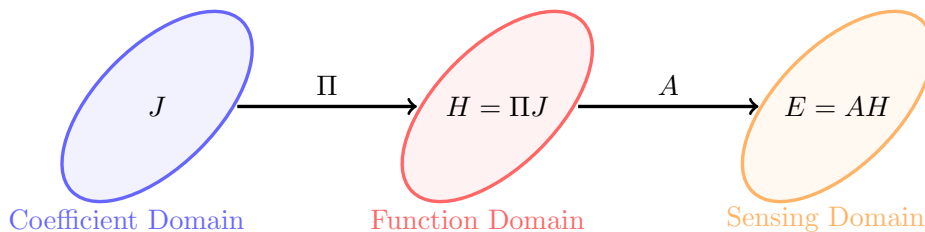
FIG. 1: Schematic of different domains involved in compressed sensing. CS with redundancy seeks to recover function H in the function domain in the center. Adapted from Candes *et al.*[23]

from being highly distorted by the measurement matrix, and similarly to the standard RIP, also prevents them from falling in its null space.

The theorem from Equation 3 essentially says that the solution to the optimization problem in Equation 2 gives accurate reconstructions of the function H, when the coefficients $\Pi^* H$ are sparse and/or decay rapidly—i.e., H is *compressible*. These results suggest that if functions of configuration for multicomponent materials are indeed *compressible*, meaning they can be represented with only a small set of functions, then accurate reconstructions can be obtained with coherent and redundant measurements. Although a rigorous proof along with bounds on the compressibility of functions of configuration—which to the best of our knowledge has not been reported in literature—would be of immense value, it is beyond the scope of the present work. We do however provide numerical indication that such functions are indeed favorably compressible. This result is not unanticipated considering established knowledge regarding the physics of locality, the success of atomic potentials with small numbers of multiple body terms, and the general tenet of parsimony in physics. This translates to our intuition that such expansions of atomic configuration should have low degree and small associated cluster diameters, and as such we should be able to represent them using a small number of terms. The crux is finding an optimal subspace spanned by a small set of functions that allows an accurate and sparse representation.

III. THE CLUSTER EXPANSION WITH SITE OCCUPANCY BASIS FUNCTIONS

Before introducing the generalized Potts frame, we give a brief exposition of the process of constructing a basis in the CEM using a particular choice of site basis functions. Although the original CEM was developed using *orthonormal* site basis functions constructed from polynomials,[1, 2] any set of functions can be used as long as they properly span the function space over the corresponding site spaces. Here we will focus on one commonly used basis set composed of site *occupancy* functions,[29] which we refer to as site *indicator* functions. A site *indicator* function simply *indicates* whether

a given species occupies a site or not,

$$\mathbf{1}_{\sigma_j}(\sigma_i) = \begin{cases} 1 & \text{if } \sigma_i = \sigma_j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Since the set of basis functions must include the constant function for the CEM to yield basis functions that follow a cluster framework, the constant function needs to replace an indicator function for one of the species at each site. In other words, one of the total n allowed species at each site will not have an associated indicator function. [29] The sets of site basis functions $\{\phi_j; j = 0, \ldots, n-1\}$ for the CEM in terms of site indicator functions are given by,

$$\phi_j(\sigma_i) = \begin{cases} 1 & \text{if } j = 0 \\ \mathbf{1}_{\sigma_j}(\sigma_i) & \text{if } j = 1, \ldots, n-1 \end{cases} \tag{5}$$

The correlation functions are then constructed from symmetry adapted averages of the following N-fold products of site basis functions for each site in a structure,

$$\Phi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) = \prod_{i=1}^{N} \phi_{\alpha_i}(\sigma_i), \tag{6}$$

where $\boldsymbol{\alpha}$ is a multi-index used to specify the particular site function, $\alpha_i = 0, \ldots, n_i - 1$, for each site in the structure. Since we can ignore all sites where the constant site basis function $\phi_0 \equiv 1$ is used, Equation 6, can be written in terms of site indicator functions as,

$$\Phi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) = \prod_{i \text{ s.t. } \alpha_i \neq 0} \mathbf{1}_{\sigma_{\alpha_i}}(\sigma_i), \tag{7}$$

where we notice that by construction, the functions in Equation 7 will indicate whether a specific occupancy of a given cluster represented by the nonzero elements of the multi-index $\boldsymbol{\alpha}$ is present in a structure. We can more briefly write Equation 7 as a *cluster* indicator function,

$$\begin{aligned} \Phi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) &= \mathbf{1}_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) \\ &= \begin{cases} 1 & \text{if cluster } \boldsymbol{\alpha} \text{ is in } \boldsymbol{\sigma} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{8}$$

Furthermore, since the final correlation functions are constructed from averages of functions given by Equation 8 over symmetrically equivalent clusters, the correlation functions represent (up to multiplicity constants) the *concentration* of a specific occupancy of clusters for the different crystallographic orbits,

$$\langle \Phi(\boldsymbol{\sigma}) \rangle_\beta = \frac{1}{N(\beta)} \sum_{\boldsymbol{\alpha} \in \beta} \mathbf{1}_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) \qquad (9)$$

where β represents an orbit of labeled clusters (i.e. with each site labeled with a particular indicator function), and $N(\beta)$ is the total number of labeled clusters in the orbit β. The sum is carried over all labeled clusters $\boldsymbol{\alpha}$ that are part of the orbit β.

We make two notable observations regarding CEM correlation functions with site indicator basis functions. The first is that the basis sets in Equation 5 are not orthogonal and as a result the resulting correlation functions in Equation 9 are not orthogonal either. This lack of orthogonality can further complicate constructing highly incoherent measurement matrices compared to using orthogonal/orthonormal basis sets.

The second observation is concerned with the set of clusters/orbits that are selected by the correlation functions in Equation 9. The set of correlation functions in Equation 9 indeed represents a basis for the function space over all possible configurations $\boldsymbol{\sigma}$ and thus the set is linearly independent.[29] As a result however, the included functions do not give the concentrations of all possible occupied clusters. Namely, the correlation functions in Equation 9 never include any occupied clusters that involve the species that do not have an associated indicator function in the corresponding site basis functions. Additionally, the number of clusters not indicated for in the CEM basis grows quickly with the number of components. Namely the number grows as $O\left(n^{N_{\boldsymbol{\alpha}}} - (n-1)^{N_{\boldsymbol{\alpha}}}\right)$, where n is the number of allowed species at a site and $N_{\boldsymbol{\alpha}}$ is the number of sites in a cluster.

Notwithstanding these observations, cluster expansions using site indicator functions formally constitute a basis for function spaces over crystalline configurations, and have been successfully used in the study of configurational thermodynamics for some time.[8, 10, 29] However, as alluded to before, the lack of orthogonality complicates obtaining incoherent measurements to maximize accurate ECI estimation from classical CS. Additionally, the choice of species left out in site basis sets is mathematically meaningless, but can often lead to precarious interpretations of fitted ECI. This begs the question of whether we can do away with using CEM basis sets and seeking maximally incoherent measurements, yet still obtain suitably sparse and accurate expansions by instead relying on redundancy by including functions labeled for all possible occupations over the included clusters.

IV. THE GENERALIZED POTTS FRAME AS A REDUNDANT REPRESENTATION

We introduce a specific redundant set of functions that spans the same function space as correlation basis sets used in the CEM. The redundant set of functions can be obtained simply by including cluster indicator functions of the form in Equation 7 for all possible occupations. More formally we build all the N-fold product functions from redundant site function sets that include $\phi_0 \equiv 1$ and site indicator functions for *all* allowed species at each site, such that for each site with n allowed species, we associate a redundant set of functions given by,

$$\phi_j(\sigma_i) = \begin{cases} 1 & \text{if } j = 0 \\ \mathbf{1}_{\sigma_j}(\sigma_i) & \text{if } j = 1, \dots, n \end{cases} \qquad (10)$$

This results in $n + 1$ total functions for a space of dimension n, where the redundancy is the trivial linear relation,

$$\phi_0(\sigma) - \sum_{i=1}^{n} \mathbf{1}_{\sigma_i}(\sigma) = 0 \qquad (11)$$

Carrying out the same operation of taking all possible N-fold products of functions from each redundant set and taking symmetry adapted averages over crystallographic orbits results in the set of symmetrized product functions which give *concentrations* for any possible cluster occupancy. This resulting set is clearly highly redundant since the total number of functions obtained are of order $O\left((n+1)^N\right)$ and the dimension of the function space is of order $O(n^N)$. In other words the combinatorics[30] for identifying symmetrically equivalent clusters involve n possible *labels* of basis functions for each site for all clusters in distinct orbits compared with the $n-1$ *labels* involved in a standard CEM basis. Nonetheless, the resulting set of functions spans the same function space as a standard CEM basis and therefore formally the set constitutes a *frame*.[31, 32] We provide a derivation for a set of frame bounds in Appendix A, but make no effort to optimize them. The resulting frame evidently has a strong connection to the well known Potts model.[33] Indeed it is a (normalized) generalization in both spatial extent and interaction size of the original nearest neighbor pair Potts model.[34] Hence we refer to the proposed frame as the *generalized Potts frame*.

The generalized Potts frame can also be seen as the union of every possible CEM correlation basis of site indicator functions. All possible CEM correlation bases can be generated by cycling over the species that is not *indicated* for in its corresponding site basis and building the corresponding CEM correlation basis for every possible combination of site basis sets. We make a special note that apart from the standard CEM basis sets, this includes correlations of *mixed* products where symmetrically equivalent sites in the underlying random structure can have distinct basis sets, i.e. a different subset of
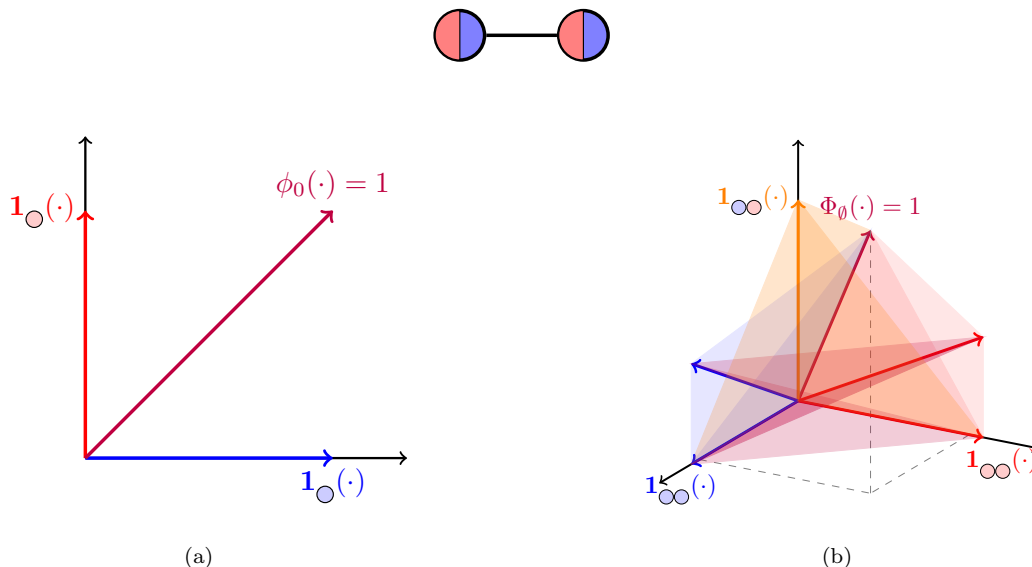
FIG. 2: Function space representations over the configurations of a single binary site and a symmetric binary diatomic molecule. (a) Function space over a single binary site space. The two different choices for site bases to construct a standard CE are colored red and blue, and each set also includes the purple $\phi_0 \equiv 1$ function. (b) Function space over symmetrically distinct configurations of the molecule. A CE basis includes either the blue colored or the red colored functions. The generalized Potts frame includes all colored functions (blue/red/yellow). All functions sets also include the magenta colored constant function. The D-RIP for a 2-sparse representation in this case is adapted to the union of all colored planes in (b).

species with associated indicator functions. Symbolically, the generalized Potts frame is obtained after taking the corresponding symmetry adapted averages for the product functions in the following set,

$$\bigcup_{\sigma} \{\Phi_{\boldsymbol{\alpha}}; \ \forall \boldsymbol{\alpha} \ \text{s.t.} \ \boldsymbol{\alpha} \ \text{is not in} \ \boldsymbol{\sigma}\} \qquad (12)$$

Though the generalized Potts frame is highly redundant, the motivation is that by introducing more expansion functions than strictly necessary and using an appropriate algorithm—such as for solving Equation 2—can yield both accurate and sparse representations of functions of crystalline configurations without the need of maximally incoherent measurements. An intuition for this can be formed by picturing the union of all subspaces spanned by size s subsets of functions in the generalized Potts frame. If the function sought lies on any of these subspaces or close enough, then the function can be accurately represented by only s terms.

As a simple illustration consider a symmetric binary diatomic system. Figure 2a shows the corresponding site spaces and site indicator functions along with the constant function ϕ_0. Additionally, the resulting symmetrized product bases are also shown in Figure 2b. The basis functions in both cases are colored to represent the possible CE bases. The union of all basis functions from these CE bases corresponds to the generalized Potts frame for this simple diatomic system. Again we highlight also the inclusion of products of *mixed* site basis sets, which corresponds to the orange vector built from

the product of a red and blue basis functions associated with each site respectively. In this case the function space over configurations of the binary diatomic molecule has three dimensions, however when a function lies close to or on any of the highlighted planes that function can be well approximated using only two terms.

V. APPLICATIONS TO MULTI-COMPONENT OXIDES

We demonstrate the performance of the generalized Potts frame by comparing fits for three fluorinated lithium-transition metal oxide systems with fits obtained using a site indicator based CE (with site basis functions given in Equation 5), and a CE using orthogonal sinusoid site basis functions.[27] These material systems are of interest as novel cathode materials made with abundant elements,[35] however their specific applications are not particularly relevant for the purpose of this work. The configuration spaces for the materials considered increase in both size and complexity (larger number of allowed species and number of symmetrically distinct sites) as shown in Figure 3.

The total number of expansion terms considered for each fit are listed as the model size in Table I. The total number of terms are obtained by using the same cut-off radius for clusters with up to four sites. Functions that evaluate to the same value (remain constant) for the training structures used in the corresponding fit are
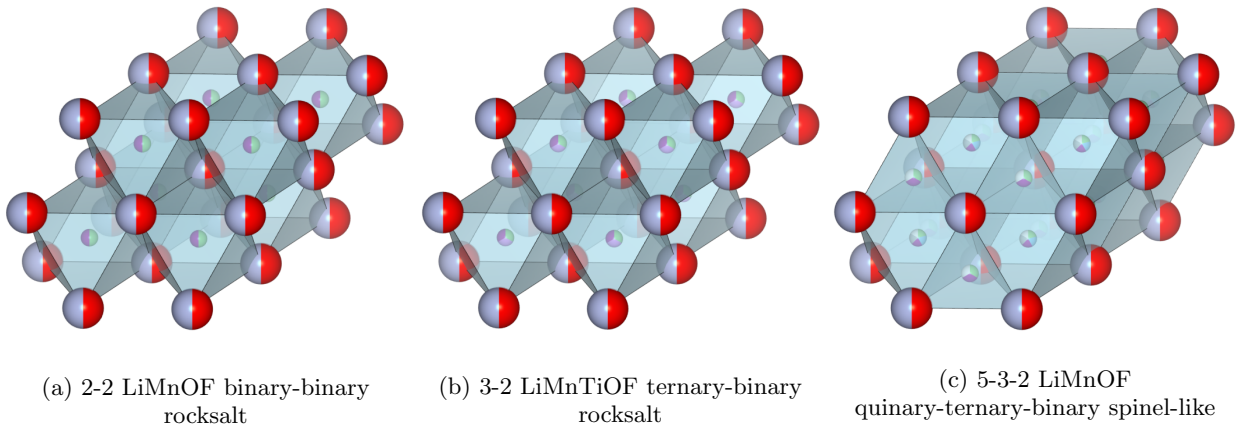
(a) 2-2 LiMnOF binary-binary rocksalt

(b) 3-2 LiMnTiOF ternary-binary rocksalt

(c) 5-3-2 LiMnOF quinary-ternary-binary spinel-like

FIG. 3: (a) System 2-2: Li-Mn-O-F rocksalt system with binary (Li+/Mn3+) cation sites and binary (O2-/F-) anion sites. A primitive cell with two sites is used as the formula unit for normalization. (b) System 3-2: Li-Ti-Mn-O-F rocksalt system with ternary (Li+/Mn3+/Ti4+) cation sites and binary (O2-/F-) anion sites. A primitive cell with two sites is used as the formula unit for normalization. (c) System 5-3-2: Li-Mn-O-F spinel-like system with quinary (Li/Mn2+/Mn3+/Mn4+/vacancy) octahedral cation sites, ternary (Li/Mn2+/vacancy) tetrahedral cation sites, binary (O2-/F-) anion sites. A primitive cell with four sites is used as the formula unit for normalization.

System	Expansion Type	Training Set Size	Test Set Size	Model Size
2-2	CEM	112	337	121
	Potts Frame	112	337	520
3-2	CEM	195	456	312
	Potts Frame	195	456	1040
5-3-2	CEM	312	56	7030*
	Potts Frame	312	56	23070*

TABLE I: Regression model and training/test data size specifications for the three fluorinated lithium-transition metal oxide systems. *Removal of correlation functions that remained constant for structures in the training set reduced the number of columns in the measurement matrices in 5-3-2 system to 4194 and 17350 for the indicator basis CEM based and Potts frame models respectively.

subsequently removed from the final measurement matrix used for each fit. Particularly, removal of constant functions was only required for the 5-3-2 system, for which the total number of columns in the measurement matrix ended up being 4194 and 17350 for the indicator basis CEM based and Potts frame models respectively. Additionally, we include an electrostatic energy term[8, 36] as an additional feature in every fit. Finally, for the fits using the Potts frame, we remove one cluster indicator function from each set associated with the same orbit; this is simply to do away with the trivial linear relation in Equation 11 applied to cluster concentrations. Doing so has a minimal effect in reducing redundancy, however we found this slightly improved efficiency for obtaining full rank measurement matrices. It is clear by construction that for the same spatial cutoffs, the number of terms in the generalized Potts frame far exceeds the number of terms in a standard cluster expansion, and this difference grows exponentially with the configuration space complexity.

The number of training structures and test structures for each fit is also listed in Table I. The structures were generated by using an initial set of structures from the Materials Project[37], and subsequently using this set and Monte Carlo to generate additional structures of up to 216 atoms. The formation energy for each training/test structure was computed with density functional theory using the Vienna ab initio simulation package[38] using the projector augmented wave (PAW) method,[39] with reciprocal space discretization of 25 k-points per Å and a plane wave energy cutoff of 520 eV. We carry out spin-polarized calculations using the SCAN meta-GGA exchange correlation [40] and pseudopotentials which include semicore states: Li_sv, Mn_pv, O, and F. Structures are converged to 10^{-6} eV in total energy and 0.01 eV/Å on atomic forces.

A total of 50 different fits for each system are computed by selecting a random set of training structures that gives a full rank underdetermined system. The remaining structures are used as a test set. The regression problem solved which is known as *Basis Pursuit Denoising* or the *LASSO* is slightly different than the ℓ_1-analysis

problem given in Equation 2. The corresponding optimization problem is,

$$\hat{\boldsymbol{J}} = \underset{\widetilde{\boldsymbol{J}}}{\text{argmin}} \; ||\widetilde{\boldsymbol{J}}||_1 \text{ subject to } ||A\Pi\widetilde{\boldsymbol{J}} - E||_2 \leq \varepsilon \quad (13)$$

The problem in Equation 13 is solved for each fit using the python package `scikit-learn`[41] in a two-step process. First a 10-fold hyperparameter cross validation optimization search is done with the training set. Subsequently a finer hyperparameter search is done centered at the previously obtained value now optimizing for out-of-sample error with respect to the test set but training only with the training set data. From the resulting fits, those with sparsity (number of nonzero coefficients) above the third quartile of the set are considered outliers and removed from the results.

Values for accuracy metrics and sparsity results for the fitted expansion obtained from the set of fits for each of the three expansion types and for each of the three materials systems are shown in boxplots in Figure 4. The average prediction accuracy metrics given include cross validation root mean squared error (CV RMSE) for the initial cross validation hyperparameter search, the out of sample RMSE for the final fit (test structures only), and the full data RMSE for both the training and test structures combined. The average sparsity value for the resulting fits is also listed.

Figure 4 shows boxplots depicting the resulting fit statistics in terms of cross validation, out of sample and full sample (both training and testing structures) root mean squared errors, along with the corresponding model sparsity values. The results depicted in Figure 4 show that for all systems, the expansions resulting from the Potts frame tend to have either a lower minimum, median and mean accuracy metrics, a lower spread in these accuracy metrics or both. In all cases the accuracy metrics are either very competitive (very close to standard cluster expansions), or better. As for the resulting model sparsity, although considerably more terms are included in the Potts frame fit, the resulting expansions have similar sparsity along with lower spread in sparsity values. These results demonstrate the applicability of using coherency and redundancy to obtain expansions that match and even exceed accuracy and sparsity of those obtained with standard cluster expansion basis sets.

The boxplots in Figure 4 provide a summary of the average values obtained for each type of fit, however they do not allow us to see which models have both high accuracy and high sparsity (low number of terms). Tables II and III list the same fit metrics as Figure 4 but for the single sparsest (smallest number of nonzero coefficients) fit and the fit resulting in the highest accuracy in terms of the full data RMSE for each materials system.

From Table II, we see that for the sparsest expansions obtained, the ones fitted using the Potts frame result in the lowest error metrics with only a few exceptions where the Potts frame fits are still of comparable accuracy. Furthermore, the expansions based on the Potts
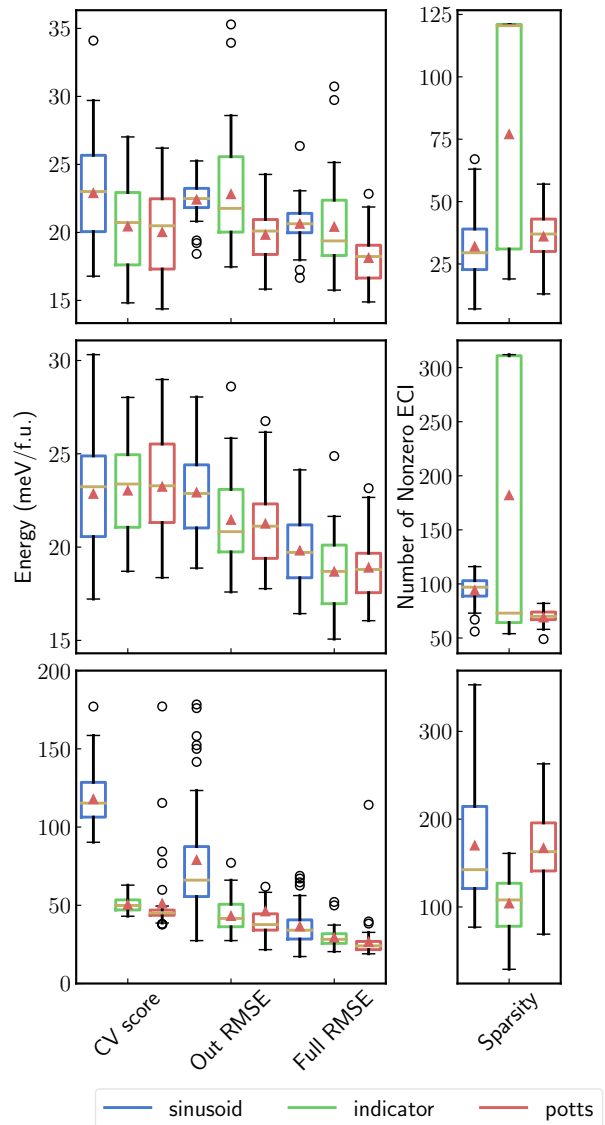

FIG. 4: Fit metric statistics for the systems tested using standard CEM with sinusoid site basis, indicator site basis and generalized Potts frame. The plotted metrics include: cross validation RMSE (CV score), out of sample RMSE (Out RMSE), full data RMSE (Full RMSE) for both the training and test structures combined, and the number of nonzero ECI in the fits (sparsity). LiMnOF binary-binary with two sites per formula unit (top), LiMnTiOF ternary-binary with two sites per formula unit (middle), LiMnOF quinary-ternary-binary with four sites per formula unit (bottom).

frame result in the lowest or second lowest sparsity for all three systems. The results for the most accurate models in terms of full dataset RMSE in Table III show that the Potts frame results in both sparse and accurate models. For the cases where one of the CEM based fits results in a lower full RMSE, the sparsity of that model is com-

System	Expansion	CV RMSE	Out RMSE	Full RMSE	Sparsity
	Sinusoid	34.10	25.26	26.36	7
2-2	Indicator	23.16	19.37	18.79	19
	Potts Frame	23.82	22.31	21.5	13
	Sinusoid	24.13	27.62	24.14	56
3-2	Indicator	25.14	21.69	19.55	54
	Potts Frame	22.26	23.43	20.93	49
	Sinusoid	131.76	158.16	68.80	77
5-3-2	Indicator	45.52	49.99	52.12	29
	Potts Frame	44.58	40.51	39.65	69

TABLE II: Fitted model accuracy metrics and sparsity of sparsest models. Cross validation RMSE (CV RMSE), out of sample RMSE (out RMSE), and full dataset RMSE (full RMSE) in meV per formula unit (random structure primitive cell).

System	Expansion	CV RMSE	Out RMSE	Full RMSE	Sparsity
	Sinusoid	25.68	18.42	16.66	50
2-2	Indicator	18.63	17.87	15.76	121
	Potts Frame	21.42	16.11	14.89	43
	Sinusoid	23.10	18.88	16.44	105
3-2	Indicator	28.02	17.89	15.07	310
	Potts Frame	23.09	17.77	16.06	72
	Sinusoid	104.39	36.72	17.24	291
5-3-2	Indicator	55.69	27.45	20.38	147
	Potts Frame	115.42	24.52	18.98	192

TABLE III: Fitted model accuracy metrics and sparsity of most accurate models in terms of the root mean squared error on the whole dataset. Cross validation RMSE (CV RMSE), out of sample RMSE (out RMSE), and full dataset RMSE (full RMSE) in formula unit (random structure primitive cell).

promised and substantially worse than the corresponding Potts frame fit. This behavior can also be observed in the results in Figure 4, where the box plot for sparsity obtained with the Potts frame has a smaller interquartile range than that of the CEM based fits (with the exception of the 5-3-2 system where it is only slightly larger than that of the indicator based cluster expansion).

Figure 5 shows the sorted magnitudes of the fitted coefficients (magnitude of the ECI times the multiplicity of the correlation function) for each expansion type and materials system for both the sparsest models and most accurate models obtained. Based on the adequacy of the resulting fit error metrics and the fast decay of coefficients shown in Figure 5, we can conclude that the configuration energy if not exactly sparse is highly compressible—since signals or functions with power law decaying (or faster) coefficients can be well approximated by a small subset of terms.[28] Specifically, a series of coefficients obey a power law decay if the sorted sequence satisfies the following,

$$|c_i| \leq Ci^{-q} \tag{14}$$

where c_i are the coefficients and $C, q > 0$ are constants. For all expansion types we see that the coefficient magnitude decay is faster than the power law decay shown.

We also see that in the results in Figure 5, the coefficient decay of fits using the Potts frame are, at worst, the second fastest decaying series for both the sparsest and

most accurate fits, such that expansions using the Potts frame are arguably more reliable in yielding sparser models than standard cluster expansions. This seems surprising considering that the total number of terms in the underdetermined system is exceedingly larger than that of the standard cluster expansions. However, considering the geometry of the union of s-dimensional subspaces as illustrated in Figure 2, we posit that indeed functions of configurational energy lie close to one of these subspaces with high probability, and we can therefore accurately represent the function with a much smaller number of terms than the total considered in the underdetermined system. Additionally, in light of the Theorem in Equation 3, the rapid decay of coefficients suggest that an s-sparse set of coefficients is close to the real coefficients and as a result the second term in the function approximation error is likely to be very small.

We can take the previous results as ex post facto evidence that configuration energy is a compressible function and that CS with redundancy and coherent measurements works well for fitting expansions of configuration energy. Additionally, we observed that expansions fitted using the Potts frame also tend to follow expectations driven by physical considerations. Figure 6 shows the number of nonzero coefficients for each crystallographic orbit considered for the most accurate indicator basis and Potts frame based models with respect to the number of nonzero coefficients obtained in the most accurate sinu-
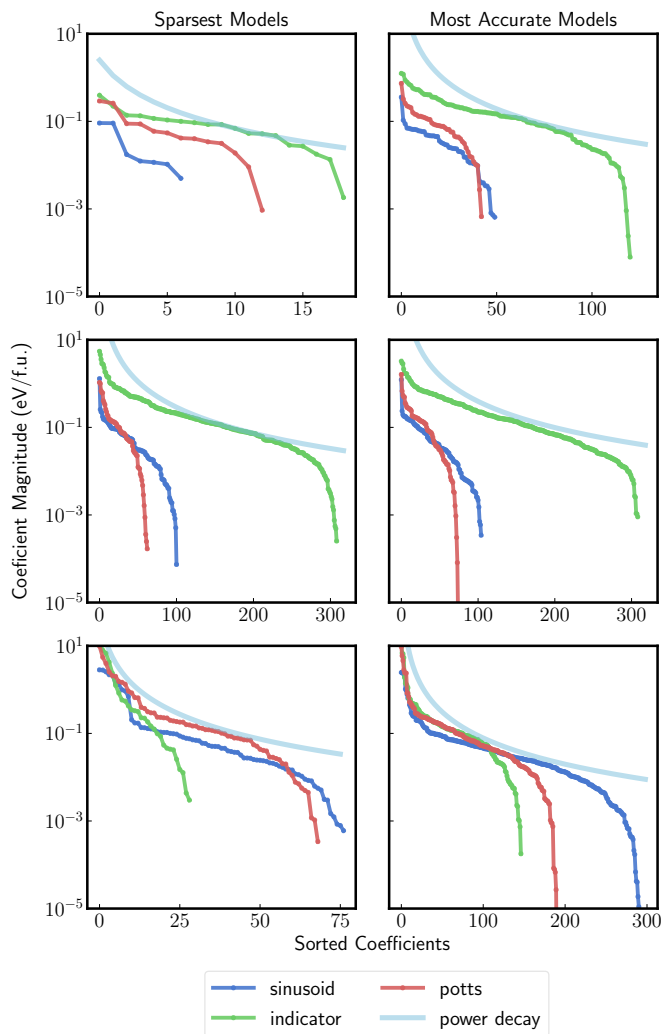
FIG. 5: Sorted fitted coefficient magnitudes (multiplicity times ECI) for sparsest and most accurate model (Full RMSE). LiMnOF binary-binary (top), LiMnTiOF ternary-binary (middle), LiMnOF quinary-ternary-binary (bottom).

FIG. 6: Number of nonzero coefficients relative to the sinusoid basis fit for each orbit, and norm of coefficients for each orbit for the most accurate models (Full RMSE). The vertical dotted lines separate the degree of orbit (pairs/triplets/quadruplets). LiMnOF binary-binary (top), LiMnTiOF ternary-binary (middle), LiMnOF quinary-ternary-binary (bottom).

soid basis based model. We notice from the plots, that using the Potts frame, despite having a much larger number of total coefficients associated with each orbit, results in fits that set a similar number of nonzero coefficients within each orbit and never exceeds three additional coefficients per orbit when compared with the sinusoid CEM fit. The significance is that, not only do we recover accurate and sparse models, but the models themselves also have similar sparsity structure to the CEM based models. Figure 6 also shows the root sum of squares or norm of the fitted ECI for each orbit. We observed that for the most accurate models the fitted coefficient weight associated with each orbit is less erratic than that of the indicator based cluster expansion. Additionally, the fitted coefficients follow the well-known CEM heuristic of coefficient decay with orbit size more fittingly than the
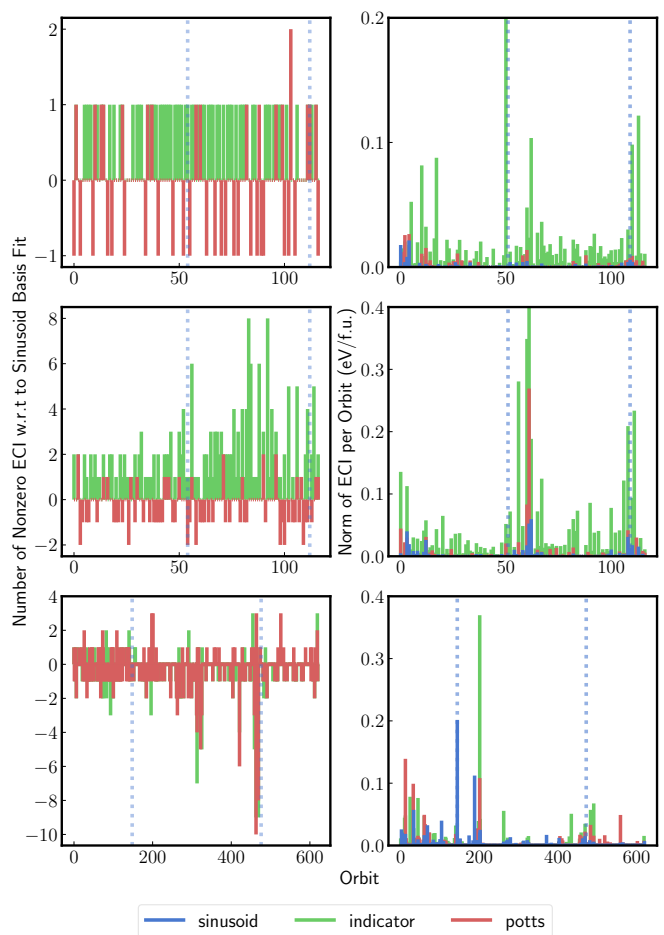
indicator based CEM. The aforementioned observations again demonstrate using the Potts frame can result in fitted expansions that are not only accurate and sparse, but also produce *well-behaved* coefficients that align with practices and heuristics based on physical insights used in the field.

VI. DISCUSSION & CONCLUSION

We have briefly introduced the concepts behind CS with redundant representations and coherent measurements within the context of the cluster expansion method, with the goal of addressing the obstacles preventing adequate sampling of training structures to satisfy requirements of classical CS when fitting CE models. We proposed the generalized Potts frame as a redundant

function set that can be used in lieu of standard CEM basis sets to represent functions over configurations of multicomponent crystalline materials. Additionally, we described the connections between the Potts frame and the commonly used nonorthogonal indicator function CEM basis set, as well as the essence of the Potts frame as a natural extension of the well known Potts model. Furthermore, we demonstrated the capability of obtaining arguably improved fitted expansions in terms of prediction accuracy, sparsity and robustness of coefficients compared to standard CEM fits. We stopped short of providing specific sensing and structure selection methods along with rigorous proofs that ensure small D-RIP constants; however the results from applications to complex lithium-transition metal oxide systems show promising results and the possibility of obtaining high accuracy expansions with such redundant representations.

Although this exposition of the power of using overcomplete frames in place of basis sets is particular in implementation and application, the concept can well be extended to many more problems in the growing space of novel methods for learning property-structure relationships in materials science. To our knowledge the concept has previously been used in the context of the CEM, although not explicitly, in the development of the variational basis cluster expansion and its connection to wavelets[42]—which can formally constitute a redundant frame. As a matter of fact, the process of fitting a cluster expansion for any heterovalent ionic material system using only a charge-neutral configurations is itself an application of fitting an expansion using a frame. The reason being that a set of CEM correlation functions used in the space of functions over constrained configurations in heterovalent systems is in itself an overcomplete or redundant representation, thus the set of CEM correlation functions is rigorously a frame for the function space over such constrained configurations. Additionally, the application of redundant frames along with CS using coherent measurements may also be worthwhile in other areas of computational materials science which could benefit from higher representation flexibility and additional robustness, particularly in methods where statistical learning methods are used to estimate basis expansions in high dimensional spaces. The simple idea of using redundant sets to construct redundant product functions can be directly applied in cluster expansions beyond discrete species to include continuous vector spaces at each site such as in the spin and atomic cluster expansion.[43, 44] Particularly in the representation and learning of interatomic potentials, function basis sets over continuous vector spaces can be replaced with corresponding frames over those spaces. Furthermore, when configurational degrees of freedom exist, frames over continuous vector spaces can be combined with finite frames, such as a generalized Potts frame; and so in such cases the resulting product basis of an atomic cluster expansion becomes a tensor product of continuous and finite frames.

With the advent of the high entropy paradigm, as more complex materials systems are explored it will become increasingly necessary to develop methodologies that can work aptly with the rapidly growing multidimensional spaces and relatively fixed training data set sizes. The compressed sensing and sparse representation framework has been developed particularly for these situations in other fields and has previously been shown to be successful in simple materials science applications.[17, 26] Continued exploration of systems with complexity beyond these systems requires continued development of methods to handle the increase in dimensionality. The generalized Potts frame developed and presented here represents one effort in said direction.

Appendix A: Frame Bounds for the Generalized Potts Frame

Definition A.1 *A countable sequence $\{\Phi_\gamma\}_{\gamma \in I}$ is said to be a **frame** for a Hilbert space $\mathcal{H}$ if there exist frame bounds $A, B > 0$ such that, [31, 32]*

$$A||H||^2 \leq \sum_{\gamma \in I} |\langle H, \Phi_\gamma \rangle|^2 \leq B||H||^2 \quad \forall H \in \mathcal{H}$$

The definition above implies that the frame $\{\Phi_\gamma\}_{\gamma \in I}$ spans $\mathcal{H}$.[31]

To derive frame bounds for the generalized Potts frame, we work with the product-basis of cluster indicator functions without taking symmetry adapted averages. The bounds obtained apply equally to the functions obtained from symmetry adapted averages. The lower frame bound A is obtained by splitting up the sum of projections into the frame elements as follows,

$$\sum_{\gamma \in I} |\langle H, \mathbf{1}_\gamma \rangle|^2 = \sum_{\gamma_{\max}} |\langle H, \mathbf{1}_{\gamma_{\max}} \rangle|^2 + \sum_{\gamma \in I \setminus \{\gamma_{\max}\}} |\langle H, \mathbf{1}_\gamma \rangle|^2$$

$$= ||H||^2 + \sum_{\gamma \in I \setminus \{\gamma_{\max}\}} |\langle H, \mathbf{1}_\gamma \rangle|^2$$

$$\geq ||H||^2$$

Where $\{\gamma_{\max}\}$ denotes the set of all maximal clusters, and thus represents the canonical orthonormal basis for the Hilbert space $\mathcal{H}$. The second sum over all smaller clusters is always greater than or equal to zero, and so the obtained lower frame bound is $A = 1$.

To obtain an upper frame bound we start by writing the Fourier expansion for cluster indicator functions,

$$
\begin{aligned}
\mathbf{1}_\gamma(\boldsymbol{\sigma}) &= \sum_\alpha \langle \mathbf{1}_\gamma, \Phi_\alpha \rangle \Phi_\alpha(\boldsymbol{\sigma}) \\
&= \sum_\alpha \prod_i \langle \mathbf{1}_{\gamma_i}, \phi_{\alpha_i} \rangle \Phi_\alpha(\boldsymbol{\sigma}) \\
&= \sum_{\substack{\alpha; \\ \mathrm{supp}(\alpha) \subseteq \mathrm{supp}(\gamma)}} \frac{1}{n^{\#\gamma}} \prod_i \phi_{\alpha_i}(\sigma_{\gamma_i}) \Phi_\alpha(\boldsymbol{\sigma}) \\
&= \sum_{\substack{\alpha; \\ \mathrm{supp}(\alpha) \subseteq \mathrm{supp}(\gamma)}} \frac{1}{n^{\#\gamma}} \Phi_\alpha(\boldsymbol{\sigma}_\gamma) \Phi_\alpha(\boldsymbol{\sigma})
\end{aligned}
$$

where $\mathrm{supp}(\cdot)$ represents the support or indices of the nonzero entries of α or γ, and $\#\gamma$ represents the total number of nonzero entries—i.e. the number of sites in a cluster. n is the number of species allowed at a site and $\boldsymbol{\sigma}_\gamma$ is any occupancy string that includes the cluster represented by γ. Note the above expression is for a system with a single type of site (i.e. with the same set of allowed species in all sites). The general expression simply involves one over the product of the different number of species for each site instead of the factor $1/n^{\#\gamma}$.

Using the expansion given above for cluster indicator functions, we obtain an upper frame bound as follows,

$$
\begin{aligned}
\sum_{\gamma \in I} |\langle H, \mathbf{1}_\gamma \rangle|^2 &= \sum_{\gamma \in I} \left| \left\langle H, \sum_{\substack{\alpha; \\ \mathrm{supp}(\alpha) \subseteq \mathrm{supp}(\gamma)}} \frac{1}{n^{\#\gamma}} \Phi_\alpha(\boldsymbol{\sigma}_\gamma) \Phi_\alpha(\boldsymbol{\sigma}) \right\rangle \right|^2 \\
&= \sum_{\gamma \in I} \left(\frac{1}{n^{\#\gamma}} \right)^2 \left| \sum_{\substack{\alpha; \\ \mathrm{supp}(\alpha) \subseteq \mathrm{supp}(\gamma)}} \hat{H}_\alpha \Phi_\alpha(\boldsymbol{\sigma}_\gamma) \right|^2 \\
&= \sum_S \frac{1}{n^{2|S|}} \sum_{\substack{\boldsymbol{\sigma}_\gamma; \\ \mathrm{supp}(\gamma) = S}} \left| \sum_{\substack{\alpha; \\ \mathrm{supp}(\alpha) \subseteq S}} \hat{H}_\alpha \Phi_\alpha(\boldsymbol{\sigma}_\gamma) \right|^2 \\
&= \sum_S \frac{1}{n^{|S|}} \sum_{\substack{\alpha; \\ \mathrm{supp}(\alpha) \subseteq S}} \hat{H}_\alpha^2 \\
&= \sum_\alpha \left(\sum_{S \supseteq \mathrm{supp}(\alpha)} \frac{1}{n^{|S|}} \right) \hat{H}_\alpha^2 \\
&\leq \left(\sum_S \frac{1}{n^{|S|}} \right) ||H||^2 \\
&= \left(1 + n^{-1} \right)^N ||H||^2
\end{aligned}
$$

where $\hat{H}_\alpha$ are the Fourier coefficients of H. The sets S contain site indices and the sum over these contains all possible subsets of site indices—i.e. all clusters of unlabeled sites. The upper frame bound obtained is $B = \left(1 + n^{-1} \right)^N$, where N is the total number of sites in the structure. The bound obtained is an improvement over the bound 2^N given by the Cauchy-Schwarz inequality.

[1] J. M. Sanchez, F. Ducastelle, and D. Gratias, Physica A: Statistical Mechanics and its Applications **128**, 334 (1984).

[2] J. M. Sanchez, Physical Review B **48**, 14013 (1993).

[3] A. Van der Ven, J. Thomas, B. Puchala, and A. Natarajan, Annual Review of Materials Research **48**, 27 (2018).

[4] A. van de Walle and G. Ceder, Journal of Phase Equilibria **23**, 348 (2002).

[5] J. M. Sanchez, Journal of Phase Equilibria and Diffusion **38**, 238 (2017).

[6] F. Otto, Y. Yang, H. Bei, and E. P. George, Acta Materialia **61**, 2628 (2013).

[7] X. Liu, J. Zhang, J. Yin, S. Bi, M. Eisenbach, and Y. Wang, Computational Materials Science **187**, 110135 (2021).

[8] W. D. Richards, Y. Wang, L. J. Miara, J. Chul Kim, and G. Ceder, Energy & Environmental Science **9**, 3272 (2016).

[9] A. Urban, J. Lee, and G. Ceder, Advanced Energy Materials **4**, 1400478 (2014).

[10] B. Ouyang, N. Artrith, Z. Lun, Z. Jadidi, D. A. Kitchaev, H. Ji, A. Urban, and G. Ceder, Advanced Energy Materials **10**, 1903240 (2020).

[11] R. J. Clément, D. Kitchaev, J. Lee, and Gerbrand Ceder, Chemistry of Materials **30**, 6945 (2018).

[12] T. Kostiuchenko, A. V. Ruban, J. Neugebauer, A. Shapeev, and F. Körmann, Physical Review Materials **4**, 113802 (2020).

[13] J. W. D. Connolly and A. R. Williams, Physical Review B **27**, 5169 (1983).

[14] V. Blum, G. L. W. Hart, M. J. Walorski, and A. Zunger, Physical Review B **72**, 165113 (2005).

[15] G. L. W. Hart, V. Blum, M. J. Walorski, and A. Zunger, Nature Materials **4**, 391 (2005).

[16] T. Mueller and G. Ceder, Physical Review B **80**, 024103 (2009).

[17] L. J. Nelson, G. L. W. Hart, F. Zhou, and V. Ozoliņš, Physical Review B **87**, 035125 (2013).

[18] Z. Leong and T. L. Tan, Physical Review B **100**, 134108 (2019).

[19] G. C. Calafiore and L. El Ghaoui, *Optimization Models* (Cambridge University Press, Cambridge, United Kingdom, 2014).

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics (Springer-Verlag, New York, 2001).

[21] M. Ångqvist, W. A. Muñoz, J. M. Rahm, E. Fransson, C. Durniak, P. Rozyczko, T. H. Rod, and P. Erhart, Advanced Theory and Simulations **2**, 1900015 (2019).

[22] E. J. Candes and M. B. Wakin, IEEE Signal Processing Magazine **25**, 21 (2008).

[23] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall, Applied and Computational Harmonic Analysis **31**, 59 (2011).

[24] E. P. George, D. Raabe, and R. O. Ritchie, Nature Reviews Materials **4**, 515 (2019).

[25] 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, Nature Materials **20**, 214 (2021).

[26] L. J. Nelson, V. Ozoliņš, C. S. Reese, F. Zhou, and G. L. W. Hart, Physical Review B **88**, 155105 (2013).

[27] A. van de Walle, Calphad Tools for Computational Thermodynamics, **33**, 266 (2009).

[28] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok, in *Compressed Sensing*, edited by Y. C. Eldar and G. Kutyniok (Cambridge University Press, Cambridge, 2012) pp. 1–64.

[29] X. Zhang and M. H. F. Sluiter, Journal of Phase Equilibria and Diffusion **37**, 44 (2016).

[30] G. L. W. Hart and R. W. Forcade, Physical Review B **77**, 224115 (2008).

[31] O. Christensen, *Frames and Bases: An Introductory Course*, Applied and Numerical Harmonic Analysis (Birkhäuser Basel, 2008).

[32] S. Waldron, *An Introduction to Finite Tight Frames*, Applied and Numerical Harmonic Analysis (Birkhäuser Basel, 2018).

[33] R. B. Potts, Mathematical Proceedings of the Cambridge Philosophical Society **48**, 106 (1952).

[34] F. Y. Wu, Reviews of Modern Physics **54**, 235 (1982).

[35] R. J. Clément, Z. Lun, and G. Ceder, Energy & Environmental Science **13**, 345 (2020).

[36] A. Seko and I. Tanaka, Journal of Physics: Condensed Matter **26**, 115403 (2014).

[37] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. a. Persson, APL Materials **1**, 011002 (2013).

[38] G. Kresse and J. Furthmüller, Computational Materials Science **6**, 15 (1996).

[39] G. Kresse and D. Joubert, Physical Review B **59**, 1758 (1999).

[40] J. Sun, A. Ruzsinszky, and J. P. Perdew, Physical Review Letters **115**, 036402 (2015).

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, Journal of Machine Learning Research **12**, 2825 (2011).

[42] J. M. Sanchez, Physical Review B **81**, 224202 (2010).

[43] R. Drautz and M. Fähnle, Physical Review B **69**, 104404 (2004).

[44] R. Drautz, Physical Review B **99**, 014104 (2019).