






An $\ell_0\ell_2$ -norm regularized regression model for construction of robust cluster expansions in multicomponent systems

Peichen Zhong ^{1,2}, Tina Chen ^{1,2}, Luis Barroso-Luque ^{1,2}, Fengyu Xie ^{1,2} and Gerbrand Ceder ^{1,2,*}

¹*Department of Materials Science and Engineering, University of California, Berkeley, California 94720, United States*

²*Materials Sciences Division, Lawrence Berkeley National Laboratory, California 94720, United States*



(Received 28 April 2022; revised 27 June 2022; accepted 13 July 2022; published 25 July 2022)

We introduce $\ell_0\ell_2$ -norm regularization and hierarchy constraints into linear regression for the construction of cluster expansions to describe configurational disorder in materials. The approach is implemented through mixed integer quadratic programming (MIQP). The ℓ_2 -norm regularization is used to suppress intrinsic data noise, while the ℓ_0 -norm is used to penalize the number of nonzero elements in the solution. The hierarchy relation between clusters imposes relevant physics and is naturally included by the MIQP paradigm. As such, sparseness and cluster hierarchy can be well optimized to obtain a robust, converged set of effective cluster interactions with improved physical meaning. We demonstrate the effectiveness of $\ell_0\ell_2$ -norm regularization in two high-component disordered rocksalt cathode material systems, where we compare the cross-validation, convergence speed, and the reproduction of phase diagrams, voltage profiles, and Li-occupancy energies with those of the conventional ℓ_1 -norm regularized cluster expansion models.

DOI: [10.1103/PhysRevB.106.024203](https://doi.org/10.1103/PhysRevB.106.024203)

I. INTRODUCTION

First-principles density functional theory (DFT) calculations have been demonstrated as a reliable tool in computational materials science. Despite the increase in computing power and accuracy of DFT methods, the scaling with the number of atoms $\sim O(n^3)$ intrinsically prohibits large-scale calculations (over 10^3 atoms) or sampling of a high-dimensional occupancy space (millions of structures) [1]. This is particularly relevant in systems with configurational degrees of freedom that need to be sampled at nonzero temperature to equilibrate states of partial order, and their associated entropy and free energy. The state of configurational order determines many materials properties, especially in systems composed of many species (high-entropy systems). It has also recently been shown to be relevant for mechanical properties in metallic alloy systems [2,3] and the energy density of complex electrode materials for energy storage application [4–7].

The cluster expansion (CE) method has been well developed to describe such configurational energetics for metallic alloys [8,9], as well as for ionic systems [10,11]. The CE method expands any property (e.g., formation energy, volume) in terms of the distribution of atoms on a set of predefined sites. When the quantity being expanded is the energy, the expansion coefficients are referred to as effective cluster interactions (ECIs). For example, in a multicomponent system, the energy is expanded as

$$E(\sigma) = \sum_{\beta} m_{\beta} J_{\beta} \langle \Phi_{\alpha \in \beta} \rangle_{\beta}, \quad \Phi_{\alpha} = \prod_{i=1}^N \phi_{\alpha_i}(\sigma_i). \quad (1)$$

A configuration σ represents a specific occupancy on all the sites of the system, where σ_i describes which species sits on the i -th site of the structure. The site basis function $\phi_{\alpha_i}(\sigma_i)$ transforms the occupancy variable σ_i into a scalar value. There are typically as many (nonconstant) cluster basis functions as possible occupancies on a site minus one. The cluster basis function label $\alpha = (\alpha_1, \alpha_2, \alpha_3, \dots)$ indicates a group of sites, each with a specific basis function on it, where each entry α_i labels the corresponding site basis function ϕ_{α_i} . Thus the cluster basis function $\Phi_{\alpha} = \prod_{i=1}^N \phi_{\alpha_i}(\sigma_i)$ can be obtained by taking the product of site basis functions.

For example, the cation sublattice of a LiMnO₂ rocksalt oxide is a binary system, where Li and Mn share the octahedral interstitial of the FCC anion framework. In such a system, Li can be encoded by $\sigma^{\text{Li}} = 0$ and Mn by $\sigma^{\text{Mn}} = 1$. The parameter α_j takes a value from $[0, 1, \dots, M - 1]$, where M is the number of allowed species defined on the sublattice (e.g., $M = 2$ for Li-Mn). While many forms of site basis function can be used [8,12,13], a sinusoid (orthogonal) basis function is applied here to transform the occupancy variable ($\sigma^{\text{Li}}, \sigma^{\text{Mn}}$) into a value [14], where

$$\phi_j(\sigma_i) = \begin{cases} 1 & \text{if } j = 0 \\ -\cos\left(\frac{\pi(j+1)\sigma_i}{M}\right) & \text{if } j \text{ is odd} \\ -\sin\left(\frac{\pi j \sigma_i}{M}\right) & \text{if } j \text{ is even} \end{cases}. \quad (2)$$

The j indicates α_i in Eq. (1) and can take a value of 0 or 1. Thus we have $\phi_{j=0} \equiv 1$, $\phi_{j=1}(\sigma^{\text{Li}} = 0) = -1$, and $\phi_{j=1}(\sigma^{\text{Mn}} = 1) = 1$. This situation corresponds to the spin variables used in a generalized Ising model [15,16]. For systems with species number $M > 2$, the basis functions take values beyond those of spin variables $\{-1, 1\}$ typically used in binary CE. Some examples of other types of site-basis functions also developed for the CE method are the

*gceder@berkeley.edu

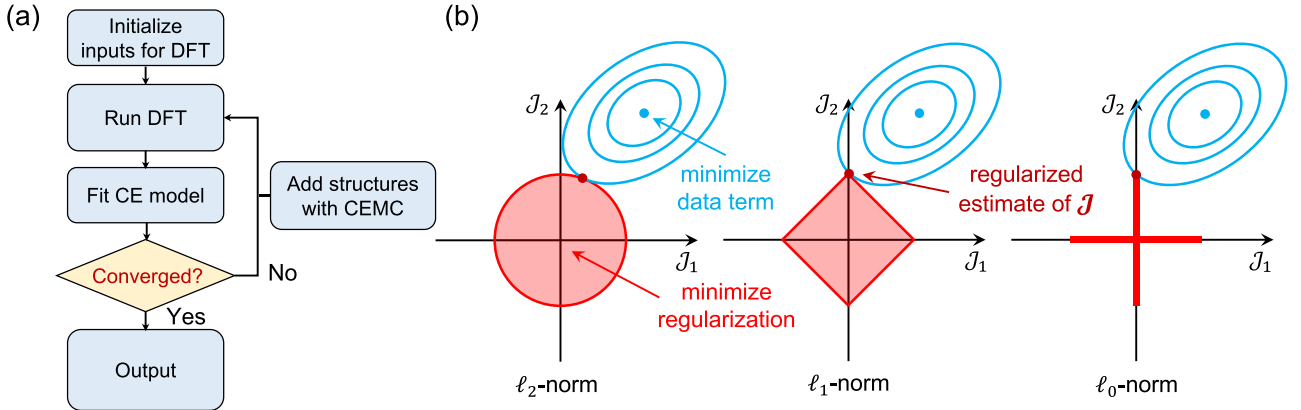


FIG. 1. (a) The general flowchart of constructing a CE model, including initialization of input structures, DFT calculations, fitting and convergence check, and cluster expansion Monte Carlo (CEMC) for sampling. (b) An illustration of ℓ_2 , ℓ_1 , and ℓ_0 -norm regularization in a two-parameter space $\mathbf{J} = (J_1, J_2)$. The blue circles represent the contours of the data term $\|\mathbf{E}_{\text{DFT},S} - \mathbf{\Pi}_S \mathbf{J}\|_2^2$ in cost function. The red regions represent the constraints of parameters (e.g., $J_1^2 + J_2^2 \leq s$ for ℓ_2 -norm, $|J_1| + |J_2| \leq s$ for ℓ_1 -norm.) The dark red point is the intersection of data term and regularization of parameters, which jointly determines the estimation of \mathbf{J} .

Chebyshev polynomials [8] and the indicator function (point delta function) [12].

In Eq. (1), the correlation function $\langle \Phi_\alpha \rangle_\beta$ is calculated by

$$\langle \Phi_\alpha(\sigma) \rangle_\beta = \frac{1}{N_\sigma m_\beta} \sum_{\alpha \in \beta} \Phi_\alpha(\sigma), \quad (3)$$

where β is an orbit representing all symmetrically equivalent cluster basis functions α , and m_β is the corresponding multiplicity. N_σ is the size of the supercell of configuration σ ; thus, the correlation function is well normalized with respect to the primitive cell. J_α is the effective cluster interaction (ECI). We refer readers to Refs. [8,14,17,18] for a more extended description of the CE. From Eq. (1), the CE energy is linearly dependent on the ECIs \mathbf{J} when the configuration σ is fixed. We can thus write

$$E_{\text{CE}}(\sigma) = \mathbf{\Pi}(\sigma) \cdot \mathbf{J}, \quad (4)$$

where $\mathbf{\Pi}(\sigma)$ is a row vector of correlation functions and \mathbf{J} is the column vector of ECIs.

Figure 1(a) presents a brief illustration of how to iteratively construct a CE Hamiltonian. In practice, the CE model is initially fitted on a small set of DFT calculations. Then, a simple CE is fitted that can be used in a Monte Carlo simulation to sample new structures. DFT calculations will be applied to a sample of the MC-obtained structures, and a new CE will be fitted. This procedure will be performed iteratively until the model is converged (i.e., the cross-validation error remains low and stable, the model reproduces DFT ground states well, etc.) [19]. In such a process, it is always desirable to achieve fewer training iterations, as DFT calculations are costly in terms of CPU time. On the other hand, fewer structures may also result in a worse fitting due to insufficient sampling of the configuration space.

Obtaining reliable ECIs \mathbf{J} from the DFT energy of a set of configurations is the central problem of CE fitting. Given a set of input occupancy configurations S , the set of correlation vectors forms a feature matrix $\mathbf{\Pi}_S = [\mathbf{\Pi}_1, \mathbf{\Pi}_2, \dots]$, and the corresponding DFT energies are used to construct the target vector $\mathbf{E}_{\text{DFT},S}$. Determining the ECIs is an inverse problem of

Eq. (4), also called linear regression. Generally, the problem can be solved by minimizing the cost function

$$\min_{\mathbf{J}} \|\mathbf{E}_{\text{DFT},S} - \mathbf{\Pi}_S \mathbf{J}\|_2^2 + \mu \|\mathbf{J}\|_p, \quad \|\mathbf{x}\|_p = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}}, \quad (5)$$

where the p -norm of \mathbf{J} is added to regularize the fit and suppress over-fitting, and μ controls the degree of regularization. Figure 1(b) shows the comparison of ℓ_2 , ℓ_1 , and ℓ_0 -norm regularization in a two-parameter space $\mathbf{J} = (J_1, J_2)$. The blue circles are the contours of the data term error $\|\mathbf{E}_{\text{DFT},S} - \mathbf{\Pi}_S \mathbf{J}\|_2^2$. The red regions represent the regularization constraints on the parameters ($\|\mathbf{J}\|_p \leq s$), which can be transformed to a Lagrangian form $\mu \|\mathbf{J}\|_p$ as shown in Eq. (5). The dark red point is the regularized estimation of \mathbf{J} , which is the intersection between the data error term and the regularization term. The ℓ_1 -norm tends to generate sparser solutions compared with the ℓ_2 -norm, because the intersection is likely to be located on the axis. The ℓ_0 -norm counts the nonzero elements of \mathbf{J} , where the intersection is exactly located on the axis and thus the ℓ_0 -norm imposes an exact sparsity constraint on \mathbf{J} .

Conventionally, ℓ_2 -norm (*ridge regression*, $p = 2$) regularization can be applied when the problem is over-determined (i.e., the number of training structures is larger than the dimension of \mathbf{J}). The ℓ_2 -norm regularized regression reduces the over-fitting caused by intrinsic noise in the training data. This can be achieved solely by introducing the ℓ_2 regularization function and additionally using the mixed-basis expansion [20–22]. Bayesian approaches have also been successfully applied to estimate the ECIs with a prior distribution in several binary systems [22–24]. However, the number of ECIs increases combinatorially with the number of species, scaling approximately as $\prod_k (M_k - 1)^{n_k}$, where M_k is the number of species on the k -th sublattice, and n_k is the number of cluster sites in the same sublattice k . The explosion in the number of basis functions when many species can occupy a site makes it difficult to predefine which cluster basis functions contribute to the expansion for high dimensional

multicomponent systems (i.e., which cluster basis function has a nonzero element in the solution of \mathbf{J}). Therefore a sparse solver for ECIs selection is required.

Rigorously, the exact sparse solution of Eq. (5) is obtained with ℓ_0 -norm regularization of \mathbf{J} . However, it is hard to compute the $\|\mathbf{J}\|_0$ in the cost function as it is an NP-hard problem. In the compressive sensing paradigm, the ℓ_0 -norm can be transformed to an ℓ_1 -norm when the feature matrix $\mathbf{\Pi}_S$ satisfies the restricted isometry property (RIP) condition [25]. To satisfy the RIP condition in CE, Nelson *et al.* [26] proposed to generate a training set in which each row is an identical independent distributed (i.i.d.) random vector. However, in practical cases, the configurations in the training set S are correlated, because structures are not randomly sampled, but are mostly part of an ensemble of configurations with low energy. Such correlations fail to satisfy the i.i.d. condition. Moreover, generating structures from a specific correlation vector is also an NP-hard problem. Though the strict compressive sensing cluster expansion is not easy to construct in practice, the ℓ_1 -norm (*lasso*, $p = 1$) regularization is widely used as feature selection, which has shown success in various alloy and ionic systems [13,26–28].

In this paper, we propose an $\ell_0\ell_2$ -norm regularization approach that incorporates hierarchy constraints to generate more robust and predictive CE models. First, we introduce the $\ell_0\ell_2$ -norm penalty term and hierarchy constraints in the paradigm of mixed-integer quadratic programming (MIQP). Second, we compare the sparseness and convergence rate of ECIs with those of the conventional ℓ_1 method in the Li-Mn-V-Ti-O-F disordered rocksalt system. Finally, we demonstrate that an $\ell_0\ell_2$ -regularized CE better reproduces the correct physical interactions by comparing with ℓ_1 -CE in terms of computed phase diagrams, voltage profiles, and related physical quantities in the Li-Mn-Ti-O system.

II. METHODS

A. The ℓ_0 -norm regularization

In Eq. (5), $p = 0$ manifests itself as a pseudonorm that counts the number of nonzero elements of \mathbf{J} :

$$\|\mathbf{J}\|_0 = \sum_i \text{Ind}(J_i), \quad \text{Ind}(J_i) = \begin{cases} 0, & J_i = 0 \\ 1, & J_i \neq 0 \end{cases}. \quad (6)$$

Adding the ℓ_0 term into the cost function directly penalizes the number of nonzero ECIs, yielding better sparseness in its solution. However, optimizing a cost function with an ℓ_0 term is an NP-hard problem and is difficult to present in a direct way [25,29]. Previously, Huang *et al.* [30] has approached the problem by rewriting ℓ_0 optimization as a mixed-integer programming problem, such that

$$\begin{aligned} \min \|\mathbf{J}\|_0 &\Leftrightarrow \min \sum_{c \in \mathcal{C}} z_{0,c} \\ \text{s.t.} \quad &Mz_{0,c} \geq J_c, \quad \forall c \in \mathcal{C}, \\ &Mz_{0,c} \geq -J_c, \quad \forall c \in \mathcal{C}, \\ &z_{0,c} \in \{0, 1\}, \quad \forall c \in \mathcal{C}, \end{aligned} \quad (7)$$

where M is a sufficiently large number (larger than the maximum possible absolute value of any ECI), and $z_{0,c}$ is a slack variable (binary integer) indicating whether the ECI of

orbit c is zero or not. J_c is constrained to 0 when the slack variable $z_{0,c} = 0$ (inactive) and to $[-M, M]$ when $z_{0,c} = 1$ (active). (For a rigorous mathematical background, refer to Ref. [30].) In practice, it is shown that one can at least obtain a sparseness-improved near-optimal solution within a reasonable CPU time cutoff. In our benchmark tests of the Li-Mn-V-Ti-O-F and Li-Mn-Ti-O systems, the optimizations of ECIs were completed within 600s using the GUROBI package [31].

B. Hierarchy constraints

In a CE, clusters are usually enumerated in an iterative, low-to-high order (i.e., from singlets to pairs, triplets, quadruplets, and so on). Practically, the CE is truncated to a maximum of n (e.g., quadruplet clusters with $n = 4$ are a typical limit), ignoring the higher-order interactions to control the model complexity. To differentiate the cluster orbits by different significance, we take one of the basic assumptions of CE that n -body cluster interactions become less important to the configurational energy (or other scalar properties) as n becomes larger. This assumption means that the majority of the fitted property can be described by the lower-order interactions and that the higher-order interactions serve as the fine-tuning part in the fitting.

Such a physically inspired concept can be introduced in the form of hierarchy constraints, as has been done successfully in some previous studies [32–34]. The hierarchy constraint manifests itself as $J_b \neq 0$ if and only if $J_a \neq 0$ ($a \subset b$), where a and b are a lower- and higher-order cluster function orbit, respectively, and b contains all the site bases of a as a subset. In the MIQP representation, the hierarchy relationship can be easily expressed as a constraint between slack variables:

$$z_{0,b} \leq z_{0,a}, \quad a \subset b. \quad (8)$$

This treatment was first proposed by Huang *et al.* [30], where it was used in the $\ell_0\ell_1$ -norm regularization paradigm.

C. The ℓ_2 -norm regularization

We propose that combining ℓ_2 -norm and ℓ_0 -norm regularization can impose true hierarchy constraints unlike the $\ell_0\ell_1$ -norm. It is to be noted that the inequality between slack variables does not necessarily impose the hierarchy relation ($J_b \neq 0$, iff $J_a \neq 0$). This is because the hierarchy constraints are defined on the magnitude of ECIs J_a and J_b , while the slack variables $z_{0,b}$, $z_{0,a}$ are intermediate to represent the presence or exclusion of the variables.

When implementing the hierarchy constraints in $\ell_0\ell_1$ -norm regularization, pseudoactive behavior can manifest itself when a $J = 0$, but its slack variable $z_0 = 1$ within the MIQP paradigm. \mathbf{J} can be regularized to zero, which is still a valid solution between $[-M, M]$, even with $z_0 = 1$. This is caused by the fact that the ℓ_1 -norm has feature-selection properties that intrinsically produce a sparse solution [35]. This pseudoactiveness can introduce excessive sparseness to the solution and break the hierarchy constraints. Figure 2 presents an example of pseudoactiveness in $\ell_0\ell_1$ -norm regularization. The excessive sparseness is introduced to the orbit β with $J_\beta = 0$, while all orbits α, β, γ has active slack variables

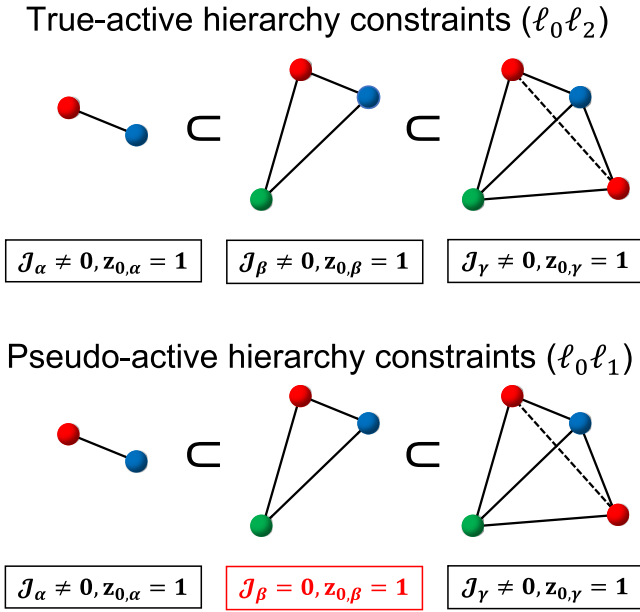


FIG. 2. Illustration of hierarchy relations ($\alpha \subset \beta \subset \gamma$) between pair, triplet, and quadruplet orbit. The different colors on the cluster sites represent the decorating species for a given site-basis function. The equation in red shows a pseudoactive hierarchy constraint that may appear in ℓ_1 -norm and its derivative methods.

$z_0 = 1$. The higher-order orbit γ is erroneously activated while $J_\beta = 0$. To avoid such a situation and ensure proper function with ℓ_0 under hierarchy constraints, a norm with no feature-selection properties is required. The ℓ_2 -norm is a natural choice.

With the introduction of the $\ell_0\ell_2$ -norm and hierarchy constraints, the final ECI optimization problem can be written as

$$\begin{aligned} \min_{\mathbf{J}} \mathbf{J}^T \mathbf{\Pi}_S^T \mathbf{\Pi}_S \mathbf{J}^T - 2\mathbf{E}_{\text{DFT}}^T \mathbf{\Pi}_S \mathbf{J} + \mu_0 \sum_{c \in \mathcal{C}} z_{0,c} + \mu_2 \|\mathbf{J}\|_2^2 \\ \text{s.t.} \quad M_{z_{0,c}} \geq J_c, \forall c \in \mathcal{C}, \\ M_{z_{0,c}} \geq -J_c, \forall c \in \mathcal{C}, \\ z_{0,b} \leq z_{0,a}, \forall a \subset b, \{a, b\} \in \mathcal{C}, \\ z_{0,c} \in \{0, 1\}, \forall c \in \mathcal{C}, \end{aligned} \quad (9)$$

where $\|\mathbf{J}\|_2^2 = \mathbf{J}^T \mathbf{J}$ penalizes the magnitude of ECIs, thus avoiding over-fitting by regularizing sampling noise while the ℓ_0 -term $\sum_c z_{0,c}$ optimizes the sparseness. The hierarchy constraints ensure correct containment relationships by manifesting lower-order ECIs first to reduce redundancy. The packages for our implementation are available in Ref. [36].

III. RESULTS

As mentioned above, for systems with many species, the number of basis functions grows rapidly. An example of such systems are the Li-excess disordered rocksalts (DRX), which are multicomponent systems that can be synthesized with a wide variety of elements [37]. Recently, high-entropy DRX materials have been synthesized with up to 12 metal species [6]. In addition, their configurational short-range order is critical to their transport properties, warranting a detailed CE approach [5,38]. Here, we provide a heuristic solution to

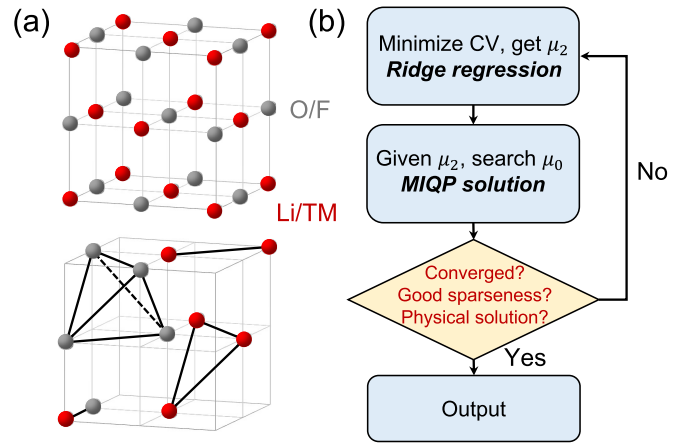


FIG. 3. (a) An illustration of the rocksalt lattice structure. The cation sites are labeled in red and can be occupied by Li^+ and transition metals (TM, including Mn^{2+} , V^{3+} , and Ti^{4+} in our example) in DRX. The anion sites are labeled in gray and can be occupied by O^{2-} and F^- . The lower panel gives some examples of n -body ($n = 2, 3$, and 4) clusters used in the CE model, including intra and inter-sublattice interactions. (b) The procedure to obtain an $\ell_0\ell_2$ -norm regularized solution, including finding the μ_2 by minimizing the CV error in ridge regression, sparseness engineering with ℓ_0 -norm using MIQP, and terminating if the solution is converged with good sparseness, as well as good-reproduction-relevant physical properties.

study configurations in such high-dimensional DRX systems by applying an $\ell_0\ell_2$ -regularized CE model to fit the formation energy of DRX compounds.

A. Robustness and convergence

The convergence of the CE when the $\ell_0\ell_2$ -norm and hierarchy constraints are enforced was tested on configurational disorder in the $\text{LiF-MnO-LiVO}_2\text{-Li}_2\text{TiO}_3$ composition space. The CE model contains pair interactions up to 7.1 \AA , triplets up to 4.0 \AA , and quadruplets up to 4.0 \AA based on a lattice parameter $a = 3 \text{ \AA}$ for the primitive cell. Figure 3(a) presents the rocksalt framework of a DRX structure. The framework contains a cation sublattice (red) and anion sublattice (gray), where the cation sites can be occupied by Li and transition metals (TM, including Mn^{2+} , V^{3+} , and Ti^{4+} in this example) and the anion sites can be occupied by O^{2-} and F^- . A species indicator where the site basis function reads $\phi_j(\sigma_i) = \delta_{i,j}$ was used [12]. The electrostatic energy (Ewald energy E_0/ϵ_r) is also included to capture long-range electrostatic interactions (E_0 is the unscreened electrostatic energy and $1/\epsilon_r$ is fitted as one of the ECI ($1/\epsilon_r \geq 0$) [39,40]. In total, 162 ECIs (including the constant term J_0) are predefined in the CE Hamiltonian. The dimension of the feature matrix $\mathbf{\Pi}_{\text{DFT},S}$ is 487×162 . The performance of the $\ell_0\ell_2$ -CE is compared with the ℓ_1 -CE. We emphasize two major improvements in the $\ell_0\ell_2$ -CE.

1. Sparseness versus cross-validation error

Cross-validation (CV) error versus model complexity is a general metric used to evaluate the robustness of a CE model.

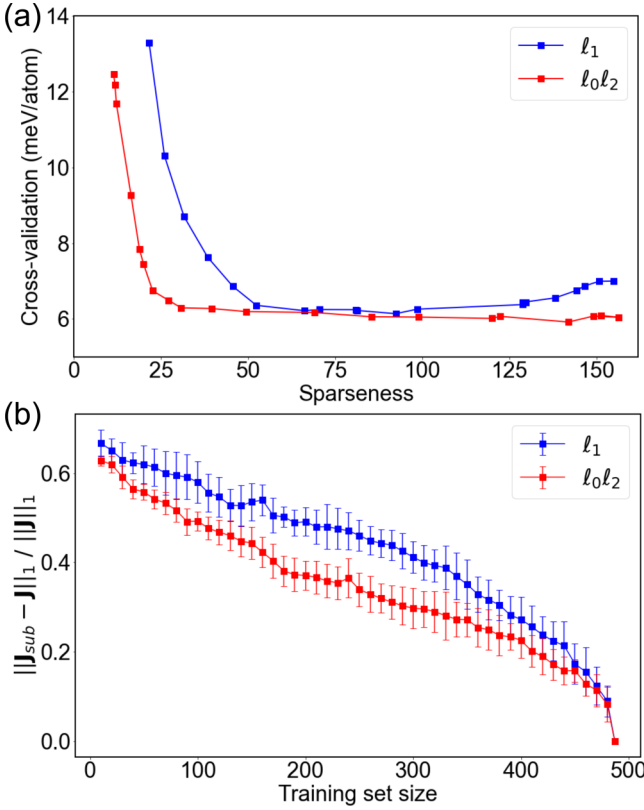


FIG. 4. (a) Cross-validation error (meV/atom) of the ℓ_1 -CE and the $\ell_0\ell_2$ -CE. The sparseness is the number of nonzero ECIs in the fit ($\|\mathbf{J}\|_0$). The curves are generated by varying hyperparameters μ_0 , μ_1 , and μ_2 in regularization. (b) ECIs convergence test vs training set size. \mathbf{J} is the ECIs fitted with full training data, and \mathbf{J}_{sub} is the ECIs fitted with a subset of corresponding size.

The optimal trade-off between under-fitting and over-fitting can be found with a CV test, where the optimal model is fitted with the regularization hyper-parameter μ that minimizes the CV error. In our test, a k -fold CV error is used,

$$CV = \sqrt{\frac{1}{k} \sum_{j=1}^k MSE_j}, \quad MSE = \frac{1}{N} \sum_{i=1}^N (E_{DFT}^i - E_{CE}^i)^2, \quad (10)$$

where CV is the cross-validation error averaged over k splits of the validation dataset, and MSE is the mean-squared-error of each validation dataset. Here, N is the size of the validation dataset, and $k = 5$ is the number of folds. In our tests, the regularization hyperparameter μ is selected from the logarithm space between $[10^{-6}, 10^{-1}]$. The sparseness is defined as the number of nonzero elements of the solution ($\|\mathbf{J}\|_0$) and represents the model complexity.

The CV error versus sparseness is presented in Fig. 4(a) for an ℓ_1 and $\ell_0\ell_2$ -norm regularized CE. For the $\ell_0\ell_2$ -CE, the CV error remains low as the sparseness varies between 25 and 150 ECIs. In this regime, the $\ell_0\ell_2$ -CE shows no sign of over-fitting as the CV error remains near the global minimum around 6 meV/atom. The ℓ_1 -CE shows a similar optimal CV error as that of $\ell_0\ell_2$ -CE near this *minimum plateau* regime from 50 to 100 in sparseness. However, as the model complexity changes, the CV error increases at both low and high sparsity,

indicating that the ℓ_1 -CE is less robust against the choice of model complexity. Therefore we conclude that the $\ell_0\ell_2$ -CE can reach low CV error with a lower complexity, which is empirically believed to result in models that better reproduce physics. A more sparse CE can increase the computational speed of energy evaluations and is also less sensitive to model complexity change as compared with the ℓ_1 -CE.

2. (2) Convergence of ECIs with a subset of training data

The second point that we want to emphasize is that the $\ell_0\ell_2$ -CE converges to its most accurate solution faster than the ℓ_1 -CE, which lowers the risk of obtaining an over-fitted result when the configuration space is undersampled. This is an important improvement in the practical use of CE constructions. To test this hypothesis and mimic the iterative sampling process, we designed a numerical experiment based on a finished DFT dataset (with 487 structures in total). Then, we evaluated the quality of fits performed on subsets of training data of increasing size. We subsequently compared the subset-fitted ECIs \mathbf{J}_{sub} with the full-set result. In such a comparison, the ground-truth (full set) solution is set as follows. (1) For the ℓ_1 -CE, the regularization parameter μ_1 is chosen at the minimum CV error according to Fig. 4(a). This solution has 99 nonzero ECIs when all 487 training structures are used in the fitting. (2) For the $\ell_0\ell_2$ -CE, to compare the convergence rate under a similar degree of model complexity, hyperparameters are selected such that the ℓ_1 -CE and $\ell_0\ell_2$ -CE have similar sparsity. The resulting $\ell_0\ell_2$ -CE has 92 nonzero ECIs with all 487 training structures included according to Fig. 4(a).

After setting the hyperparameters for both models, we compared the normalized absolute difference $\|\mathbf{J}_{sub} - \mathbf{J}\|_1 / \|\mathbf{J}\|_1$ between the ℓ_1 -CE (blue line) and $\ell_0\ell_2$ -CE (red line) in Fig. 4(b). For each subset size, ten randomly selected subsets with the same size were evaluated and averaged. The solid square represents the average, and the error bar represents the standard deviation resulting from different subsets. Figure 4(b) indicates that the ℓ_1 -CE demonstrates higher deviation from the ground-truth solution and converges more slowly to it than the $\ell_0\ell_2$ -CE as the training set is increased. This result unambiguously demonstrates the robustness of $\ell_0\ell_2$ -CE to work with small input data sets.

B. ECIs with improved physics

From a general perspective of machine learning (ML), the predictions of energies are made by fitting statistical models on a group of data points. The statistical models can predict the absolute energy with high accuracy by minimizing the cost function, which is constructed by the difference between prediction and observation of the training data. However, in materials science, relative energy quantities are of greater significance than the absolute one (such as energy above the hull, phase diagram and the derivatives of formation energy with respect to the compositional variables). Bartel *et al.* [41] critically examined several ML models for energetics prediction, and found that while the models predict the formation energy (ΔH_f) of materials well, they failed to predict the relative phase stability. Such a dilemma indicates that the prediction error (CV or RMSE) is not the only thing one should consider when constructing a statistical model for the energy.

To demonstrate that the $\ell_0\ell_2$ -CE also leads to a more physically informed solution, we studied a multicomponent system: Li-Mn-Ti-O oxide in an fcc rocksalt framework, with Li^+ - Mn^{2+} - Mn^{3+} - Mn^{4+} - Ti^{4+} -vacancy disorder on the octahedral cation sites and Li^+ - Mn^{2+} - Mn^{3+} -vacancy disorder on the interstitial tetrahedral sites. The Li-Mn-Ti-O composition space contains a number of battery-relevant systems [4,5]. These battery systems are charged and discharged by adding or removing lithium (i.e., lithiation or delithiation) and a charge-compensating electron, which reduces or oxidizes a transition metal. As a result, an important physical property to correctly model in the Li-Mn-Ti-O system is the energetics of Li in octahedral vs. tetrahedral sites. One significant battery-relevant system in which the effects of Li local environment preference are especially presented is the LiMn_2O_4 spinel. When fully lithiated to $\text{Li}_2\text{Mn}_2\text{O}_4$, Li ions occupy octahedral sites while the Li ions occupy tetrahedral sites for compositions $\text{Li}_x\text{Mn}_2\text{O}_4$ when $x \leq 1$.

Thus we design two additional tests to ensure that the CE models well represent the physics of the Li octahedral versus tetrahedral site preferences. Specifically, we compare how well the CE model reproduces: (1) energy differences between the Li in the tetrahedral and octahedral sites in layered MnO_2 and spinel MnO_2 frameworks and (2) a simplified spinel voltage profile against the DFT ground truths. The simplified spinel voltage profile includes the fully lithiated rocksalt-like $\text{Li}_2\text{Mn}_2\text{O}_4$, the spinel LiMn_2O_4 , the commonly seen $\text{Li}_{0.5}\text{Mn}_2\text{O}_4$ ordering, and the fully delithiated Mn_2O_4 and is calculated by taking the average voltage between each set of adjacent orderings. The average voltage is calculated using DFT and the following equation [42–44]:

$$\bar{V}(x_1, x_2) \approx -\frac{E_{\text{Li}_{x_1}\text{Mn}_2\text{O}_4} - E_{\text{Li}_{x_2}\text{Mn}_2\text{O}_4} - (x_1 - x_2)E_{\text{Li}}}{F(x_1 - x_2)}, \quad (11)$$

where x_1 and x_2 are adjacent Li contents with $x_1 > x_2$, E_{Li} is the DFT energy of bcc Li metal, and F is the Faraday constant.

The CE was generated with pair interactions up to 7.1 Å, triplet interactions up to 4.0 Å, and quadruplet interactions up to 3.0 Å based on a primitive cell of the rocksalt structure with lattice parameter $a = 3$ Å. A sinusoid site basis was used [Eq. (2)]. In total, 1475 ECIs (including the constant term J_0) were predefined in the CE Hamiltonian. The dimension of the feature matrix is 1137×1475 . Because of the high compositional dimensionality, the possible number of ECIs within the interaction cutoffs is large. In addition, there are some constraints on the occupancies in the Li-Mn-Ti-O system, such as (1) the total number of Li, transition metals, and vacancies is fixed between octahedral and tetrahedral cation sublattice; (2) the net charge of the system must be neutral, etc. These relations and the inability to sample all possible configurations with DFT reduce the rank of the feature matrix below the dimension [$\text{rank}(\mathbf{\Pi}_S) = 557$], which indicates that a sparse solution is required.

From the test results in Fig. 4, we notice that when the sparseness varies, the variation of the CV error is smaller when the CE is regularized with the $\ell_0\ell_2$ -norm than with the ℓ_1 -norm. This result indicates that $\ell_0\ell_2$ has a hyperparameter space that is larger and more tunable, whereas the ℓ_1 -CE is

more deterministic with a small range of optimal μ_1 obtained by minimizing the CV error. Motivated by this observation, the selection of ECIs for the Li-Mn-Ti-O system was completed as follows.

The regularization strength μ_1 in the ℓ_1 -CE was selected from the stable plateau region when minimizing the CV error in *lasso* (e.g., the μ_1 associated with points between sparseness of 50 to 100 in Fig. 4). For the $\ell_0\ell_2$ -norm, the μ_2 was selected from the stable plateau region by minimizing the CV error in *ridge regression*, similar to what is done for ℓ_1 -CE. After obtaining the optimal μ_2 , the solution for $\ell_0\ell_2$ -CE was further determined by searching μ_0 for a solution with the proper sparseness (at least $\|\mathbf{J}\|_0 < \text{rank}(\mathbf{\Pi}_S)$, $\mu_1, \mu_2, \mu_0 \in [10^{-6}, 10^{-1}]$). For both ℓ_1 -CE and $\ell_0\ell_2$ -CE, several models with low CV error were tested for their ability to well reproduce physical properties, such as minimal violation of DFT ground states in the phase diagram, voltage profile comparison against DFT, as well as the Li-site energy difference between tetrahedral and octahedral occupancy. The best performing models for both ℓ_1 and $\ell_0\ell_2$ are presented in Fig. 5, respectively.

Figure 5(a) presents a comparison of ground-state phase diagrams with the ℓ_1 -CE predictions, $\ell_0\ell_2$ -CE predictions, and DFT calculations. The phase diagrams were generated with in-sample training data (all 1137 structures evaluated with DFT) for both DFT and CE models. We take the DFT phase diagram as the ground truth. In formation-energy prediction, the phase diagram is a key quantity that directly demonstrates the correct physics near the ground states. As the ground states are formed variationally, they are particularly discerning towards spurious ECIs, as the nonphysical noisy interactions often create new ground states leading one to miss the true ground states. Thus a well-reproduced phase diagram is desirable for a CE model. In our tests, the ℓ_1 -CE creates 12 new ground states, indicating that the correct physics in terms of cluster interactions is not well captured. However, the $\ell_0\ell_2$ -CE preserves most of the DFT ground states, with only four spurious “ground states” in the $\ell_0\ell_2$ -CE phase diagram.

The improvement in the physics of the predictions associated with applying the $\ell_0\ell_2$ -norm with hierarchy constraints is further demonstrated by the voltage profile and Li-occupancy energy. In Fig. 5(b), the voltage profiles generated by prediction using the ℓ_1 -CE and $\ell_0\ell_2$ -CE (blue lines) are compared with those from DFT (orange lines), taken as the ground truth. We can see that the ℓ_1 -CE incorrectly predicts the voltage plateau between $x = 0.5$ to 1 in the $\text{Li}_x\text{Mn}_2\text{O}_4$ spinel-like structure such that the $x = 0.5$ configuration is no longer stable (the voltage between $x = 0.5$ and 1.0 is higher than that between $x = 0.0$ and 0.5). In contrast, the $\ell_0\ell_2$ -CE matches very well with the DFT-generated voltage profiles. The erroneous predictions of the ℓ_1 -CE are further confirmed by the Li-occupancy energy. In Fig. 5(c), the energy difference between Li in octahedral and tetrahedral occupancy was evaluated in the layered- MnO_2 and spinel- MnO_2 frameworks. The absolute error compared with DFT is 0.52 eV (layered) and 0.18 eV (spinel) for the ℓ_1 -CE, whereas that for the $\ell_0\ell_2$ -CE is 0.09 eV (layered) and 0.09 eV (spinel), respectively. A significant reduction of prediction error is observed with the $\ell_0\ell_2$ -norm regularized CE.

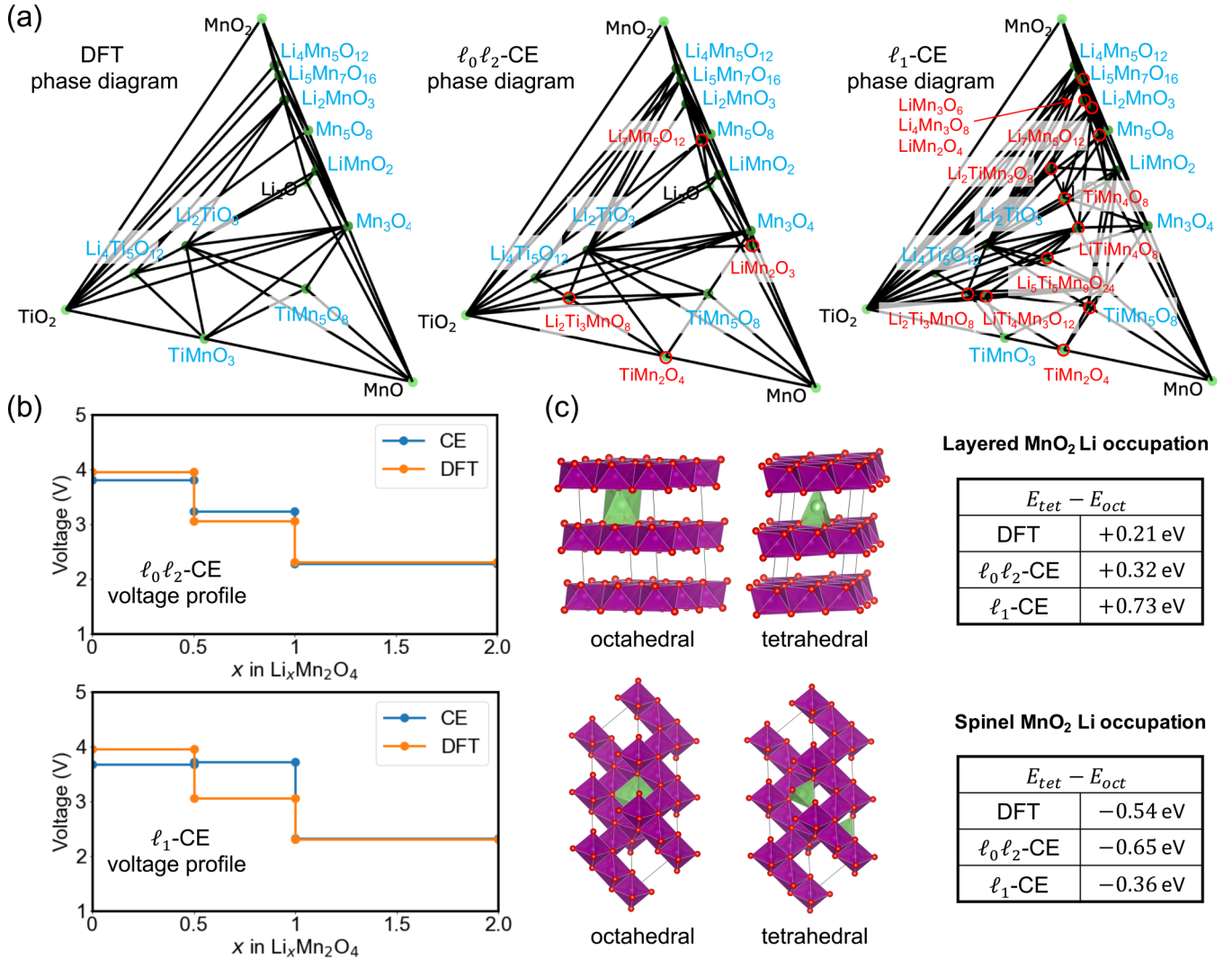


FIG. 5. (a) Phase diagram generated with DFT, $\ell_0\ell_2$ -CE, and ℓ_1 -CE. The DFT ground states are labeled in blue text. The incorrectly predicted ground states are labeled with red circles and text. (b) The simplified spinel voltage profile (blue line) generated by ℓ_1 -CE and $\ell_0\ell_2$ -CE for spinel orderings in Li_xMn₂O₄ is compared with the DFT ground-truths (orange line). (c) Energy difference of Li occupation in octahedral and tetrahedral sites in layered MnO₂ (top) and spinel MnO₂ framework (bottom).

IV. DISCUSSION AND SUMMARY

In two complex oxide systems, we showed that the $\ell_0\ell_2$ -CE with hierarchy constraints outperforms the conventional ℓ_1 -CE in terms of sparseness against CV error, convergence rate with a subset of training-data, and some critical physical quantities in Li intercalation materials. More generally, the optimization of the ECIs is not deterministic within a single method, and the successful construction of a CE model typically relies on two aspects: (1) choosing a valid interaction space by truncating the clusters or orbits and (2) applying a proper optimization algorithm to obtain the ECIs. The results in this paper show that for the second step, the $\ell_0\ell_2$ -norm method is the superior choice for a robust and physical solution compared to the conventional ℓ_1 method.

We note that one limitation of the $\ell_0\ell_2$ -norm method in the MIQP paradigm is the computational efficiency. As solving the ℓ_0 -norm is an NP-hard problem, more computational time is required to solve the MIQP when more predefined ECIs

are included. The $\ell_0\ell_2$ -CE works well for relatively small or well-predefined systems [$\dim(\Pi_S) \leq 2000$]. Therefore the most applicable way to use $\ell_0\ell_2$ -norm regularized CE with hierarchy constraints is likely to be as follows: (1) define a CE within a relatively small cutoff and truncate to quadruplet or quintuplet clusters at most [ideally staying within $\dim(\Pi_S) \leq 2000$] and (2) follow the procedure described in Fig. 3(b) to determine the optimal hyperparameter to obtain the ECIs. However, we note that $\dim(\Pi_S) \leq 2000$ applies to virtually all known published CE.

To obtain a model that represents the physics of a system well, the relative difference of energies between configurations is of greater significance than the absolute ones. In ordinary least-squares fitting, the cost function only focuses on the global averaged error of the training set, which leads to over-fitting. Adding regularization of the ECIs can alleviate this issue by constraining the optimization space of parameters, but our results show that not all regularization creates physically meaningful solutions. We propose that it is

beneficial to include the physically inspired constraints into the design of the cost function, such as adding hierarchy constraints with $\ell_0\ell_2$ -norm implementation. The $\ell_0\ell_2$ -CE can improve the physical meaning of the solution and break the correlation between coupled clusters, which is achieved by directly penalizing the number of nonzero ECIs for feature selection and enforcing hierarchy relations between ECIs via the slack variables in the MIQP paradigm. The $\ell_0\ell_2$ -CE gives an estimation of the ECIs with reasonable physics near the ground-states, but does not strictly enforce the preservation of ground states. Additionally, the ground-state preservation can be further achieved by adding inequality constraints on the energies into Eq. (9) as shown in previous work [19].

In summary, our method sheds light on how to obtain good ECIs for simulations in complex and coupled multicomponent systems, with several proposed criteria in ECIs optimization: (1) minimize the CV error under general regression level (e.g., ridge regression); (2) as the sparseness describes the complexity of and number of independent variables, the sparseness of the solution shall be improved (reduced) with reasonable in-sample training error; and (3) check the near-ground-state behavior and related physics for the optimal ECIs selection.

ACKNOWLEDGMENTS

This work was funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under Contract No. DE-AC0205CH11231 (Materials Project program KC23MP). The work was also supported with computational resources provided by the Extreme Science and Engineering Discovery Environment (XSEDE), supported by National Science Foundation Grant No. ACI1053575; 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; the Center for Functional Nanomaterials (CFN), which is a U.S. Department of Energy Office of Science User Facility, at Brookhaven National Laboratory under Contract No. DE-SC0012704; and the Lawrence computational cluster resource provided by the IT Division at the Lawrence Berkeley National Laboratory.

APPENDIX: DFT CALCULATIONS

DFT calculations were performed with the *Vienna ab initio simulation package* (VASP) using the projector-augmented wave method [45,46], a plane-wave basis set with an energy cutoff of 520 eV, and a reciprocal space discretization of 25 *k* points per Å. All the calculations were converged to 10^{-6} eV in total energy for electronic loops and 0.02 eV/Å in interatomic forces for ionic loops. In the LiF-MnO-LiVO₂-Li₂TiO₃ system, we used the Perdew-Burke-Ernzerhof (PBE) generalized gradient approximation exchange-correlation functional [47] with rotationally averaged Hubbard *U* correction (GGA + *U*) to compensate for the self-interaction error on all transition-metal atoms except titanium [48]. The *U* parameters were obtained from the literature, where they were calibrated to transition-metal oxide formation energies (3.9 eV for Mn and 3.1 eV for V). The GGA + *U* computational framework is believed to be reliable in determining the formation enthalpies of similar compounds [49]. In the Li-Mn-Ti-O oxide system, the strongly constrained and appropriately normed (SCAN) meta-GGA exchange-correlation functional was used [50]. The SCAN functional is believed to have better performance at capturing charge transfer due to better redox and atomic coordination prediction [51,52], which would improve the accuracy of energetics involving introducing vacancies on octahedral and interstitial tetrahedral sublattices in the rocksalt framework.

-
- [1] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder *et al.*, *APL Mater.* **1**, 011002 (2013).
 - [2] R. Zhang, S. Zhao, J. Ding, Y. Chong, T. Jia, C. Ophus, M. Asta, R. O. Ritchie, and A. M. Minor, *Nature (London)* **581**, 283 (2020).
 - [3] S. Yin, Y. Zuo, A. Abu-Odeh, H. Zheng, X.-G. Li, J. Ding, S. P. Ong, M. Asta, and R. O. Ritchie, *Nat. Commun.* **12**, 4873 (2021).
 - [4] 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, *Nat. Commun.* **10**, 592 (2019).
 - [5] B. Ouyang, N. Artrith, Z. Lun, Z. Jadidi, D. A. Kitchaev, H. Ji, A. Urban, and G. Ceder, *Adv. Energy Mater.* **10**, 1903240 (2020).
 - [6] 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, *Nat. Mater.* **20**, 214 (2021).
 - [7] P. Zhong, Z. Cai, Y. Zhang, R. Giovine, B. Ouyang, G. Zeng, Y. Chen, R. Clément, Z. Lun, and G. Ceder, *Chem. Mater.* **32**, 10728 (2020).
 - [8] J. Sanchez, F. Ducastelle, and D. Gratias, *Physica A* **128**, 334 (1984).
 - [9] J. M. Sanchez, *Phys. Rev. B* **99**, 134206 (2019).
 - [10] P. D. Tapesch, G. D. Garbulsky, and G. Ceder, *Phys. Rev. Lett.* **74**, 2272 (1995).
 - [11] A. Seko, Y. Koyama, and I. Tanaka, *Phys. Rev. B* **80**, 165122 (2009).
 - [12] X. Zhang and M. H. F. Sluiter, *J. Phase Equilibria Diff.* **37**, 44 (2016).
 - [13] L. Barroso-Luque, J. H. Yang, and G. Ceder, *Phys. Rev. B* **104**, 224203 (2021).
 - [14] A. van de Walle, *Calphad* **33**, 266 (2009).
 - [15] G. Ceder, *Comput. Mater. Sci.* **1**, 144 (1993).
 - [16] W. Huang, D. A. Kitchaev, S. T. Dacek, Z. Rong, A. Urban, S. Cao, C. Luo, and G. Ceder, *Phys. Rev. B* **94**, 134424 (2016).
 - [17] D. De Fontaine, *Solid State Phys.* **34**, 73 (1979).
 - [18] M. Ångqvist, W. A. Muñoz, J. M. Rahm, E. Fransson, C. Durniak, P. Rozyczko, T. H. Rod, and P. Erhart, *Adv. Theory Simulations* **2**, 1900015 (2019).
 - [19] W. Huang, A. Urban, Z. Rong, Z. Ding, C. Luo, and G. Ceder, *npj Comput. Mater.* **3**, 30 (2017).

- [20] D. B. Laks, L. G. Ferreira, S. Froyen, and A. Zunger, *Phys. Rev. B* **46**, 12587 (1992).
- [21] V. Blum and A. Zunger, *Phys. Rev. B* **70**, 155108 (2004).
- [22] T. Mueller and G. Ceder, *Phys. Rev. B* **80**, 024103 (2009).
- [23] L. J. Nelson, V. Ozolinš, C. S. Reese, F. Zhou, and G. L. W. Hart, *Phys. Rev. B* **88**, 155105 (2013).
- [24] E. Cockayne and A. van de Walle, *Phys. Rev. B* **81**, 012104 (2010).
- [25] E. J. Candes and M. B. Wakin, *IEEE Signal Processing Magazine* **25**, 21 (2008).
- [26] L. J. Nelson, G. L. W. Hart, F. Zhou, and V. Ozolinš, *Phys. Rev. B* **87**, 035125 (2013).
- [27] A. Seko, A. Takahashi, and I. Tanaka, *Phys. Rev. B* **90**, 024101 (2014).
- [28] Z. Leong and T. L. Tan, *Phys. Rev. B* **100**, 134108 (2019).
- [29] M. Elad, in *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, edited by M. Elad (Springer, New York, NY, 2010), pp. 35–54
- [30] W. Huang, A. Urban, P. Xiao, Z. Rong, H. Das, T. Chen, N. Artrith, A. Toumar, and G. Ceder, [arXiv:1807.10753](https://arxiv.org/abs/1807.10753).
- [31] Gurobi Optimization, LLC, Gurobi Optimizer Reference Manual (2021).
- [32] M. H. F. Sluiter and Y. Kawazoe, *Phys. Rev. B* **71**, 212201 (2005).
- [33] N. A. Zarkevich and D. D. Johnson, *Phys. Rev. Lett.* **92**, 255702 (2004).
- [34] A. van de Walle and G. Ceder, *J. Phase Equilib.* **23**, 348 (2002).
- [35] R. Tibshirani, *J. R. Stat. Soc.: Series B (Methodological)* **58**, 267 (1996).
- [36] <https://github.com/CederGroupHub/smol>, <https://github.com/CederGroupHub/sparse-lm>.
- [37] R. J. Clément, Z. Lun, and G. Ceder, *Energy Environ. Sci.* **13**, 345 (2020).
- [38] J. Huang, P. Zhong, Y. Ha, D.-h. Kwon, M. J. Crafton, Y. Tian, M. Balasubramanian, B. D. McCloskey, W. Yang, and G. Ceder, *Nature Energy* **6**, 706 (2021).
- [39] A. Seko and I. Tanaka, *J. Phys.: Condens. Matter* **26**, 115403 (2014).
- [40] W. D. Richards, S. T. Dacek, D. A. Kitchaev, and G. Ceder, *Adv. Energy Mater.* **8**, 1701533 (2018).
- [41] C. J. Bartel, A. Trewartha, Q. Wang, A. Dunn, A. Jain, and G. Ceder, *npj Comput. Mater.* **6**, 97 (2020).
- [42] A. Urban, D.-H. Seo, and G. Ceder, *npj Comput. Mater.* **2**, 16002 (2016).
- [43] D.-H. Seo, A. Urban, and G. Ceder, *Phys. Rev. B* **92**, 115118 (2015).
- [44] M. Aydinol and G. Ceder, *J. Electrochem. Soc.* **144**, 3832 (1997).
- [45] G. Kresse and J. Furthmüller, *Comput. Mater. Sci.* **6**, 15 (1996).
- [46] G. Kresse and D. Joubert, *Phys. Rev. B* **59**, 1758 (1999).
- [47] J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).
- [48] L. Wang, T. Maxisch, and G. Ceder, *Phys. Rev. B* **73**, 195107 (2006).
- [49] A. Jain, G. Hautier, S. P. Ong, C. J. Moore, C. C. Fischer, K. A. Persson, and G. Ceder, *Phys. Rev. B* **84**, 045115 (2011).
- [50] J. Sun, A. Ruzsinszky, and J. P. Perdew, *Phys. Rev. Lett.* **115**, 036402 (2015).
- [51] J. H. Yang, D. A. Kitchaev, and G. Ceder, *Phys. Rev. B* **100**, 035132 (2019).
- [52] Y. Zhang, D. A. Kitchaev, J. Yang, T. Chen, S. T. Dacek, R. A. Sarmiento-Pérez, M. A. Marques, H. Peng, G. Ceder, J. P. Perdew *et al.*, *npj Comput. Mater.* **4**, 9 (2018).