npj | computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

# Approaches for handling high-dimensional cluster expansions of ionic systems

Julia H. Yang[1,2], Tina Chen[1,2], Luis Barroso-Luque[1], Zinab Jadidi[1,2] and Gerbrand Ceder[1,2✉]

Disordered multicomponent systems attract great interest due to their engineering design flexibility and subsequent rich space of properties. However, detailed characterization of the structure and atomic correlations remains challenging and hinders full navigation of these complex spaces. A lattice cluster expansion is one tool to obtain configurational and energetic resolution. While in theory a cluster expansion can be applied to any system of any dimensionality, the method has primarily been used in binary systems or ternary alloys. Here we apply cluster expansions in high-component ionic systems, setting up the largest cluster expansion ever attempted to our knowledge. In doing so, we address and discuss challenges specific to high-component ionic systems, namely charge state assignments, structural relaxations, and rank-deficient systems. We introduce practical procedures to make the fitting and analysis of complex systems tractable, providing guidance for future computational studies of disordered ionic systems.

*npj Computational Materials* (2022)8:133 ; https://doi.org/10.1038/s41524-022-00818-3

## INTRODUCTION

Disordered and partially disordered systems can contain a relatively high number of components. In recent years, active research in these high-component disordered systems has spanned a range of breakthrough technologies[1,2], including high-temperature, strong, and lightweight high-entropy alloys[3], superionic lithium conductors[4], ultra-high temperature ceramics for structural applications in extreme environments[5], and sustainable battery design with improved performance[6,7].

It is known that local configurations are important for some materials properties. For instance, in multi-principal element entropy alloys (MPEAs), magnetic interactions can drive atomic orderings which explain otherwise anomalous material properties[8]. Given the challenges in modeling multicomponent alloys[9], coarse-grained Hamiltonians such as the cluster expansion (CE) approach have been remarkably useful, leading to the discovery of hierarchical ground state orderings[10], prediction of configurational energetics[11], and generation of mesoscale phase-field models[12].

The CE approach which maps the configurational problem in a crystalline solid on that of a lattice model has been used in pseudo-binary and ternary ionic systems to predict solid state phase diagrams in the CaO–MgO system[13], understand fluorine solubility[14], observe lithium (Li)-gettering in fluorinated cathodes[15], and characterize short-range-order[16]. In this work, we apply the CE approach to study a new class of partially disordered spinel (PDS) materials which exhibit ultrahigh energy and power density[17] and discuss new challenges specific to high-component ionic CE. Since ionic systems have greater formation energy than do metallic systems, prediction errors tend to be higher[18], motivating methodology studies such as this one to aid in reducing sources of error and developing predictive CE models. We first explain the complexity of PDS and summarize the theory of multicomponent, multi-sublattice systems introduced in detail elsewhere[19,20]. Next, we discuss ab-initio data generation and preparation specific to ionic CEs, namely species charge assignments and structural relaxations that maintain sublattice topology.

We then introduce new methods for fitting the CE by grouping the thousands of possible effective cluster interactions which are the expansion coefficients of the basis functions that describe the configurational arrangement. We demonstrate how to group site interactions required to address the compositional constraints arising from the charge neutrality requirement in ionic systems. Rank deficiency problems occur within groups of basis functions on the same lattice figure because it is not possible to sample all configurations with ab-initio calculations. We handle this by applying sparse group lasso regularization when the energetics of unsampled configurations is represented in lower-order features. Finally, we show that models of high-component systems are prone to higher errors compared to models of lower-dimensional systems which have been well-explored, and bring the new perspective that model predictability should instead scale with configuration space size.

## BACKGROUND: MOTIVATION TO USE CE TO STUDY HIGH-COMPONENT IONIC SYSTEMS

Our work was inspired by a new class of Li-Mn-oxyfluorides in the PDS structure, which have demonstrated ultrahigh power and energy density in Li-ion batteries, delivering over 900 Wh kg$^{-1}$ [17,21]. These materials are approximately based on an $AB_2X_4$ spinel structure which consists of a face-centered cubic (FCC) anion (X) framework with half the octahedral sites occupied by metal B (the “16d” sites) and the other half (“16c” sites) unoccupied. A small number of tetrahedral sites (“8a” sites) are occupied by the A metal. It is the requirement that these occupied tetrahedral sites have no face-sharing with octahedral sites that creates the 16d/16c cation ordering on the octahedral sites. PDS is a significant departure from this classic stoichiometric spinel both because it has a higher cation/anion ratio than spinel and is partially disordered. For example, the PDS compound of Ji et al.[17] has stoichiometry $Li_{1.68}Mn_{1.6}O_{3.7}F_{0.3}$ with 84% of the Mn in 16d sites and 16% of the Mn in 16c sites. Li, which in $LiMn_2O_4$ would

[1]Department of Materials Science and Engineering, Berkeley, CA 94720, USA. [2]Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. ✉email: gceder@berkeley.edu

solely occupy the 8a site fills only 52% of the tetrahedral 8a sites, 30% of the octahedral 16d sites, and 28% of the octahedral 16c sites. Note that in PDS the cation/anion ratio is higher than in a stoichiometric spinel, where it is 3/4.

The structure of the PDS is challenging to understand because the cation-excess space removes baseline understanding of how cations can be arranged in the structure. For instance, it is unclear how the Mn occupancy of the 16c site affects the occupancy of the nearest neighbor 8a sites with which it is face-sharing, which is usually not preferable in oxides. Because the 8a sites form a percolating transport channel for Li one would expect their blockage to lead to poor Li transport, but this is contrary to what is observed experimentally.

In principle, a well-parametrized configurational CE would enable equilibration of the local structure in the system with Monte Carlo (MC) techniques, as is done to identify chemical short-range order[22–24], compute phase diagrams[25–27], and find ground states[26,28,29]. We show that in practice, simple assumptions and typical approaches to obtain the CE are difficult for this system with such high configurational degrees of freedom. For the PDS materials, which have the stoichiometries $Li_{1.68}Mn_{1.6}O_{3.7}F_{0.3}$ and $Li_{1.68}Mn_{1.6}O_{3.4}F_{0.6}$, the anion FCC lattice hosting binary disorder ($O^{2-}$, $F^{-}$) forms two types of symmetrically distinct cation sites with different allowed species on them: an octahedral site with quinary disorder taken from the space of ($Li^{+}$, $Mn^{2+}$, $Mn^{3+}$, $Mn^{4+}$, Vacancy) and a tetrahedral site with ternary disorder among ($Li^{+}$, $Mn^{2+}$, Vacancy). Without symmetrizing, the configuration space in a primitive cell has dimension 90, obtained by taking the product of all site spaces from the anion, octahedral, and two tetrahedral sites[30]. It is clear that this CE transcends the usual complexity of CE models which are typically done for dimension two or three, resulting from one site space with binary or ternary disorder[31,32].

## Background: introduction to CE theory

We provide only a brief introduction to the mathematics of CE, referring the reader to classic works by Sanchez, Ducastelle, and Gratias[19] and van de Walle[20] for a comprehensive explanation of multi-component CE. For multi-sublattice ionic CE, we refer the reader to work by Tepesch, Garbulsky, and Ceder[30] and our recent review[33].

The CE approach assumes an underlying well-defined set of sites ("the lattice") over which species can distribute. The lattice can be partitioned into "sublattices" with different allowed species decorations. For instance, in ionic systems there are typically at least two such sublattices: one for the cation species and another for the anion species. Here, the terminology "lattice" is used in a broader sense than in crystallography where in the strictest sense of the term it only refers to the Bravais lattice of a structure.

The basic principle of the CE is that a relaxed DFT structure is represented by an occupation string $\boldsymbol{\sigma}$ which describes exactly which species occupy each site on all sublattices. A CE representation of the energy is possible as long as this mapping between a DFT-relaxed structure and occupation string $\boldsymbol{\sigma}$ is one-to-one. This distinction is necessary because a lattice cluster expansion model cannot capture the exact spatial positions of atoms. Rather, it strictly specifies the decoration $\boldsymbol{\sigma}$ of a lattice.

Any scalar extensive quantity $q$ can be represented as a function of its decoration $\boldsymbol{\sigma}$ as:

$$q(\boldsymbol{\sigma}) = \sum_{\beta} m_{\beta} J_{\beta} \langle \Phi_{\alpha}(\boldsymbol{\sigma}) \rangle_{\beta} \tag{1}$$

where $\beta$ are symmetrically distinct groupings of site basis functions on the lattice. The cluster $\alpha$ is a multi-index array with entries which label the corresponding single-site basis functions. $m_{\beta}$ is the number of clusters $\alpha$ equivalent by symmetry in whatever normalizing unit scalar $q$ is taken (e.g. per cell, per site, etc.), $J_{\beta}$ is the effective cluster interaction (ECI), and $\Phi_{\alpha}$ is the cluster function. The 48 symmetry operations for the CE lattice are four $C_3$ rotations, three $C_4$ rotations, and an inversion, corresponding to the point group $m\bar{3}m$. Lastly, the average of cluster functions evaluated over a crystal is the correlation function and the concatenation of all correlation functions is referred to as the correlation vector.

As an example, we demonstrate the construction of a single cluster function $\Phi_{\alpha}$ and evaluate it for a LiF structure using the orthogonal sinusoidal basis[20]. The $n$ number of site basis functions, indexed from $a_j = 0, \ldots, n-1$, for a single site $\sigma_i$ with $n_i$ possible species are:

$$\phi_{a_j,n_i}(\sigma_i) = \begin{cases} 1, & \text{if } a_{\mathrm{j}} = 0 \\ -\cos\left(\frac{2\pi \left\lceil \frac{a_j}{2} \right\rceil \sigma_i}{n_i}\right), & \text{if } a_{\mathrm{j}} > 0 \text{ and odd} \\ -\sin\left(\frac{2\pi \left\lceil \frac{a_j}{2} \right\rceil \sigma_i}{n_i}\right), & \text{if } a_{\mathrm{j}} > 0 \text{ and even} \end{cases} \tag{2}$$

Given a set of single-site basis functions $\{\phi_{a_j,n_i}\}$, the cluster function in Eq. (3) is the tensor product of the $n_i$ single-site basis functions on each possible site in $\boldsymbol{\sigma}$:

$$\Phi_{\alpha}(\boldsymbol{\sigma}) = \prod_{i=1}^{N} \phi_{\alpha_i,n_i}(\sigma_i) \tag{3}$$

The product is a "cluster-like"[19] because only the occupancies on which the site function is not equal to the constant "1" are relevant.

To be explicit, we write the cluster function for a specific octahedral-tetrahedral "geometric cluster". (A geometric cluster is strictly a set of crystallographic sites, whereas a cluster $\alpha$ is, in full technicality, a cluster of functions. However, for simplicity we refer to $\alpha$ as a cluster.) We then evaluate the cluster function for the occupancy string of $Li_1F_1$, which is $\boldsymbol{\sigma}_{LiF} = [\sigma_1 = 0, \sigma_2 = 2, \sigma_3 = 2, \sigma_4 = 1]$ because the species on the octahedral (site 1), tetrahedral (sites 2 and 3), and anion (site 4) are ["Li+", "Vacancy", "Vacancy", and "F-"]. In this example, we have chosen the site variables for an octahedral Li, a tetrahedral vacancy, and an anion fluorine to be 0, 2, and 1 respectively, but other site variables can be chosen.

The cluster function for sites 1 and 2 is:

$$\begin{aligned} \Phi_{\alpha}(\boldsymbol{\sigma}) = \Bigg[ 1, -\cos\left(\frac{2\pi \left\lceil \frac{1}{2} \right\rceil \sigma_1}{5}\right), -\sin\left(\frac{2\pi \left\lceil \frac{2}{2} \right\rceil \sigma_1}{5}\right), \\ -\cos\left(\frac{2\pi \left\lceil \frac{3}{2} \right\rceil \sigma_1}{5}\right), -\sin\left(\frac{2\pi \left\lceil \frac{4}{2} \right\rceil \sigma_1}{5}\right) \Bigg] \otimes \Bigg[ 1, -\cos\left(\frac{2\pi \left\lceil \frac{1}{2} \right\rceil \sigma_2}{3}\right), -\sin\left(\frac{2\pi \left\lceil \frac{2}{2} \right\rceil \sigma_2}{3}\right) \Bigg] \end{aligned} \tag{4}$$

This tensor product yields a basis set for the geometric cluster comprising site 1 and site 2. The basis functions for sites 1 and 2 are indexed as $(a_j, a_{j'})$, where $a_j$ indexes the basis functions for the octahedral site (with $n_i = 5$) and $a_{j'}$ indexes the basis functions for the tetrahedral site (with $n_i = 3$). So, this set of multi-indices for cluster $\alpha$ is the Cartesian product of basis function indices, specifically: $(a_j, a_{j'}) \in \{(1, 1), (2, 1), (3, 1), (4, 1), (1, 2), (2, 2), (3, 2), (4, 2)\}$. The relevant set of basis functions has contracted multi-indices, meaning that all labels that are 0 (i.e. not part of the cluster in general) are dropped. When translational symmetry is included as well, these contracted multi-indices make up a set we call **B**, which is the set of symmetrically distinct orbits $\boldsymbol{\beta}$. We will demonstrate the use of **B** in our regularization scheme later on.

Using Eq. (4), we calculate that $\Phi_{a_j,a_{j'}=(1,1)}(\sigma_{LiF}) = -0.5$. The entire set of correlation functions for the set of contracted multi-indices are then: [−0.5, 0, −0.5, 0, −0.866, 0, −0.866, 0].

## RESULTS: HIGH-COMPONENT CE IN OXIDES: EXAMPLE OF PARTIALLY DISORDERED SPINELS (PDS)

Besides the CE to capture energy configurational energy dependence of the PDS, we also add an explicit term to capture the electrostatic energy in Ewald form[34]. The Ewald summation is a technique to efficiently sum up long-range electrostatic interactions and their periodic images and is a sum of the direct space, constant, and reciprocal space terms. The proportionality constant for this term is also fitted and can be thought of as representing the dielectric constant. We use Pymatgen[35] to calculate the total Ewald energy. Lastly, we apply a form of the structure inversion method proposed by Connolly and Williams[36] to determine the ECI and dielectric constant for the Ewald energy by fitting to DFT energies.

In building our CE, we consider relevant geometric clusters arising from multi-body interactions within a certain distance from one another. Our geometric clusters consist of pairs of sites less than 7 Å apart, triplets with points less than 5 Å apart, quadruplets 4 Å apart, and quintuplets 3 Å apart. Given that our lattice model fixes the nearest-neighbor octahedral-tetrahedral cation distance to 1.82 Å; the octahedral-anion bond length to 2.1 Å, and the tetrahedral-anion bond length to 1.82 Å, the correlation vector for a given structure has length 4587. Adding the Ewald energy adds one more dimension to our feature vector, resulting in a total length of 4588.

We use an in-house developed Python package, Statistical Mechanics On Lattices (smol), to generate the correlation vectors on the orthogonal sinusoidal basis in Eq. 2. Because of the large number of possible ECI in these high-component systems, even when limiting their interaction range to 7 Å, the fitting of cluster interactions to DFT energies always starts off as an under-determined system because the number of DFT-relaxed structures used as training data will be fewer than the number of ECI. Well-known statistical tools based on regularized regression exist to handle model generation in under-determined systems: lasso[37], group lasso[38], and sparse group lasso (SGL)[39] all techniques which we will discuss later.

### Result: data preparation – automatic, optimized charge assignments

In ionic systems, the same ion can behave differently in terms of their size, site coordination preference, or local interactions when it has a different formal valence. For instance, crystal field effects lead to a strong preference for $Mn^{2+}$ to be tetrahedral, which is not observed for $Mn^{3+}$ or $Mn^{4+}$. This site and interaction preference cannot easily be captured when all Mn ions are treated as the same "Mn" species, as would be done in the CE of metallic systems, and therefore different charge states of Mn ions must be treated as different species. Prior work in ionic CE have also explicitly treated these charge states[40,41]. In this section, we describe how to optimally assign charge states to ions from electronic structure data. The details of DFT calculations are provided in Methods.

The charge density around a transition metal ion itself is often remarkably invariant with respect to the formal valence[42,43], due to the hybridization shift with the anion that takes places when an electron is removed from the metal[44,45]. For example, total charge density integration upon Li insertion in $\lambda$-$Mn_2O_4$ (to spinel $LiMn_2O_4$) reveals greater charge-transfer to the oxygen anion, in that, upon Li insertion the Mn ion gains 0.136 electronic charge per electron, whereas the oxygen accepts 0.171 electronic charge[43]. Thus, there is a strong electron exchange with oxygen and for this reason magnetic moments have instead been found to be a much better guidance for the formal valence of an ion[46].

We use the magnetic moment arising from *d*-orbital contributions to identify Mn charge states. These magnetic moments are obtained by integrating the local (spin up minus spin down) moments in a sphere around each Mn atom. Charge assignment is non-trivial because the moment distribution around Mn ions varies depending on its environment. For instance, we find that in $MnF_3$ and $Mn_2O_3$ the magnetic moments for $Mn^{3+}$ are 3.770 $\mu_B$ and 3.797 $\mu_B$ respectively, reflecting little difference between a F and a O environment for $Mn^{3+}$. Yet, in $Mn_3OF_5$, which contains $Mn^{2+}$ and $Mn^{3+}$, the moment on $Mn^{3+}$ is 4.077 $\mu_B$ which is significantly higher than in $MnF_3$ and $Mn_2O_3$. (The moments on the two $Mn^{2+}$, 4.351 $\mu_B$ and 4.393 $\mu_B$, are clearly different from that on $Mn^{3+}$.) Evidently, knowing the Mn moments in the pure oxide and pure fluorine reference states is not enough to assign charges in mixed-valence Mn-oxyfluoride compositions.

The cation configurations may also influence the magnetic moment distribution in non-obvious ways. To see this, we provide the Mn moments for three different polymorphs of $Li_6Mn_4O_{10}$, in which the average Mn oxidation state is 3.5+, in Table 1, along with their nearest neighbor (NN) cation environments. Table 1 shows that all three polymorphs of $Li_6Mn_4O_{10}$ are assigned to be "charge-balanced" if appropriate differentiation between moments for $Mn^{3+}$ and $Mn^{4+}$ is made. The magnetic moments for high-spin $Mn^{3+}$ and $Mn^{4+}$ are expected to be 4 $\mu_B$ ($t_{2g}{}^3e_g{}^1$) and 3 $\mu_B$ ($t_{2g}{}^2e_g{}^1$), which are reasonably represented in Polymorphs A and C, given that moments in reality are lower on the metal center since the surrounding oxygen hybridizes and shares some of the magnetic moment[45,47,48].

**Table 1.** Description of three polymorphs of composition $Li_6Mn_4O_{10}$.

| DFT structure information | | | | Cation environment around metal, i.e. number of nearest neighbor $Li^+$, $Mn^{3+}$, $Mn^{4+}$ | | | |
|---|---|---|---|---|---|---|---|
| Polymorph | $Mn^{3+}$ moments | $Mn^{4+}$ moments | Energy above hull (meV/atom) | $Mn^{3+}$ | $Mn^{3+}$ | $Mn^{4+}$ | $Mn^{4+}$ |
| A | 3.501, 3.504 | 2.732, 3.005 | 21.7 | 8 $Li^+$<br>2 $Mn^{3+}$<br>2 $Mn^{4+}$ | 8 $Li^+$<br>2 $Mn^{3+}$<br>2 $Mn^{4+}$ | 8 $Li^+$<br>2 $Mn^{3+}$<br>2 $Mn^{4+}$ | 8 $Li^+$<br>2 $Mn^{3+}$<br>2 $Mn^{4+}$ |
| B | 3.232, 3.232 | 3.169, 3.169 | 45.9 | 8 $Li^+$<br>1 $Mn^{3+}$<br>3 $Mn^{4+}$ | 8 $Li^+$<br>1 $Mn^{3+}$<br>3 $Mn^{4+}$ | 7 $Li^+$<br>3 $Mn^{3+}$<br>2 $Mn^{4+}$ | 7 $Li^+$<br>3 $Mn^{3+}$<br>2 $Mn^{4+}$ |
| C | 3.442, 3.629 | 2.724, 3.027 | 113.5 | 8 $Li^+$<br>1 $Mn^{3+}$<br>3 $Mn^{4+}$ | 8 $Li^+$<br>1 $Mn^{3+}$<br>3 $Mn^{4+}$ | 8 $Li^+$<br>2 $Mn^{3+}$<br>2 $Mn^{4+}$ | 6 $Li^+$<br>4 $Mn^{3+}$<br>2 $Mn^{4+}$ |

The DFT structure information for the three polymorphs is Mn *d*-orbital magnetic moments and energy above hull per atom and the valence is classified by the Bayesian optimization model. The 12 edge-sharing nearest neighbor cations, identified using the Bayesian optimization assignments, are also described for each polymorph.

Within the $Mn^{4+}$ environments, we observe that the magnetic moments on the $Mn^{4+}$ ion are relatively rigid when there are six or eight surrounding $Li^+$ in polymorphs A and C. This is because the moment is always about 2.7 or 3.0 $\mu_B$ even when the surrounding environment is more Mn-rich as seen in polymorph C. However, the $Mn^{4+}$ moment can be higher (3.169 $\mu_B$) if the environment around $Mn^{4+}$ has seven $Li^+$, as seen in polymorph B.

Within the $Mn^{3+}$ environments, the effects are even less clear: Having eight surrounding $Li^+$ is associated with having a range of moments: from 3.232 $\mu_B$ to 3.629 $\mu_B$. When $Mn^{3+}$ is surrounded by equal $Mn^{3+}$ and $Mn^{4+}$ the moment is about 3.5 $\mu_B$ (polymorph A), but a more oxidized environment around $Mn^{3+}$ can lead to either lower (3.232 $\mu_B$ in polymorph A) or higher (3.629 $\mu_B$ in polymorph C) moments. It may be necessary to know details of how the NN cations are arranged around $Mn^{3+}$ to systematically understand how the moment is distributed.

Lastly, as a final indication of the effects which can influence magnetic moment distribution, we observe that Polymorph B does not have as well-separated magnetic moments, indicating some degree of self-interaction error which could be reduced by applying a Hubbard *U* correction[49].

Given hundreds of relaxed DFT structures with moments arising from the various effects described (chemical, configurational, and remnant self-interaction), the challenge is to find an optimal solution for differentiating among $Mn^{2+}$, $Mn^{3+}$, and $Mn^{4+}$. Our approach is to use Bayesian optimization via Gaussian Processes[50] and assign charges to moments via some black box mapping function *f* under the condition of maximizing the total number of charge-neutral DFT structures. Black box optimization is particularly useful in this situation where each set of magnetic moments is computationally expensive, and the exact form of *f* is neither known nor necessarily differentiable[51]. We formulate *f* to depend on three magnetic moment upper cutoffs (corresponding to upper cutoffs for the three Mn valence states) that determine the charge for each Mn atom. The solution which minimizes the loss, the sum of the absolute value of each structure's charge, is the final solution. We apply the Bayesian Optimization module in scikit-learn[52] to charge-balance 642 out of 775 structures. The upper cutoffs are $Mn^{3+}$: 4.082 $\mu_B$, $Mn^{4+}$: 3.228 $\mu_B$, and $Mn^{2+}$: 4.973 $\mu_B$. Explanation S1 describes the approach in more detail. All magnetic moments in all DFT structures and their Bayesian-optimized cutoffs are plotted in Fig. 1.

## Results: data preparation – structure mapping

As mentioned earlier, the rigorous implementation of the cluster expansion to model configurational disorder relies on a one-to-one mapping between relaxed DFT structures and a lattice occupation[53]. Typically, mapping back to the lattice configuration is done by performing structure matching after density rescaling, such that the density of the relaxed DFT structure is a multiple of the primitive cell[14,16]. Such mapping can be performed using the StructureMatcher functionality in Pymatgen[35]. In this structure mapping, an attempt is made to map all atoms from the relaxed DFT structure onto a subset of the sites of a supercell of the primitive cell within a set tolerance. Because the sites of the supercell of the primitive cell are the ideal "rigid" lattice sites, the mapping allows for each atom (and its species) in the relaxed DFT structure to be associated with a lattice site, and the remaining lattice sites (with no associated relaxed atom) are assumed to be vacant.

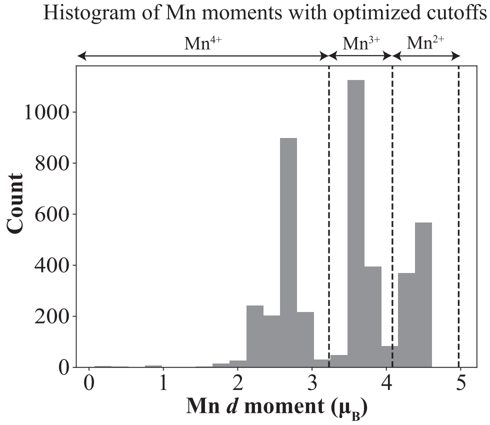

**Fig. 1 All Mn moments in 775 DFT-SCAN structures with Bayesian-optimized moments (dashed line).** The optimized cutoffs results in 642 out of 775 charge-balanced structures.

However, in ionic systems significant relaxation may occur in the DFT calculation of a structure. This may include distortions of the anion lattice due to size differences of the cations, vacancies, Jahn–Teller effects, and off-center relaxations of the cations in their anion coordination polyhedron. As long as the relaxed DFT structures maintain the topology of the CE lattice, they can in principle be mapped onto the lattice model. In the previously described structure mapping method, atoms that distort outside of the set tolerance can no longer be mapped to lattice sites. For example, in Fig. 2a, the relaxed $Li^+$ (green sphere) should be associated with the cation lattice site (white sphere) at the center of the anion coordination polyhedra because it still sits within the anion coordination polyhedra but is outside the set tolerance for structure mapping. While one could attempt to include the case in Fig. 2a by simply increasing the tolerance for mapping, such increased tolerance can result in the mis-mapping of other atoms. Because in ionic systems the identification of a cation with a specific anion polyhedron is a key topological element, we propose a new method to properly map moderately distorted cations to cation lattice sites based on their anion coordination polyhedra.

Figure 2b demonstrates how we can obtain mappings from the relaxed DFT structures to the lattice configuration. Because the anion FCC framework of the spinel materials defines the cation sites, we first map only the anion sites ($a_i$) in the relaxed structure ($s_{relaxed}$) to the anion lattice sites ($s_{lattice}$) directly using the traditional StructureMatcher approach. This mapping must be successful for the relaxed structure to be considered as having an FCC anion lattice. For the cations, which can undergo larger relaxations, we associate each cation ($c_i$) in $s_{relaxed}$ to its anion polyhedra by finding the set of nearest neighbor $a_i$ whose convex hull is not broken by $c_i$. Because we can map the anions from structure $s_{relaxed}$ to its anion lattice sites (the anions in $s_{lattice}$), and we can also locate the cations in $s_{relaxed}$ in their anion polyhedra, we can map the cations to their cation lattice sites via an intermediate mapping based on the anion polyhedra of the cation sites in both $s_{relaxed}$ and $s_{lattice}$.

Using a combination of the StructureMatcher method and the method for mapping cations based on their anion polyhedra in the FCC lattice, we successfully obtain lattice configurations for 448 relaxed DFT structures, resulting in an overall efficiency of 70%. Of the 194 structures that fail to map, we are unable to map 106 structures due to a failure in the anion mapping (i.e., the anion FCC lattice is not adequately maintained). An additional 16 structures contain mappings of species to cation lattice sites where they are disallowed (i.e., $Mn^{3+}$ or $Mn^{4+}$ on the tetrahedral sites). The remaining 72 structures cannot be mapped due to improper identification of the anion polyhedra in the $s_{relaxed}$. Improper identification of the anion polyhedra can result when relaxation of the cation is so severe that it distorts so far (>3.1 Å) away from one or more of the anions constituting its polyhedra that the neighboring anion is no longer identified as a possible member of the cation's anion polyhedra, as in Fig. 2c. In this case the Mn ion in the octahedral has taken on a 2+ valence state which

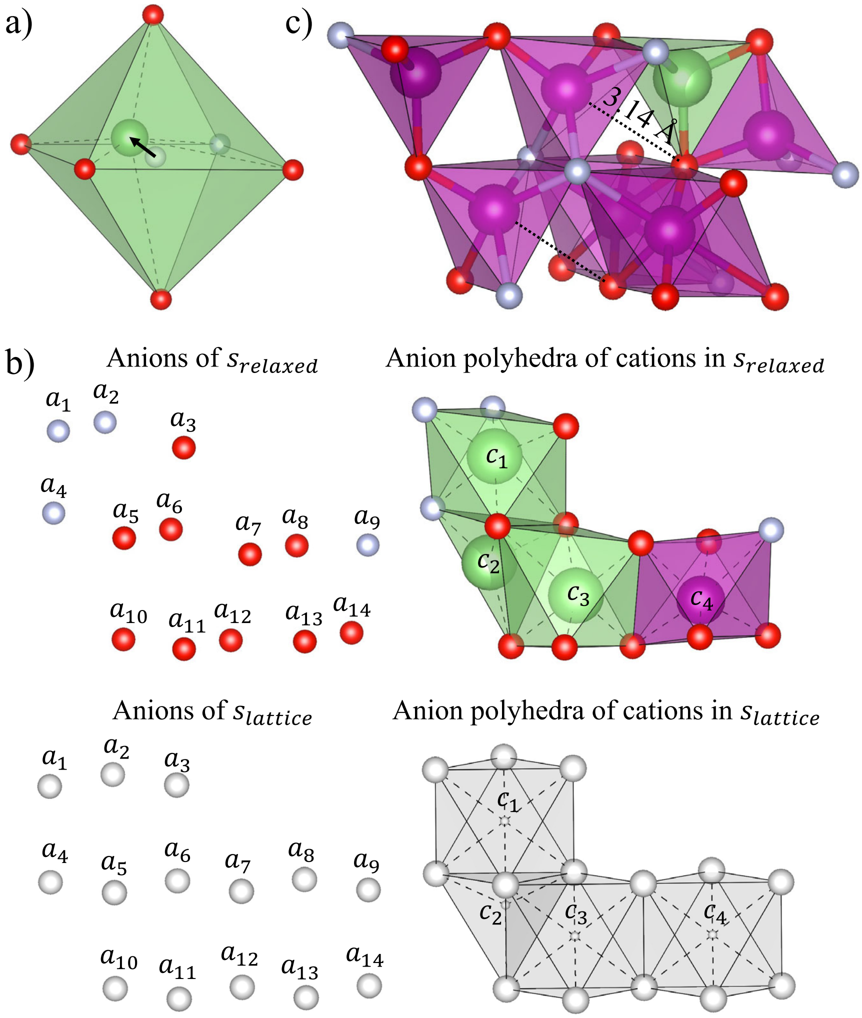
a)
c)
3.14 Å
b)
Anions of $s_{relaxed}$
$a_1$ $a_2$ $a_3$ $a_4$ $a_5$ $a_6$ $a_7$ $a_8$ $a_9$ $a_{10}$ $a_{11}$ $a_{12}$ $a_{13}$ $a_{14}$
Anion polyhedra of cations in $s_{relaxed}$
$c_1$ $c_2$ $c_3$ $c_4$
Anions of $s_{lattice}$
$a_1$ $a_2$ $a_3$ $a_4$ $a_5$ $a_6$ $a_7$ $a_8$ $a_9$ $a_{10}$ $a_{11}$ $a_{12}$ $a_{13}$ $a_{14}$
Anion polyhedra of cations in $s_{lattice}$
$c_1$ $c_2$ $c_3$ $c_4$

**Fig. 2 Details of structural mapping in Li-Mn-O-F rocksalt system. a** Example of $Li^+$ (green sphere) and its anion polyhedra in relaxed DFT structure which cannot be mapped to its proper cation site (white sphere), but which can be mapped using the new mapping technique. **b** Diagram of new structure mapping process, which involves mapping the anions ($a_i$) of the relaxed structure $s_{relaxed}$ to the anion sites of the lattice configuration $s_{lattice}$ (left), followed by mapping the cations of $s_{relaxed}$ to the proper cation sites in $s_{lattice}$ by matching the anions in their anion polyhedra. **c** Example structure which fails to map using new mapping technique due to an $Mn^{2+}$ that has relaxed too far from one of the $O^{2-}$ anions in its anion polyhedra.

strongly prefers tetrahedral coordination. In an octahedron, a pseudo-tetrahedral environment can be achieved by relaxing to the center of the pyramid that constitutes half of the octahedron.

Lastly, we de-duplicate all 448 structures by their correlation vectors, finding a total of 428 distinct structures to be used for training and testing. Figure S2a shows the 0 K DFT ground states. The ground states are consistent with low-temperature experimental phases reported in the phase diagram of Li-Mn-O spinel in air by Paulsen and Dahn[54].

## Result: charge constraints on point basis functions in ionic CE

Given DFT structures which are charge-balanced and mapped to all sublattices, we next describe fitting procedures specific to reducing error in ionic CE. In statistics and machine learning, it is standard to center the target vector, i.e. train on $E(\vec{\sigma}) - \langle E(\vec{\sigma})\rangle$, so we propose here that $J_0$ can be fitted to the average energy of the training set. However, note that since the zeroth basis function is defined as 1, the true value of $J_0$ is the average energy of the random sample with sampled centered basis functions.

Next, the charge neutrality constraint limits the rank of the point basis functions. By writing the charge constraint for the number of species $N_i$, of each type $i$:

$$N_{Li^+} + 2N_{Mn^{2+}} + 3N_{Mn^{3+}} + 4N_{Mn^{4+}} = N_{O^{2-}} + N_{F^-} \tag{5}$$

it is clear any function of $N$, such as the occupation mapping function, $f(N_{Li^+}, \ldots) = \sigma$, and functions of $\boldsymbol{\sigma}$, such as the single site basis functions, will also be constrained and have its rank reduced by one arising from Eq. (5). This is why with the fitting of the point ECI, the correct degrees of freedom need to also be enforced such that one ECI is set to 0. Otherwise, overfitting of the point ECI will result in higher out-of-sample error.

## Results: applying structured sparsity due to rank deficiency

In principle a CE is always under-determined because there exist an infinite number of basis functions for a finite number of training data. In simple binary systems, we can sometimes posit that a subset of basis functions are relevant and solve for their corresponding ECI by fitting to an over-determined system. However, with high-component systems this procedure becomes

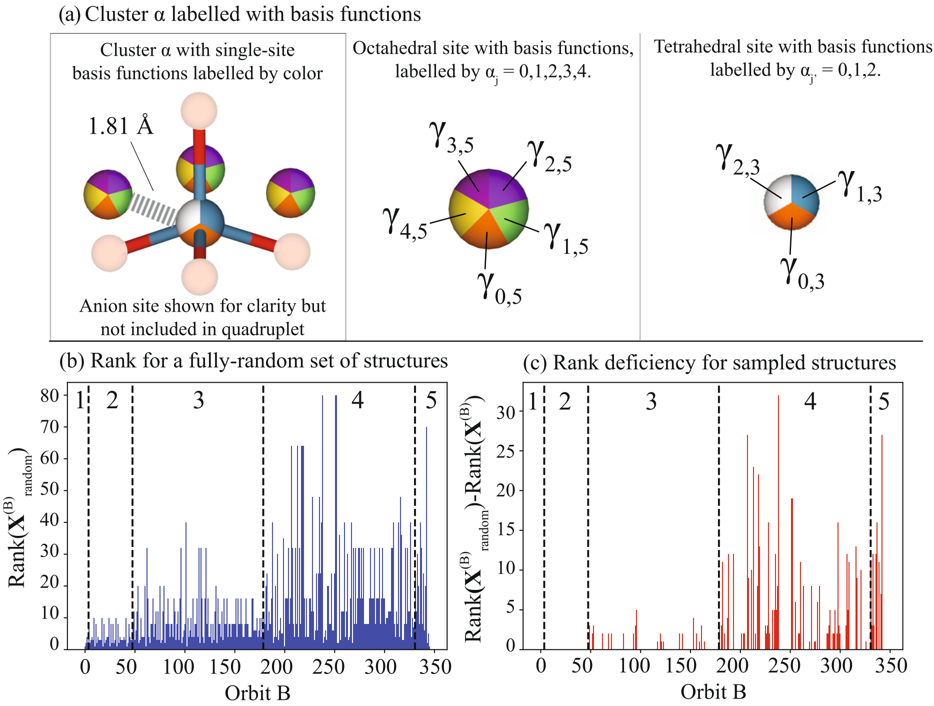

**Fig. 3 Rank deficiency in high-component ionic systems. a** Illustration of the cluster $\alpha$ with single site basis functions labelled by color. Since the zeroth label is independent of occupation and is always 1, the colors are the same (orange). However, since the rest of the basis functions ( $\gamma_{\alpha_j>0,n_j}(\sigma_i)$) are dependent on occupation, they are differently colored. **b** Rank of all orbits **B** for the fully random set of structures and the (**c**) rank deficiency of all orbits **B** for the set of structures in this study. The size of the cluster generating orbit B is indicated for points (1), pairs (2), triplets (3), quadruplets (4), and quintuplets (5).

complicated because even with a small set of clusters we have a large number of ECI. In these cases, we will always start from an under-determined system and use statistical approaches to enforce sparse solutions. One might add a constraint to the least-squares error function using Lagrange multipliers to penalize the $L_1$ norm of solutions, an approach known as lasso regularization[37]. Lasso regularization returns more sparse models than least-squares regression which always returns dense solutions. When coefficients are set to zero in lasso regularization, their corresponding basis functions play no role in energy prediction. Lasso regularization has been used to study Ag-Pt, model protein folding in the zinc-finger motif[55], and construct models for Cu-Pt, Ag-Pt, and Ag-Pd via reweighted Bayesian compressive sensing[56]. We find that applying lasso regularization to our system results in higher average training and testing errors (see Figure S1 for the pure lasso case). Other approaches to select clusters in fitting CE include using genetic algorithms[57–59] or the steepest descent algorithm to add or remove clusters one at a time as a function of cross-validation score[60].

In this section we introduce another regularization approach, sparse group lasso (SGL)[39], which builds on lasso regularization by applying structured sparsity: starting from the usual penalized lasso regression framework with an $n$ by $p$ covariate matrix $\mathbf{X}$ (made of p−1 correlation functions and the Ewald energy, for $n$ structures) and a response vector with centered energies $E'$, SGL further breaks down $\mathbf{X}$ into sub-matrices $\mathbf{X}^{(B)}$, where each sub-matrix has dimension $n$ by $p_B$ where $p_B$ is the size of a member B in **B**. (Remember that $p_B$ is the number of contracted multi-indices labeling a geometric cluster, which symmetrized and evaluated over the random structure on this CE lattice, produces member B.) $p_B$ is effectively a weighted penalization. The ECI $J_\beta$ are chosen such that they minimize the objective function to solve the convex optimization problem:

$$min_{J_\beta}\left(\frac{1}{2n}\left\|E' - \sum_B \mathbf{X}^{(B)}J_\beta^{(B)}\right\|_2^2 + \lambda\alpha\left\|J_\beta\right\|_1 + (1-\lambda)\alpha\sum_B \sqrt{p_B}\left\|J_\beta^{(B)}\right\|_2\right) \tag{6}$$

The penalty parameter $\alpha > 0$ bounds both the $l_1$ norm of all ECI $J_\beta$ and the $l_2$ norm of the vector of ECI that are within each orbit B, $J_\beta^{(B)}$. $\lambda \in [0, 1]$ is a mixing parameter. In the limiting cases when $\lambda =$ 1, the objective function becomes that of lasso; when $\lambda = 0$, the objective function becomes that of group lasso[38], an approach which enforces orbit-wise sparsity. Intermediate values of $\lambda$ enforce sparsity in $J_\beta^{(B)}$. The mixing parameter $\lambda$ is set to 0.5 and $\alpha$ is 0.056 in this study, and details of the hyperparameter optimization are given in Methods and Figure S1.

Our approach to enforce structured sparsity by applying SGL is fundamentally different than enforcing structured sparsity via hierarchical cluster selection rules by applying group lasso, the approach used by Leong and Tan to study the ternary Mo-V-Nb alloy[61]. In their work, cluster functions are selected only after their sub-clusters are also selected. Here, we do not employ such hierarchical constraints. Instead, structured sparsity is obtained by grouping ECI by their corresponding orbit B, obtaining orbit-wise sparse solutions when entire groups of ECI are set to zero.

Furthermore, within an orbit B, sparsity in $J_\beta^{(B)}$ is attained. This is a necessary approach to handle under-determined sub-matrices $\mathbf{X}^{(B)}$, which is common when including larger geometric clusters such as quadruplets. Consider the tetrahedral site with basis functions $\{\gamma_{\alpha_{j'},3}\}$ face-sharing with three of its nearest-neighbor octahedral sites each with basis functions $\{\gamma_{\alpha_j,5}\}$. The cluster $\alpha$ labelled with these basis functions are shown in Fig. 3a. In total, there are 80 contracted multi-indices, which can be obtained after

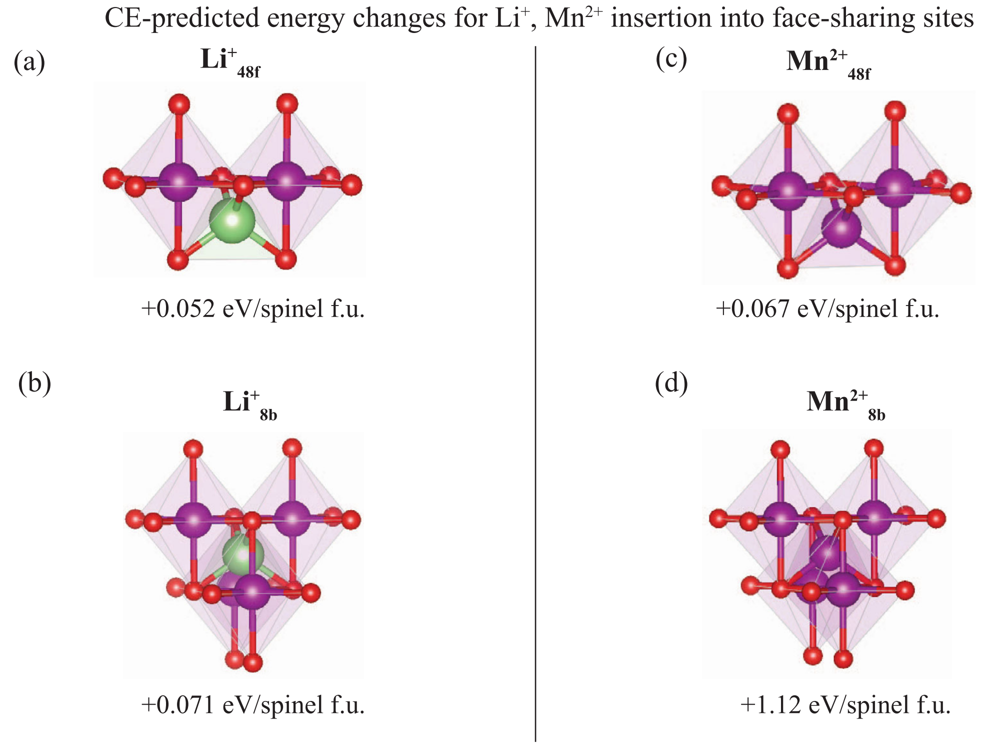

**Fig. 4 Examples of high-energy cation configurations predicted by the CE per spinel formula unit (f.u.), $LiMn_2O_4$, where the scenarios are $Li^+$ or $Mn^{2+}$ insertions onto vacant sites in spinel (48f or 8b).** The defect energies are calculated as either $E^{final} - E^{initial} - \mu_{tet}^{Li^+}$ for $Li^+$ insertion and $E^{final} - E^{initial} - \mu_{tet}^{Mn2+}$ for $Mn^{2+}$ insertion. The chemical potentials are calculated, starting from the spinel structure, as: $E(Li_8Mn_{16}O_{32}) - E(Li_7Mn_{16}O_{32}) = \mu_{Li^+}$ and $E(Li_7Mn_1^{2+}Mn_{16}O_{32}) - E(Li_7Mn_{16}O_{32}) = \mu_{Mn^{2+}}$. **a** The $Li^+$-occupied 48f site face-shares with two Mn, resulting in a +0.052 eV/spinel f.u. increase in energy. **b** Adding $Li^+$ to a more metal-rich cluster, the 8b site, results in an even higher increase in energy: +0.071 eV/spinel f.u. **c** The $Mn^{2+}$-occupied 48f site, face-sharing with two Mn, has a +0.067 eV/ spinel f.u. increase while the (**d**) $Mn^{2+}$-occupied 8b site increases by +1.12 eV/ spinel f.u.

taking the Cartesian product of the single site basis functions, removing all labels where $\alpha_j = 0$ or $\alpha_{j'} = 0$, and applying translation symmetry. During fitting of $J_\beta^{(B)}$, in order for the submatrix corresponding to orbit B to be full rank, we need to train with at least 80 unique, symmetrized decorations so that $n_{unique} = p_B = 80$. We call the submatrix for our training data $\boldsymbol{X}^{(B)}$ and the submatrix of the fully random set of structures on this CE lattice $\boldsymbol{X}_{random}^{(B)}$. The fully random set of structures is the set that contains every possible lattice configuration in a supercell with size up to the largest cutoff (7 Å).

However, the rank in our DFT input set of this quadruplet is only 61 so $\boldsymbol{X}^{(B)}$ is clearly rank-deficient. The complexity in part comes from the ability of $Li^+$ and $Mn^{2+}$ to occupy the central tetrahedral site when the neighboring octahedral sites host $Mn^{2+}$, $Mn^{3+}$, or $Mn^{4+}$. Such configurations exist in theory, but in reality, tetrahedral $Li^+$ or $Mn^{2+}$ will only occur if its three nearest-neighbor octahedral sites are also vacant or hosting $Li^+$, since the octahedral site and tetrahedral site are very close together (around 1.8 Å apart) and experience strong electrostatic repulsion when both are occupied. Between these two closely situated cation sites there is little charge shielding. Thus, the configuration space of this quadruplet cluster spans a larger space than physically sampleable, so rank deficiency, defined as $rank\left(\boldsymbol{X}_{random}^{(B)}\right) - rank\left(\boldsymbol{X}^{(B)}\right)$, is observed.

This symptom of under-sampling is evident in submatrices $\boldsymbol{X}^{(B)}$ for other members B. Figure 3b shows the rank of the submatrices for each member in **B** for the fully random set of structures on this CE lattice, $\boldsymbol{X}_{random}^{(B)}$, and Fig. 3c shows the rank deficiency observed in the physical set of structures used in this study. There is clearly rank deficiency across triplets, quadruplets, and quintuplets ranging from five for triplets, to almost 30 for quadruplets. To avoid overfitting in all cases, sparsity of the ECI within an orbit can be enforced by using sparse group lasso. The lack of information on the energetics of these configurations is not a problem as long as their energies as represented by lower order clusters are high enough so that they are never sampled in MC simulations with the CE. Figure 4 shows examples of clusters with face-sharing cations, where occupancy of the 8b or 48f site results in face-sharing with octahedral sites. When $Li^+$ or $Mn^{2+}$ occupy the 48f site which face-shares with two occupied octahedral sites (Fig. 4a, c), the defect energies predicted by the CE are +0.052 eV/spinel formula unit and +0.067 eV/spinel formula unit, respectively. Thus, the even more cation-rich clusters in Fig. 4b, d, where $Li^+$ or $Mn^{2+}$ face-shares with four octahedral sites, are unlikely to be sampled during MC, since their CE-predicted energies are even higher.

## Results: applying sparse group lasso

The testing and training error depend on the number of training examples (known as the learning curve) and model complexity (known as the capacity curve)[62], and both evaluations are shown in Fig. 5. The learning curve, which compares training data size against a loss (root mean squared error (RMSE) per primitive cell in our case), is a widely used metric to assess model convergence. During this process which is shown in Fig. 5a, we conduct 50 cross-validation trials, setting aside 80% of the total sample size for training and testing on the remaining 20%. Since the learning curve converges in training and testing RMSE, SGL is neither under-fit nor over-fit, and the validation dataset is representative[63]. The mark of an under-fit model is that training and cross-validation RMSE continue to decrease with increasing examples,

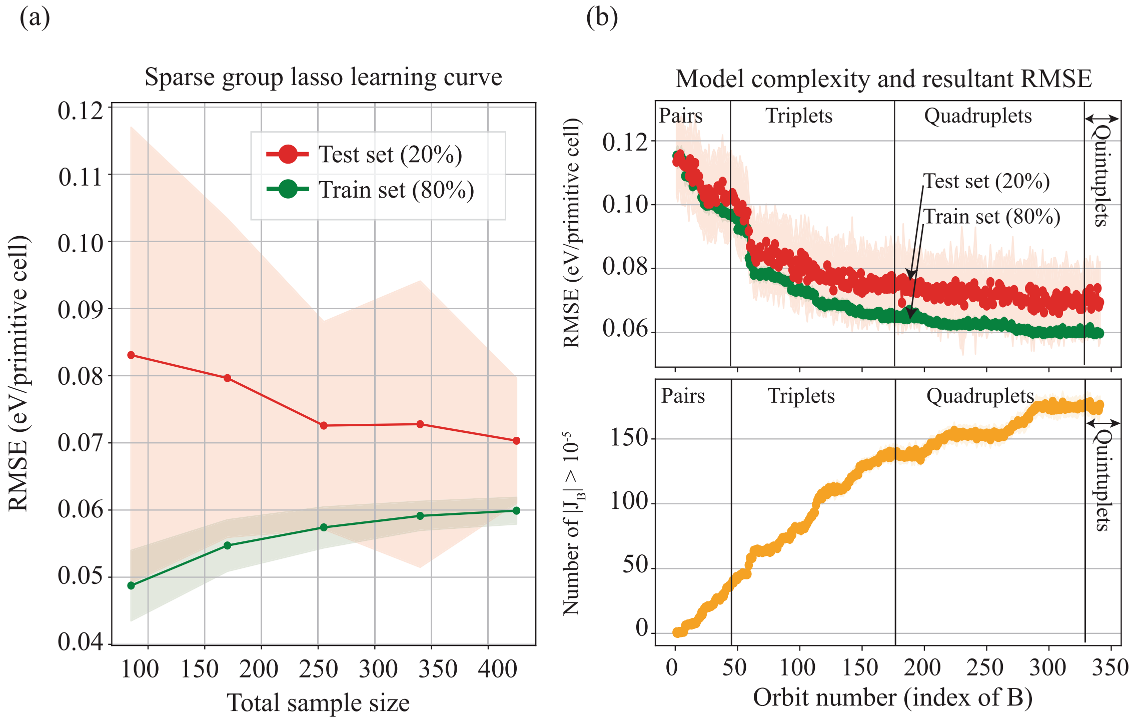

**Fig. 5 Loss as a function of sample size and model complexity.** The green (red) colors indicate the average and standard deviation of the loss for training (testing) in 50 cross-validation trials, setting aside 80% of the 428 structures for training and 20% for testing. **a** The learning curve for SGL with a loss function of root mean squared error (RMSE) per primitive cell as a function of sample size. The chosen hyperparameters are $\alpha = 0.056$ and $\lambda = 0.5$. **b** The RMSE as a function of model complexity, starting from including only the first orbit in a pair cluster and ending with including all geometric clusters up to quintuplets. Each individual model always uses the Ewald energy and all features in orbits up to the orbit number indicated. The number of significant features selected for each model is in yellow, showing how the number of ECI increases to over 150 in the last model.

indicating the model fitting was halted prematurely. On the other hand, over-fit models diverge in training and cross-validation RMSE because the model has been over-fit to the training samples.

The learning curve in Fig. 5a shows the typical behavior of testing and training[64], where with few training samples the model has enough free parameters to completely model the training set, so the training error is small. This sampling is not indicative of the test set so the out-of-sample error is high. With increasing training set size, the test error decreases. As we further increase the training set size, a testing RMSE of around 70 meV/primitive cell and training error of around 60 meV/primitive cell are reached. Figure S3 applies early theoretical work by Cortes et al.[65] in the convergence of learning curves, to project that the asymptotic RMSE convergence for PDS is 70 meV/primitive cell, indicating that more structure sampling is not expected to reduce the error in Fig. 5a. This convergence is not a property of the group lasso, but the best achievable performance among “all” models[62,65].

In CE, it is generally known and demonstrated through examples that better predictability can be achieved if more interactions are considered[66]. We apply this concept in the capacity curve in Fig. 5b, carrying out 50 cross-validation trials, again setting aside 80% of the 428 structures for training and 20% for testing. We build increasingly complex models by including more orbits, and always fitting with the Ewald energy. By adding more orbits in B, we find that both training and testing RMSE converge to 60 meV/primitive cell and 70 meV/primitive cell, respectively, approaching the limiting performance or asymptotic performance of the data[62].

In fact, this convergence is almost achieved using solely orbits from pairs and triplets. The lack of continuously decreasing RMSE indicates that even with information from quadruplets or quintuplets, the predictability does not improve, suggesting that in capturing the configurational energetics of Li-Mn-O-F the most critical information is contained in pairs and triplets. Figure S2 shows the convex hull of the CE which reproduces most of the ground states predicted by DFT. However, the CE also stabilizes additional ground state configurations along the $MnO$-$MnO_2$ tie line (Figure S2c). The depth of the hull in the CE and DFT phase diagram are similar (−0.3 eV/atom).

Figure 5 shows conclusively that in training a model for this high-component system, the out-of-sample average RMSE converges to around 70 meV/primitive cell or 35 meV/atom, with on average 176 selected features. This error is higher than errors reported for CE fits in multicomponent rocksalt systems: less than 8 meV/atom in ternary disorder for Li–Mn–Zr–O and Li–Mn–Ti–O[16], 18 meV/oxygen in ternary Li–Ni–Vac–O[67], and 21 meV/atom in ternary-binary disorder for rocksalt $Li^+$–Vac–$Cr^{3+}$–$O^{2-}$–$F^-$[18]. However, as we will discuss, this error may be reasonable given the dimensionality of PDS system, as lower-dimensional fits to subspaces of this dataset provide comparable RMSE to those in the literature.

## DISCUSSION: HIGH PREDICTION ERROR IN HIGH-COMPONENT IONIC SYSTEMS

We demonstrated that charge constraints on point-orbits limit their rank and showed that rank deficiency in higher-order orbits can be handled by applying SGL. However, even with increasing model complexity as shown in Fig. 5b, we are unable to converge to lower RMSE than around 35 meV/atom for out-of-sample RMSE. In fact, Fig. 5b shows that including orbits after pairs and triplets offer little improvement in RMSE.

Here we show data indicating that this higher RMSE may be reasonable for high-component systems due to their dimensionality.

**Table 2.** The performance of various composition models using only pairs up to 7 Å and triplets up to 5 Å and the Ewald energy.

| Composition space (system, total configurational space) | Avg. out-of-sample RMSE in meV/atom, rocksalt composition (RMSE in literature) | Avg. features | Number of structures |
|---|---|---|---|
| Oct: $Li^+$-$Mn^{3+}$-$Mn^{4+}$<br>Anion: O<br>(ternary, 3) | **13** (8[16], 18[67]) | 24 | 126 |
| Oct: $Li^+$-Vac-$Mn^{3+}$-$Mn^{4+}$<br>Anion: O (quaternary, 4) | **13** | 44 | 161 |
| Oct: $Li^+$-$Mn^{3+}$-$Mn^{4+}$<br>Anion: O-F<br>(ternary-binary, 6) | **25** (21[18], 24[68]) | 42 | 165 |
| Oct: $Li^+$-Vac-$Mn^{3+}$-$Mn^{4+}$<br>Anion: O-F (quaternary-binary, 8) | **28** | 71 | 222 |
| Oct: $Li^+$-Vac-$Mn^{2+}$-$Mn^{3+}$-$Mn^{4+}$<br>Anion: O-F<br>(quinary-binary, 10) | **27** | 106 | 343 |
| Oct: $Li^+$-Vac-$Mn^{2+}$-$Mn^{3+}$-$Mn^{4+}$<br>Anion: O-F<br>Tet: Li-Vac<br>(binary-quinary-binary, 40) | **34** | 136 | 377 |
| Oct: $Li^+$-Vac-$Mn^{2+}$-$Mn^{3+}$-$Mn^{4+}$<br>Anion: O-F<br>Tet: Li-Vac-$Mn^{2+}$<br>(ternary-quinary-binary, 90) | **35** | 150 | 428 |

The out-of-sample root mean squared error (RMSE) and number of features are averaged over 50 CV trials, setting aside 80% of the number of structures for training and testing on the remaining 20%. The composition space, level of disorder, and total configurational space are also indicated in the first column.

Table 2 illustrates the RMSE for increasing chemical complexity in ternary, quaternary, ternary-binary, quaternary-binary, quinary-binary, binary-quinary-binary, and ternary-quinary–binary systems, subspaces of our cluster expansion model. We fit each system using pairs up to 7 Å and triplets up to 5 Å because Fig. 5b suggests predictability is mostly achieved with these interactions in the Li–Mn–O–F space. We again perform 50 cross-validation trials, training SGL on 80% of the number of structures available and testing on the remaining 20% of structures. The average number of features, average RMSE, and total number of applicable structures (of the 428) are shown for each composition space modeled. Where possible, we juxtapose with reported RMSE in the literature in parentheses.

Even when using 126 structures, for the ternary $Li^+$–$Mn^{3+}$–$Mn^{4+}$–O system, we achieve a respectable RMSE of 13 meV/atom, compared to values of 8 meV/atom for Li–Mn–Ti–O[16] and 18 meV/atom for Li–Ni–Vac–O[67]. Including another level of complexity, vacancies on the octahedral site, results in a similar level of error of 13 meV/atom. The error increases to 25 meV/atom for a ternary-binary system, which is similar to 22 meV/atom[18] for Li–Cr–O–F and 24 meV/atom for Li–V–O–F[68]. For quaternary-binary and quinary-binary disorder, the RMSE are 28 meV/atom and 27 meV/atom. Lastly, we observe the RMSE of 35 meV/atom with the inclusion of binary or ternary disorder on the third sublattice.

Table 2 shows that adding sublattices in ionic CE always increases the RMSE, but the level of error in lower-dimensional systems is still comparable to those reported in literature. This finding is remarkable as it shows that ternary and ternary-binary ionic systems do not require multiple hundreds of DFT data and that pair and triplet interactions are reasonably sufficient starting CE models, provided that they utilize the approaches described here.

As this is the first high-component ionic CE that uses three sublattices with 10 species, the high RMSE may be reasonable in that the same approach is able to represent lower-dimensional systems well. The analysis suggests that high-component systems may be limited to higher RMSE compared to those in lower component systems. New compressed sensing approaches, such as one which employs coherency and redundancy to utilize the compressibility of configurational energy, may be promising alternative routes to increase predictability[69].

## CONCLUSIONS

We have described practical and theoretical advances in high-component ionic CE models. Automated charge assignments and modified structure mapping procedures enable more complex data to be included during fitting. We show that electroneutrality constraints decrease the rank of charge-constrained orbits, and rank deficiency in orbits can be handled by using sparse group lasso regularization. This lack of information is not a problem as long as the energetics of high-energy configurations are represented in lower order clusters so that they are never sampled during Monte Carlo simulations. We discuss that the new approaches predicting higher RMSE in this work still predict lower RMSE consistent with those in literature dealing with lower-dimensional systems, and suggest that, considering practical limitations, the high RMSE may be unavoidable for high-component ionic CE. In summary, the approaches outlined in this work provide critical guidance for meticulous understandings of other high-dimensional ionic systems not just limited to the FCC anion lattice.

## METHODS: FIRST-PRINCIPLES DATA GENERATION

We use Density Functional Theory (DFT) with the semi-local SCAN meta-generalized gradient density functional approximation for the exchange-correlation correction. Previous studies found SCAN[70] to be most suitable for ground state structure prediction in ionic systems[71] due to its ability to capture medium-range Van der Waals interactions[72]. In addition, internal coordinate relaxations are closer to experimentally reported values for SCAN than those observed in PBE and PBE+U[73]. These reasons make the DFT-SCAN approximation a rational choice for parametrizing the effective cluster interactions in an ionic system, despite its higher computational cost.

For our system 775 DFT-SCAN structures are calculated using the Vienna Ab Initio Simulation Package (VASP)[74,75], using the projector augmented wave (PAW) method[76,77], with reciprocal space discretization of 25 k-points per $Å^{-1}$ and a plane wave energy cutoff of 520 eV. Calculations use the VASP-recommended pseudopotentials (Li_sv, Mn_pv, O, and F) are converged to $10^{-6}$ eV in total energy and 0.01 eV/Å on atomic forces. The initial set of structures were generated by scraping an internal database for structures within the Li-Mn-O-F composition space containing fewer than 50 atoms to limit computational cost, ionic substitution for $Mn^{4+}$ onto spinel-like Li-Ti-O structures from another work[78], the Inorganic Crystal Structure Database for defect spinels, and Monte Carlo CE searches for ionic configurations with low Ewald energy. The typical iterative approach to refine structures[79,80] was completed using CE-Monte Carlo, concluding the search for new structures when the cross-validation (CV) score is 65 meV/primitive cell, equivalently 33 meV/atom, assuming a rocksalt composition. The smallest and largest cell sizes sampled are 1, corresponding to $Li_2O$, and 64, corresponding to $Li_{32}Mn_{32}O_{64}$.

## Methods: hyperparameter optimization in lasso, sparse group lasso, and group lasso

*Algorithm overview*. Deviating from the algorithm described in[39] which cyclically iterates through all groups, we iterate through each member in B only once, from the first member $B_1$ to the last member.

(1) (Outer loop) For the current group $B_i$, execute step 2.
(2) Check if the coefficients are identically 0 by seeing if they obey the sub-gradient equations in ref. [39]. If not, apply step 3.

(3) (Inner loop) Solve for the coefficients $J_\beta^{(B)}$ using Elastic Net regularization, choosing a random coefficient to be updated every iteration.

We test the hyperparameter $\alpha$ and degree of mixing $\lambda$ for the three approaches (Lasso, Group Lasso, and Sparse Group Lasso), and show the results in Figure S1. We sample $\alpha$ from 25 evenly spaced intervals on the log scale from $10^{-1.5}$ to $10^{-0.5}$ and $\lambda$ for 0 (pure lasso), 0.25, 0.50, 0.75, and 1.0 (pure group lasso). The root-mean squared errors, setting aside 80% for the training and 20% for the testing, and number of features are averaged over 50 cross-validation trials. We find that the sparsest solutions and lowest error result from even-mixing of lasso and group lasso ($\lambda = 0.5$) and $\alpha = 0.056$.

## DATA AVAILABILITY

The feature matrix, energies, groups, and 775 DFT-SCAN data used to train the sparse group lasso are available at github.com/juliayang/high-component-ce-tools. Within the public repository, we also make available codes for training Sparse Group Lasso and for using the BayesianChargeAssigner. Finally, we include jupyter notebook tutorials for using these codes.

## CODE AVAILABILITY

The statistical mechanics on lattices (smol) package is a Ceder group repository and is available at: https://github.com/CederGroupHub/smol.

Received: 8 March 2022; Accepted: 25 May 2022;
Published online: 27 June 2022

## REFERENCES

1. Yeh, J.-W. & Lin, S.-J. Breakthrough applications of high-entropy materials. *J. Mater. Res.* **33**, 3129–3137 (2018).
2. Ma, Y. et al. High-entropy energy materials: challenges and new opportunities. *Energy Environ. Sci.* **14**, 2883–2905 (2021).
3. Feng, R. et al. High-throughput design of high-performance lightweight high-entropy alloys. *Nat. Commun.* **12**, 4329 (2021).
4. Bérardan, D., Franger, S., Meena, A. K. & Dragoe, N. Room temperature lithium superionic conductivity in high entropy oxides. *J. Mater. Chem. A* **4**, 9536–9541 (2016).
5. Gild, J. et al. High-entropy metal diborides: a new class of high-entropy materials and a new type of ultrahigh temperature ceramics. *Sci. Rep.* **6**, 37946 (2016).
6. Wang, Q. et al. Multi-anionic and -cationic compounds: new high entropy materials for advanced Li-ion batteries. *Energy Environ. Sci.* **12**, 2433–2442 (2019).
7. Lun, Z. et al. Cation-disordered rocksalt-type high-entropy cathodes for Li-ion batteries. *Nat. Mater.* **20**, 214–221 (2021).
8. Walsh, F., Asta, M. & Ritchie, R. O. Magnetically driven short-range order can explain anomalous measurements in CrCoNi. *Proc. Natl Acad. Sci.* **118**, e2020540118 (2021).
9. Widom, M. Modeling the structure and thermodynamics of high-entropy alloys. *J. Mater. Res.* **33**, 2881–2898 (2018).
10. Goiri, J. G. & Van der Ven, A. Recursive alloy Hamiltonian construction and its application to the Ni-Al-Cr system. *Acta Mater.* **159**, 257–265 (2018).
11. Zhang, J. et al. Robust data-driven approach for predicting the configurational energy of high entropy alloys. *Mater. Des.* **185**, 108247 (2020).
12. Kleiven, D. & Akola, J. Precipitate formation in aluminium alloys: multi-scale modelling approach. *Acta Mater.* **195**, 123–131 (2020).
13. Tepesch, P. D. et al. A model to compute phase diagrams in oxides with empirical or firs}_principles energy methods and application to the solubility limits in the Cao_MgO system. *J. Am. Ceram. Soc.* **79**, 2033–2040 (1996).
14. Richards, W. D., Dacek, S. T., Kitchaev, D. A. & Ceder, G. Fluorination of lithium-excess transition metal oxide cathode materials. *Adv. Energy Mater.* **8**, 1701533 (2018).
15. Kitchaev, D. A. et al. Design principles for high transition metal capacity in disordered rocksalt Li-ion cathodes. *Energy Environ. Sci.* **11**, 2159–2171 (2018).
16. Ji, H. et al. Hidden structural and chemical order controls lithium transport in cation-disordered oxides for rechargeable batteries. *Nat. Commun.* **10**, 592 (2019).
17. Ji, H. et al. Ultrahigh power and energy density in partially ordered lithium-ion cathode materials. *Nat. Energy* **5**, 213–221 (2020).
18. Chang, J. H. et al. {CLEASE}: a versatile and user-friendly implementation of cluster expansion method. *J. Phys. Condens. Matter* **31**, 325901 (2019).
19. Sanchez, J. M., Ducastelle, F. & Gratias, D. Generalized cluster description of multicomponent systems. *Phys. A Stat. Mech. its Appl.* **128**, 334–350 (1984).
20. van de Walle, A. Multicomponent multisublattice alloys, nonconfigurational entropy and other additions to the Alloy Theoretic Automated Toolkit. *Calphad* **33**, 266–278 (2009).
21. Cai, Z. et al. Realizing continuous cation order-to-disorder tuning in a class of high-energy spinel-type Li-ion cathodes. *Matter* **4**, 3897–3916 (2021).
22. Clément, R. J., Kitchaev, D., Lee, J. & Ceder, G. Short-range order and unusual modes of nickel redox in a fluorine-substituted disordered rocksalt oxide lithium-ion cathode. *Chem. Mater.* **30**, 6945–6956 (2018).
23. Sobieraj, D. et al. Chemical short-range order in derivative Cr–Ta–Ti–V–W high entropy alloys from the first-principles thermodynamic study. *Phys. Chem. Chem. Phys.* **22**, 23929–23951 (2020).
24. Lavrentiev, M. Y., Drautz, R., Nguyen-Manh, D., Klaver, T. P. C. & Dudarev, S. L. Monte Carlo study of thermodynamic properties and clustering in the bcc Fe-Cr system. *Phys. Rev. B* **75**, 14208 (2007).
25. van de Walle, A., Asta, M. & Ceder, G. The alloy theoretic automated toolkit: a user guide. *Calphad* **26**, 539–553 (2002).
26. Liu, S., Martínez, E. & LLorca, J. Prediction of the Al-rich part of the Al-Cu phase diagram using cluster expansion and statistical mechanics. *Acta Mater.* **195**, 317–326 (2020).
27. Žguns, P. A., Ruban, A. V. & Skorodumova, N. V. Phase diagram and oxygen–vacancy ordering in the CeO2–Gd2O3 system: a theoretical study. *Phys. Chem. Chem. Phys.* **20**, 11805–11818 (2018).
28. Seko, A., Yuge, K., Oba, F., Kuwabara, A. & Tanaka, I. Prediction of ground-state structures and order-disorder phase transitions in II-III spinel oxides: a combined cluster-expansion method and first-principles study. *Phys. Rev. B* **73**, 184117 (2006).
29. Drautz, R., Singer, R. & Fähnle, M. Cluster expansion technique: an efficient tool to search for ground-state configurations of adatoms on plane surfaces. *Phys. Rev. B* **67**, 35418 (2003).
30. Tepesch, P. D., Garbulsky, G. D. & Ceder, G. Model for configurational thermodynamics in ionic systems. *Phys. Rev. Lett.* **74**, 2272–2275 (1995).
31. Wolverton, C., Ozolins, V. & Zunger, A. Short-range-order types in binary alloys: a reflection of coherent phase stability. *J. Phys. Condens. Matter* **12**, 2749–2768 (2000).
32. Van der Ven, A. & Ceder, G. Vacancies in ordered and disordered binary alloys treated with the cluster expansion. *Phys. Rev. B* **71**, 54102 (2005).
33. Barroso-luque, L. et al. Cluster expansions of multicomponent ionic materials: formalism & methods. pp. 1–44, (2022).
34. Ewald, P. Evaluation of optical and electrostatic lattice potentials. *Ann. Phys.* **64**, 253–287 (1921).
35. Ong, S. P. et al. Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
36. Connolly, J. W. D. & Williams, A. R. Density-functional theory applied to phase transformations in transition-metal alloys. *Phys. Rev. B* **27**, 5169–5172 (1983).
37. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1996).
38. Yuan, M. & Lin, Y. Model selection and estimation in regression with grouped variables,". *J. R. Stat. Soc. Ser. B (Statistical Methodol)* **68**, 49–67 (2006).
39. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. A Sparse-Group Lasso. *J. Comput. Graph. Stat.* **22**, 231–245 (2013).
40. Lee, J. et al. Reversible Mn2+/Mn4+ double redox in lithium-excess cathode materials. *Nature* **556**, 185–190 (2018).
41. Huang, J. et al. Non-topotactic reactions enable high rate capability in Li-rich cathode materials. *Nat. Energy* **6**, 706–714 (2021).
42. Mackrodt, W. C. & Williamson, E.-A. First-principles Hartree-Fock description of the electronic structure of monoclinic C2/m LixMnO2 (1 ≥ x ≥ 0). *Philos. Mag. B* **77**, 1077–1092 (1998).

43. Aydinol, M. K. & Ceder, G. First-principles prediction of insertion potentials in Li-Mn oxides for secondary Li batteries. *J. Electrochem. Soc.* **144**, 3832–3835 (1997).
44. Wolverton, C. & Zunger, A. First-principles prediction of vacancy order-disorder and intercalation battery voltages in Li$_x$CoO$_2$. *Phys. Rev. Lett.* **81**, 606–609 (1998).
45. Ceder, G. et al. Identification of cathode materials for lithium batteries guided by first-principles calculations. *Nature* **392**, 694–696 (1998).
46. Reed, J. & Ceder, G. Role of electronic structure in the susceptibility of metastable transition-metal oxide structures to transformation. *Chem. Rev.* **104**, 4513–4534 (2004).
47. Gwon, H., Seo, D.-H., Kim, S.-W., Kim, J. & Kang, K. Combined first-principle calculations and experimental study on multi-component olivine cathode for lithium rechargeable batteries. *Adv. Funct. Mater.* **19**, 3285–3292 (2009).
48. Jang, Y. et al. LiAly Co$_{1\ -y}$ O$_2$ (R $\bar{3}$m) intercalation cathode for rechargeable lithium batteries. *J. Electrochem. Soc.* **146**, 862–868 (1999).
49. Anisimov, V. I., Aryasetiawan, F. & Lichtenstein, A. I. First-principles calculations of the electronic structure and spectra of strongly correlated systems {theLDA}$ \mathplus$Umethod". *J. Phys. Condens. Matter* **9**, 767–808 (1997).
50. Rasmussen, C. E. "Gaussian Processes in Machine Learning," in *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 — 14, 2003, Tübingen, Germany, August 4 — 16, 2003, Revised Lectures*, O. Bousquet, U. von Luxburg, and G. Rätsch, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 63–71.
51. Bajaj, I., Arora, A. & Hasan, M. M. F. "Black-Box Optimization: Methods and Applications," in *Springer Optimization and Its Applications*, vol. 170, P. M. Pardalos, V. Rasskazova, and M. N. Vrahatis, Eds. Cham: Springer International Publishing, 2021, pp. 35–65.
52. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
53. Ceder, G. A derivation of the Ising model for the computation of phase diagrams. *Comput. Mater. Sci.* **1**, 144–150 (1993).
54. Paulsen, J. M. & Dahn, J. R. Phase Diagram of Li—Mn—O Spinel in Air. *Chem. Mater.* **11**, 3065–3079 (1999).
55. Nelson, L. J., Hart, G. L. W., Zhou, F., & Ozoli, V. Compressive sensing as a paradigm for building physics models. *Phys. Rev. B* **87**, 35125 (2013).
56. Nelson, L. J., Ozoli\, V., Reese, C.S., Zhou, F. & Hart, G. L. W. Cluster expansion made easy with Bayesian compressive sensing. *Phys. Rev. B* **88**, 155105 (2013).
57. Blum, V., Hart, G. L. W., Walorski, M. J. & Zunger, A. Using genetic algorithms to map first-principles results to model Hamiltonians: application to the generalized Ising model for alloys. *Phys. Rev. B* **72**, 165113 (2005).
58. Hart, G. L. W., Blum, V., Walorski, M. J. & Zunger, A. "Evolutionary approach for determining first-principles hamiltonians,". *Nat. Mater.* **4**, 391–394 (2005).
59. Drautz, R., & Díaz-Ortiz, A. Obtaining cluster expansion coefficients in ab initio thermodynamics of multicomponent lattice-gas systems. *Phys. Rev. B* **73**, 224207 (2006).
60. Schmidt, D. J., Chen, W., Wolverton, C. & Schneider, W. F. Performance of cluster expansions of coverage-dependent adsorption of atomic oxygen on Pt(111). *J. Chem. Theory Comput.* **8**, 264–273 (2012).
61. Leong, Z. & Tan, T. L. Robust cluster expansion of multicomponent systems using structured sparsity. *Phys. Rev. B* **100**, 134108 (2019).
62. Mohr F. & van Rijn, J. N. Learning curves for decision making in supervised machine learning – A Survey. *CoRR*, abs/2201.12150, (2022).
63. Brownlee, J. *Better deep learning: train faster, reduce overfitting, and make better predictions*. Machine Learning Mastery, (2018).
64. Cortes, C., Jackel, L. D. & Chiang, W.-P. Limits on learning machine accuracy imposed by data quality. In *Advances in Neural Information Processing Systems*, 1994, vol. 7.
65. Cortes, C., Jackel, L. D., Solla, S., Vapnik, V., & Denker, J. Learning curves: asymptotic values and rate of convergence. In *Advances in Neural Information Processing Systems*, 1993, vol. 6.
66. Cheng, Y.-H., Liao, J.-H., Zhao, Y.-J. & Yang, X.-B. An extended cluster expansion for ground states of heterofullerenes. *Sci. Rep.* **7**, 16211 (2017).
67. Das, H., Urban, A., Huang, W. & Ceder, G. First-principles simulation of the (Li–Ni–Vacancy)O phase diagram and its relevance for the surface phases in Ni-rich Li-ion cathode materials. *Chem. Mater.* **29**, 7840–7851 (2017).
68. Chang, J. H. et al. Superoxide formation in Li2VO2F cathode material––a combined computational and experimental investigation of anionic redox activity. *J. Mater. Chem. A* **8**, 16551–16559 (2020).
69. L. Barroso-Luque, J. H. Yang, & G. Ceder. Sparse expansions of multicomponent oxide configuration energy using coherency \& redundancy. *arXiv Prepr. arXiv2109.06905*, (2021).
70. Sun, J., Ruzsinszky, A. & Perdew, J. P. Strongly constrained and appropriately normed semilocal density functional. *Phys. Rev. Lett.* **115**, 36402 (2015).
71. Zhang, Y. et al. Efficient first-principles prediction of solid stability: towards chemical accuracy. *npj Comput. Mater.* **4**, 9 (2018).
72. Yang, J. H., Kitchaev, D. A. & Ceder, G. Rationalizing accurate structure prediction in the meta-GGA SCAN functional. *Phys. Rev. B.* **100**, 35132 (2019).
73. Hinuma, Y., Hayashi, H., Kumagai, Y., Tanaka, I. & Oba, F. Comparison of approximations in density functional theory calculations: energetics and structure of binary oxides. *Phys. Rev. B* **094102**, 1–24 (2017).
74. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
75. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).
76. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
77. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
78. Zhang, W. et al. Kinetic pathways of ionic transport in fast-charging lithium titanate. *Sci. (80-.)* **367**, 1030 LP–1031034 (2020).
79. Seko, A., Togo, A., Oba, F. & Tanaka, I. Structure and stability of a homologous series of tin oxides. *Phys. Rev. Lett.* **100**, 45702 (2008).
80. Kelly, T. D., & Matos, G. R., comps., 2014, Historical statistics for mineral and material commodities in the United States (2016 version): U.S. Geological Survey Data Series 140, accessed [May 28, 2021].

## ACKNOWLEDGEMENTS

The project was primarily funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under Contract No. DE-AC02-05-CH11231 (Materials Project program KC23MP). This work was also supported by the Energy Efficiency and Renewable Energy, Vehicle Technologies Office, under the Applied Battery Materials Program, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. L.B.L and Z.J. gratefully acknowledge financial support from the NSF Graduate Research Fellowship Program (GRFP) under contract no. 1752814. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The research was performed using computational resources sponsored by the Department of Energy's Office of Energy Efficiency and Renewable Energy and located at the National Renewable Energy Laboratory.

## AUTHOR CONTRIBUTIONS

G.C. conceived and supervised the project. J.H.Y. did the magnetic charge assignments, SCAN calculations, regularization, and drafted the manuscript. T.C. did the structural mapping and corresponding write-up. L.B.L. refactored and wrote the smol package. All authors discussed, reviewed, and edited the manuscript.

## COMPETING INTERESTS

The authors declare no competing interests.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41524-022-00818-3.

**Correspondence** and requests for materials should be addressed to Gerbrand Ceder.