



Universiteit  
Leiden  
The Netherlands

## Construction of ground-state preserving sparse lattice models for predictive materials simulations

Huang, W.; Urban, A.; Rong, Z.; Ding, Z.; Luo, C.; Ceder, G.

### Citation

Huang, W., Urban, A., Rong, Z., Ding, Z., Luo, C., & Ceder, G. (2017). Construction of ground-state preserving sparse lattice models for predictive materials simulations. *Npj Computational Materials*, 3, 30. doi:10.1038/s41524-017-0032-0

Version: Not Applicable (or Unknown)

License: [Leiden University Non-exclusive license](#)

Downloaded from: <https://hdl.handle.net/1887/57892>

**Note:** To cite this publication please use the final published version (if applicable).

## ARTICLE OPEN

## Construction of ground-state preserving sparse lattice models for predictive materials simulations

Wenxuan Huang<sup>1</sup>, Alexander Urban<sup>2</sup>, Ziqin Rong<sup>1</sup>, Zhiwei Ding<sup>1</sup>, Chuan Luo<sup>3</sup> and Gerbrand Ceder<sup>1,2,4</sup>

First-principles based cluster expansion models are the dominant approach in *ab initio* thermodynamics of crystalline mixtures enabling the prediction of phase diagrams and novel ground states. However, despite recent advances, the construction of accurate models still requires a careful and time-consuming manual parameter tuning process for ground-state preservation, since this property is not guaranteed by default. In this paper, we present a systematic and mathematically sound method to obtain cluster expansion models that are guaranteed to preserve the ground states of their reference data. The method builds on the recently introduced compressive sensing paradigm for cluster expansion and employs quadratic programming to impose constraints on the model parameters. The robustness of our methodology is illustrated for two lithium transition metal oxides with relevance for Li-ion battery cathodes, i.e.,  $\text{Li}_{2x}\text{Fe}_{2(1-x)}\text{O}_2$  and  $\text{Li}_{2x}\text{Ti}_{2(1-x)}\text{O}_2$ , for which the construction of cluster expansion models with compressive sensing alone has proven to be challenging. We demonstrate that our method not only guarantees ground-state preservation on the set of reference structures used for the model construction, but also show that out-of-sample ground-state preservation up to relatively large supercell size is achievable through a rapidly converging iterative refinement. This method provides a general tool for building robust, compressed and constrained physical models with predictive power.

*npj Computational Materials* (2017)3:30; doi:10.1038/s41524-017-0032-0

## INTRODUCTION

First-principles density functional theory (DFT) calculations have established themselves as a routine and reliable tool in computational materials science research<sup>1–4</sup> and have enabled important advancements in materials discovery.<sup>1, 2, 5</sup> Although implementations with increasing numerical efficiency and growing computational power have made it possible to simulate ever larger structures with DFT, the method's intrinsic scaling with the number of electrons prevents applications that require large structures (thousands of atoms) and intensive sampling (millions of configurations). Approximate energy models fitted to DFT reference data, such as cluster expansion (CE) lattice models<sup>6–9</sup> or machine learning regression,<sup>10, 11</sup> can overcome these limitations by constructing computationally more efficient models with accuracies that are close to DFT for a chosen structural and chemical space. One prototypical application for approximate energy models is the prediction of ordered ground states based on an underlying lattice topology.<sup>12, 13</sup>

The concept of CEs goes back to the Ising model,<sup>14</sup> which describes the magnetic phases of an atomic lattice in terms of pair interactions. A CE model, or generalized Ising model, is the discrete sum representation of materials properties in terms of lattice site topologies and site interactions, such as site pairs, triplets, quadruplets, and so on. CEs have broad applications in different fields of science,<sup>6, 15</sup> such as magnetism<sup>15</sup> and alloy thermodynamics.<sup>6</sup>

The key challenge in constructing CE models is to determine the expansion coefficients, the effective cluster interactions (ECIs),

in a robust fashion through a fit to reference configurations.<sup>1, 16, 17</sup> Conventional ECI fitting procedures<sup>1, 16, 17</sup> focus on minimizing the overall difference between the CE fit and the input configurations with respect to the expanded quantity, such as the energy. In many cases, that input quantity may be determined by an accurate *ab-initio* method such as Density Functional Theory. One essential requirement that each CE fit must meet for practical applications is ground-state preservation, i.e., a physically accurate CE model must reproduce the ground states of the input if only the input configurations are considered. This requirement is important as the ground states usually govern the material properties at relevant temperature,<sup>18</sup> such as finite temperature voltage profiles<sup>18</sup> and phase diagrams.<sup>18</sup> In this article, we revisit the ECI fitting problem with a focus on constraints that guarantee ground-state preservation. We propose a robust and efficient scheme to construct ground-state preserving CE models based on compressive sensing<sup>17, 19</sup> and quadratic programming.<sup>20</sup>

The manuscript is organized as follows: First, we briefly review the CE formalism and define the ECI fitting problem in a rigorous mathematical manner. We then derive ground-state preserving constraints that can be used in conjunction with a quadratic programming solver. Afterwards, we consider the phase diagrams of two prototypical oxides of practical relevance as benchmark cases. Finally, we compare the strengths and weaknesses of our approach with established methods.

<sup>1</sup>Department of Material Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA; <sup>2</sup>Department of Materials Science and Engineering, UC Berkeley, Berkeley, CA, USA; <sup>3</sup>Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China and <sup>4</sup>Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Correspondence: Gerbrand Ceder (gceder@lbl.gov)

Received: 9 May 2017 Revised: 23 June 2017 Accepted: 5 July 2017

Published online: 07 August 2017

## RESULTS AND DISCUSSION

### Compressed sensing and cluster expansion

For a rigorous mathematical introduction to cluster expansions and their formal relationship to the partition function of crystalline solids, we refer the readers to references.<sup>6, 9, 21</sup> Here, we only illustrate the key features of cluster expansions that are of relevance to the present work.<sup>17</sup>

The general expression of a cluster expansion Hamiltonian is

$$\begin{aligned} E_{CE}(\sigma) = & J_0 + \sum_i^{\text{sites}} J_i \sigma_i + \sum_{i,j}^{\text{pairs}} J_{i,j} \sigma_i \sigma_j + \sum_{i,j,k}^{\text{triplets}} J_{i,j,k} \sigma_i \sigma_j \sigma_k \\ & + \dots = \sum_{c \in \mathbf{C}} J_c \sigma_c \end{aligned} \quad (1)$$

where  $\sigma$  is the spin representation of an atomic configuration in which each component  $\sigma_i$  (a spin variable) denotes the occupancy of site  $i$ . Following the Ising model convention,  $\sigma_i$  takes on values of  $\pm 1$  in a binary system, encoding the atomic species on site  $i$ . Each product of spin variables,  $\sigma_i \sigma_j \dots$ , (spin product) corresponds to a cluster of lattice sites, and the cluster expansion energy  $E_{CE}$  is a polynomial of the spin variables weighted by the expansion coefficients  $\mathbf{J}$ , the ECIs. For brevity, we denote the set of interacting clusters as  $\mathbf{C}$ . For any cluster  $c \in \mathbf{C}$ ,  $J_c$  is the corresponding ECI and  $\sigma_c$  is the corresponding spin product. Note that typically multiple clusters of the same *type* exist (e.g., the point term for each equivalent site or the cluster corresponding to the nearest-neighbor pair interaction), and symmetry requires the coefficients of equivalent clusters have to be identical.<sup>22</sup> The summation in Eq. (1) is therefore actually over cluster types, and the individual spin products can be replaced by their average over all equivalent clusters, the *cluster correlations*.

From Eq. (1) it is obvious that the CE energy is linearly dependent on the ECIs,  $\mathbf{J}$ , when the configuration  $\sigma$  is fixed. We can thus write

$$E_{CE}(\sigma) = \mathbf{\Pi}(\sigma) \mathbf{J} \quad (2)$$

where  $\mathbf{\Pi}(\sigma)$  is the row vector of cluster correlations (with multiplicity incorporated) corresponding to configuration  $\sigma$ . Given a set of input atomic configurations  $S$  and their DFT energies  $\mathbf{E}_{\text{DFT},S}$ , the problem of determining the ECIs can then be naively expressed as minimization of the  $L_2$  norm

$$\min_{\mathbf{J}} \|\mathbf{E}_{\text{DFT},S} - \mathbf{\Pi}_S \mathbf{J}\|_2 \quad (3)$$

where the rows of the feature matrix  $\mathbf{\Pi}_S$  are the cluster correlations of the configurations in  $S$ . Note that the  $L_2$  norm is the conventional Euclidean norm, and the general  $L_p$  norm  $\|\mathbf{u}\|_p$  is defined as:

$$\|\mathbf{u}\|_p = \left( \sum_i |u_i|^p \right)^{1/p} \quad (4)$$

Simply minimizing the  $L_2$  norm in Eq. (3) essentially means that the ECIs are fitted such that the average squared difference between the DFT energies and the CE-predicted energies of all structures is minimized. However, such a direct minimization of the error function leads to overfitting when the number of ECIs (the model parameters) exceeds or becomes close to the number of reference configurations (the fitting parameters), i.e., when the system of linear equations Eq. (3) is underdetermined. Overfitting means that the ECIs accurately reproduce the energies of the reference structures (in-sample data) but deliver poor generalization, i.e., the CE model does not reliably predict the energy of other unseen structures (out-of-sample data). A standard method to avoid overfitting is regularization,<sup>23</sup> i.e., the simultaneous minimization of the sum of the error function and the magnitude of the model parameters. Compressive sensing<sup>17, 19</sup> implements  $L_1$  norm regularization, which has been shown to be a nearly optimal

and robust way to reconstruct signals from a small number of data points.<sup>24</sup> The compressive sensing formulation of the cluster expansion problem is:

$$\min_{\mathbf{J}} \|\mathbf{E}_{\text{DFT},S} - \mathbf{\Pi}_S \mathbf{J}\|_2^2 + \mu \|\mathbf{J}\|_1 \quad (5)$$

where  $\mu$  is a parameter controlling the sparseness of the fit. A higher value of  $\mu$  shifts the weight towards minimizing the  $L_1$  norm, when  $\mu$  is small the minimization of the  $L_2$  error dominates. The  $L_1$  norm of a vector is a measure of the vector's sparseness,<sup>24</sup> thus larger  $\mu$  values result in fewer ECIs not equal to zero and thereby reduce overfitting. An optimal  $\mu$  value can be determined through minimizing the error of the CE model on unseen data.<sup>17</sup>

### Constrained cluster expansion models

For practical applications it is often desirable that a CE model preserves some invariants on the input data. For example, predicting the qualitative features of a phase diagram may require that the energetic order of all structures is exactly preserved while quantitative errors in the structural energies might be tolerable.<sup>25</sup> This is because the set of ground states and the ranking of excited states close in energy determines the topology of a phase diagram more than the actual energies themselves.<sup>26</sup> As the energy difference between competing structures is typically small, minimization of the average error in reproducing the DFT energy does not by itself enforce the structural energy order one wants to preserve. As a result even very small energy errors in the CE can qualitatively change a phase diagram when it leads to new ground states.<sup>26</sup> We have found practically that trying to preserve the structural ordering and ground states by increasing the relative weights of these input data rapidly leads to overfitting in the CE. In the following, we will develop a methodology that allows including constraints in the ECI optimization problem in a systematic, unbiased fashion and without overfitting.

In recent years, mathematical programming has been a rapidly growing field that enables the highly efficient, systematic and rigorous solution of problems in different standard forms.<sup>27</sup> One rapidly growing area is quadratic programming (QP),<sup>28</sup> for which robust solvers exist,<sup>20</sup> and a variety of different approaches have been researched and implemented, such as the interior point method, the active set method and the augmented Lagrangian method.<sup>28</sup> In essence, quadratic programming is a mathematical optimization technique for problems of the following specific form:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{c}^T \mathbf{x} \\ \text{s.t.} \quad & \mathbf{A} \mathbf{x} \leq \mathbf{b} \\ & \mathbf{C} \mathbf{x} = \mathbf{d} \end{aligned} \quad (6)$$

where  $\mathbf{Q}$  is a positive semidefinite matrix,  $\mathbf{A}$  and  $\mathbf{C}$  are real matrices, and  $\mathbf{b}$ ,  $\mathbf{c}$  and  $\mathbf{d}$  are real vectors. Note that a matrix is positive semidefinite if and only if for all real vectors  $\mathbf{X}$ ,  $\mathbf{x}^T \mathbf{Q} \mathbf{x} \geq 0$ . The semidefinite property is essential so that the optimization problem is convex. Also note that when  $\mathbf{Q} = 0$  Eq. (6) reduces to a standard linear programming problem which was introduced to CE optimization in reference.<sup>25</sup>

Our key strategy for CE fitting is to cast the compressive sensing problem Eq. (5) into a quadratic programming problem Eq. (6)<sup>29</sup> and to add constraints that guarantee ground-state preservation.

Explicitly, Eq. (5) can be rewritten as:

$$\begin{aligned}
 & \min_{\mathbf{J}} \|\mathbf{E}_{\text{DFT},S} - \Pi_S \mathbf{J}\|_2^2 + \mu \|\mathbf{J}\|_1 \\
 \Leftrightarrow & \min_{\mathbf{J}} \|\mathbf{E}_{\text{DFT},S} - \Pi_S \mathbf{J}\|_2^2 + \mu \sum_{c \in \mathbf{C}} |J_c| \\
 \Leftrightarrow & \min_{\mathbf{J}, \mathbf{z}} \|\mathbf{E}_{\text{DFT},S} - \Pi_S \mathbf{J}\|_2^2 + \mu \sum_{c \in \mathbf{C}} z_c \\
 & \text{s.t. } z_c \geq J_c \quad \forall c \in \mathbf{C} \\
 & z_c \geq -J_c \quad \forall c \in \mathbf{C}
 \end{aligned} \tag{7}$$

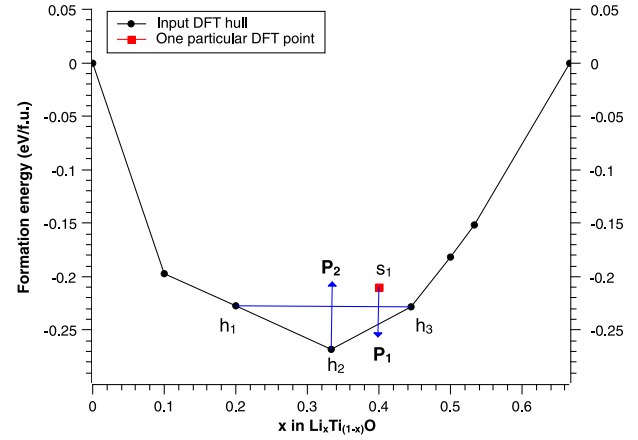
$$\begin{aligned}
 \Leftrightarrow & \min_{\mathbf{J}, \mathbf{z}} \mathbf{J}^T \Pi_S^T \Pi_S \mathbf{J} - 2 \mathbf{E}_{\text{DFT},S}^T \Pi_S \mathbf{J} + \mu \sum_{c \in \mathbf{C}} z_c + \mathbf{E}_{\text{DFT},S}^T \mathbf{E}_{\text{DFT},S} \\
 & \text{s.t. } z_c \geq J_c \quad \forall c \in \mathbf{C} \\
 & z_c \geq -J_c \quad \forall c \in \mathbf{C}
 \end{aligned} \tag{8}$$

In the conversion step in Eq. (7) auxiliary variables  $z_c$ , corresponding to constraints on  $\mathbf{J}$ , have been introduced to remove the  $L_1$  norm of Eq. (5). The equivalence in Eq. (7) holds because every  $z_c$  can be independently minimized while it is constrained to be larger than  $\pm J_c$ . Note that the QP formulation in Eq. (6) does not allow absolute value operations, so that two separate linear constraints are required in Eq. (7),  $z_c \geq J_c$  and  $z_c \geq -J_c$ , even though they are in combination essentially expressing the absolute value constraint  $z_c \geq |J_c|$ . The conversion step in Eq. (8) is a direct expansion of the  $L_2$  norm into vector multiplication. Note that  $\Pi_S^T \Pi_S$  is always positive semidefinite for every vector  $\mathbf{x}$ , since  $\mathbf{x}^T \Pi_S^T \Pi_S \mathbf{x} = (\Pi_S \mathbf{x})^T \Pi_S \mathbf{x} \geq 0$ . Hence, we have arrived at a formulation of the compressive sensing ECI problem Eq. (5) in terms of a QP problem.

The second key step of our methodology is to include suitable constraints for ground-state preservation in the QP formulation. Ground states, i.e., those configurations that are thermodynamically stable at zero temperature (0 K), can be identified by constructing the lower convex hull of the formation energies.<sup>30</sup> When the energy of a configuration is above the ground-state hull it is thermodynamically unstable with respect to decomposition into neighboring ground states.

Note that there are 2 different scenarios that lead to inconsistent ground states from an ECI fit: The first type of ground-state inconsistency occurs when the energy of some non-ground-state configuration is underestimated so much that it erroneously becomes a ground-state of the CE model. This problem is illustrated in Fig. 1 (labeled with  $P_1$ ), where the energy of configuration  $s_1$  is predicted to be below its decomposition line in the input data, i.e., below the convex combination of configurations  $h_2$  and  $h_3$  (shown as the line connecting the points). To constrain the QP system such that no inconsistency of type 1 occurs, we add the first constraint:

**(C1)** for each configuration that is not on the ground-state hull (i.e., a configuration that would thermodynamically decompose into ground states), we require that its CE configuration energy is greater than its CE decomposition line. To express this condition formally, we denote the  $i$ -th ground-state configuration (i.e., the  $i$ -th configuration on the lower convex hull) by  $\sigma_{h,i}$  with  $i \in H$ . With this notation, the decomposition of an unstable configuration  $\sigma_s$ ,



**Fig. 1** Schematic of the two types of ground-state inconsistencies that may arise during the fit of a cluster expansion (CE) model to DFT reference data. ( $P_1$ ) illustrates the situation in which one particular configuration,  $s_1$ , that is unstable based on the DFT input data becomes a stable ground-state of the CE model, as its CE energy is below the convex combination of its decomposition line defined by the ground-state configurations  $h_2$  and  $h_3$ . ( $P_2$ ) illustrates the converse situation in which the CE energy of one ground-state configuration,  $h_2$ , is greater than the convex combination of the neighboring ground states,  $h_1$  and  $h_3$ , which causes  $h_2$  to be unstable in the CE model

into the stable ground states can be expressed as  $Dec_H(\sigma_s) = \{x_i(\sigma_s)\sigma_{h,i}\}_{i \in H}$  where  $x_i(\sigma_s)$  is the fraction of  $\sigma_{h,i}$  in the decomposition products. The constraint to remove ground-state inconsistencies of type 1 becomes

$$\Pi(\sigma_s)\mathbf{J} \leq \sum_{i \in H} x_i(\sigma_s)\Pi(\sigma_{h,i})\mathbf{J} + \varepsilon \text{ for all unstable configurations } \sigma_s \tag{9}$$

where  $\varepsilon$  is some small number used as numerical tolerance.

Introducing constraint (C1) in Eq. (9) to the QP problem in Eq. (8) guarantees that all ground states of the CE model are also ground states of the DFT input data. However, the converse is not necessarily true, i.e., a DFT ground-state configuration might not be a ground-state of the CE model. This scenario is shown in Fig. 1 ( $P_2$ ), where configuration  $h_2$  has a greater CE energy than its convex hull decomposition line defined by  $h_1$  and  $h_3$ . To remove this second type of ground-state inconsistency, we introduce a second constraint:

**(C2)** for each ground-state configuration  $\sigma_h$  (i.e., for each configuration  $\sigma_h$  on the lower convex hull), we require that its energy is smaller than the energy of a modified hull that results from removing  $\sigma_h$  from the set of input ground states. Formally, given a ground-state configuration  $\sigma_h$  on the input hull, we consider its decomposition into a modified ground-state hull as  $Dec_{H \setminus h}(\sigma_h) = \{x_{i, H \setminus h}(\sigma_h)\sigma_{h,i}\}_{i \in H \setminus h}$  where  $H \setminus h$  is the index set of all input hull configurations not including  $\sigma_h$ , and  $x_{i, H \setminus h}(\sigma_h)$  is the fraction of decomposition product  $\sigma_{h,i}$ . The constraint to remove ground-state inconsistencies of type 2 thus becomes

$$\Pi(\sigma_h)\mathbf{J} \leq \sum_{i \in H \setminus h} x_{i, H \setminus h}(\sigma_h)\Pi(\sigma_{h,i})\mathbf{J} - \varepsilon \tag{10}$$

for all ground state configurations  $\sigma_h$

Constraint (C2) in Eq. (10) guarantees that all ground-state configurations in the (DFT) input data are also ground states of the CE model. Consequently, by combining (C1) and (C2), a configuration is a ground-state of the resulting CE model if and only if it is a ground-state of the input data. The full quadratic

4 programming formulation for ground-state preserving CE fitting is

$$\min_{\mathbf{J}, \mathbf{z}} \mathbf{J}^T \Pi_S^T \Pi_S \mathbf{J} - 2 \mathbf{E}_{\text{DFT}, S}^T \Pi_S \mathbf{J} + \mu \sum_{c \in \mathbf{C}} z_c + \mathbf{E}_{\text{DFT}, S}^T \mathbf{E}_{\text{DFT}, S}$$

$$\text{s.t. } z_c \geq J_c \quad \forall c \in \mathbf{C}$$

$$z_c \geq -J_c \quad \forall c \in \mathbf{C}$$

$$\Pi(\sigma_s) \mathbf{J} \geq \sum_{i \in H} x_i(\sigma_{h,i}) \Pi(\sigma_{h,i}) \mathbf{J} + \varepsilon \quad (11)$$

for all unstable configurations  $\sigma_s$

$$\Pi(\sigma_h) \mathbf{J} \leq \sum_{i \in H \setminus h} x_{i, H \setminus h}(\sigma_{h,i}) \Pi(\sigma_{h,i}) \mathbf{J} - \varepsilon$$

for all ground state configurations  $\sigma_h$

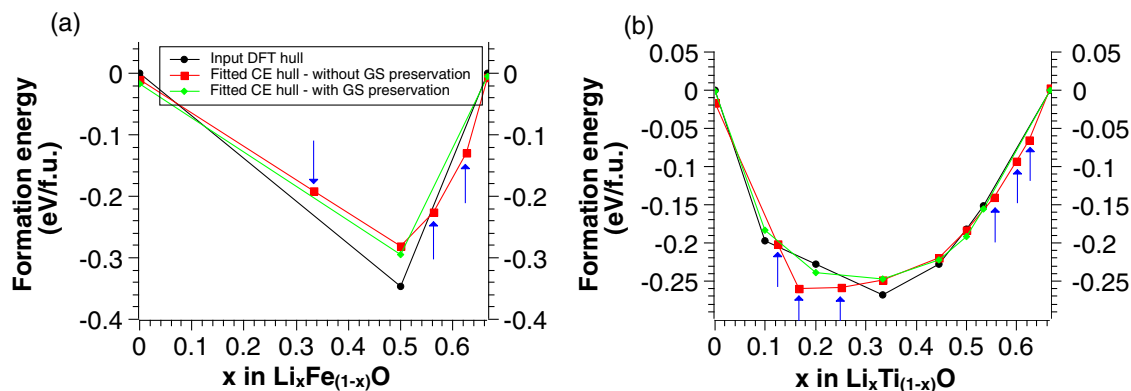
Cation ordering in the rocksalt-type lithium transition metal oxide systems  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$  and  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$

We demonstrate the effectiveness of the QP approach to cation ordering in two oxide systems. Rocksalt-type lithium transition metal oxides,  $\text{LiMO}_2$  ( $M = \text{one or more transition metal species}$ ), are the most important class of cathode materials for lithium-ion batteries in consumer electronics.<sup>31</sup> During the last decade, materials with lithium excess compositions,  $\text{Li}_{(1+x)}\text{M}_{(1-x)}\text{O}_2$ , have attracted much interest owing to their high lithium storage capacities.<sup>32, 33</sup> One criterion for the suitability of  $\text{Li}_{(1+x)}\text{M}_{(1-x)}\text{O}_2$  as cathode material is whether the material is a sufficiently good conductor for Li ions, which critically depends on the cation (Li, M) ordering in the structure.<sup>34, 35</sup> While conventional oxide-based cathode materials form in ordered crystal structures (such as layered  $\text{LiCoO}_2$ <sup>36</sup>), several cation-disordered lithium-excess materials with high practical capacities have recently been discovered.<sup>35, 37</sup> Some of these new compositions contain Ti<sup>38</sup> and Fe<sup>37</sup> which makes them attractive for technological applications because of the metals' high abundance and nontoxicity. However,  $\text{LiTiO}_2$ <sup>39, 40</sup> and  $\text{LiFeO}_2$ <sup>41, 42</sup> are the only  $\text{LiMO}_2$  with single transition metal species that form in cation-disordered structures

in solid-state synthesis, and consequently their configurational phase diagrams are challenging to investigate experimentally.

In the following we employ the ground-state preserving QP methodology developed above to investigate the phase diagrams of  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$  and  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$ <sup>39-42</sup> to obtain a better understanding of the relevant atomic configurations. The input consists of 863 and 602 reference configurations for  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$  and  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$  respectively. DFT calculations for  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$  configurations were performed within the Hubbard-U corrected Generalized Gradient Approximation (GGA+U), using the PBE exchange-correlation functional.<sup>43, 44</sup> The U values are taken from the work of Jain et al.<sup>45</sup> DFT calculations for  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$  configurations did not employ a Hubbard-U correction. For both systems, an initial set of configurations at  $x = 0.5$  with supercell sizes up to 8 sites was generated using the enumerating algorithms by Hart et al.,<sup>46</sup> and the reference sets were subsequently refined by including ground-state configurations of preliminary cluster expansions determined using a recently published ground-state search algorithm for lattice models.<sup>47</sup> The corresponding ground-state input hulls are shown in Fig. 2 as black dots and lines. We note that both systems,  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$  and  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$ , cannot easily be fitted using the conventional (unconstrained) compressive sensing technique, as the approach gives rise to a number of spurious ground states as shown in Fig. 2 (with optimal  $\mu$  parameter as will be discussed below). Specifically, some  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$  configurations with  $x = 1/3, 5/8$  and  $9/16$  and  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$  configurations with  $x = 1/8, 1/6, 1/4, 5/9, 3/5, 5/8$  are erroneously predicted to be ground states (i.e., inconsistencies of type 1 as defined above). These over-stabilized configurations are marked with arrows in Fig. 2. In addition, the actual  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$  ground-state configurations with  $x = 1/10, 1/5, 8/15$  become unstable in the compressive sensing CE model (inconsistencies of type 2). These examples demonstrate that ground-state preservation is not an automatic feature inherent to the compressive sensing approach, and the problem needs to be addressed before predictive simulations of materials systems are possible.

As seen in Fig. 2, the QP fitting scheme achieves ground-state preservation for both materials and yields CE hulls that are spanned by the same configurations as the input DFT hulls. However, it is worth noting that the sparseness parameter,  $\mu$ , of Eq. (11) needs to be carefully selected to arrive at this result. In the following section we will show that  $\mu$  should be chosen such that the cross-validation error is minimized. The discussion of the cross-validation error is essential in that it provides a standard measure of predictive power of our fitting scheme which sets the method apart from other approaches for ground-state preservation, such



**Fig. 2** Input hull as obtained from density functional theory (black lines and circles), cluster expansion hull fitted using the conventional (unconstrained) compressive sensing method (red lines and squares), and ground-state preserving cluster expansion hull fitted with the quadratic programming method of the present work (green lines and diamonds) for (a)  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$  and (b)  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$ . The input ground-state configurations for (a)  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$  correspond to  $\text{FeO}$ ,  $\text{LiFeO}_2$  and  $\text{Li}_2\text{FeO}_3$  and for (b)  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$  correspond to  $\text{TiO}$ ,  $\text{Li}_1\text{Ti}_9\text{O}_{10}$ ,  $\text{Li}_2\text{Ti}_8\text{O}_{10}$ ,  $\text{Li}_3\text{Ti}_6\text{O}_9$ ,  $\text{Li}_4\text{Ti}_5\text{O}_9$ ,  $\text{Li}_2\text{TiO}_3$ ,  $\text{LiTiO}_2$ ,  $\text{Li}_8\text{Ti}_7\text{O}_{15}$  and  $\text{Li}_2\text{TiO}_3$ . Arrows are used to mark the over-stabilized unstable configurations. The sparseness parameter,  $\mu$ , was chosen to be 0.144 for both systems



as the adjustment of configuration weights, which will also be shown below.

Cross-validation of the choice of the sparseness parameter

Cross-validation is the standard way to decide the optimal sparseness of a numerical model, which is generally referred to as bias variance trade-off in statistical inference.<sup>48</sup> To determine the sparseness parameter  $\mu$  by means of cross-validation we randomly split the DFT data,  $D$ , into  $N = 10$  equal parts. For each part  $D_i$ , we define its complement  $\bar{D}_i$  as all the DFT data points except those belonging to  $D_i$  (formally,  $\bar{D}_i \equiv D - D_i$ ). Next, the QP scheme of Eq. (11) is applied to the complement set  $\bar{D}_i$  to obtain a CE fit without the information in part  $D_i$ , so that an out-of-sample validation can be performed by calculating the root mean square error (RMSE) of the unseen data  $D_i$ . We denote the resulting out-of-sample RMSE as  $e_{i,\mu}$ . The cross-validation (cv) score  $cv_\mu$  given a sparseness parameter  $\mu$  is then defined as the root mean square of the out-of-sample RMSE over all  $N$  data parts, i.e., formally

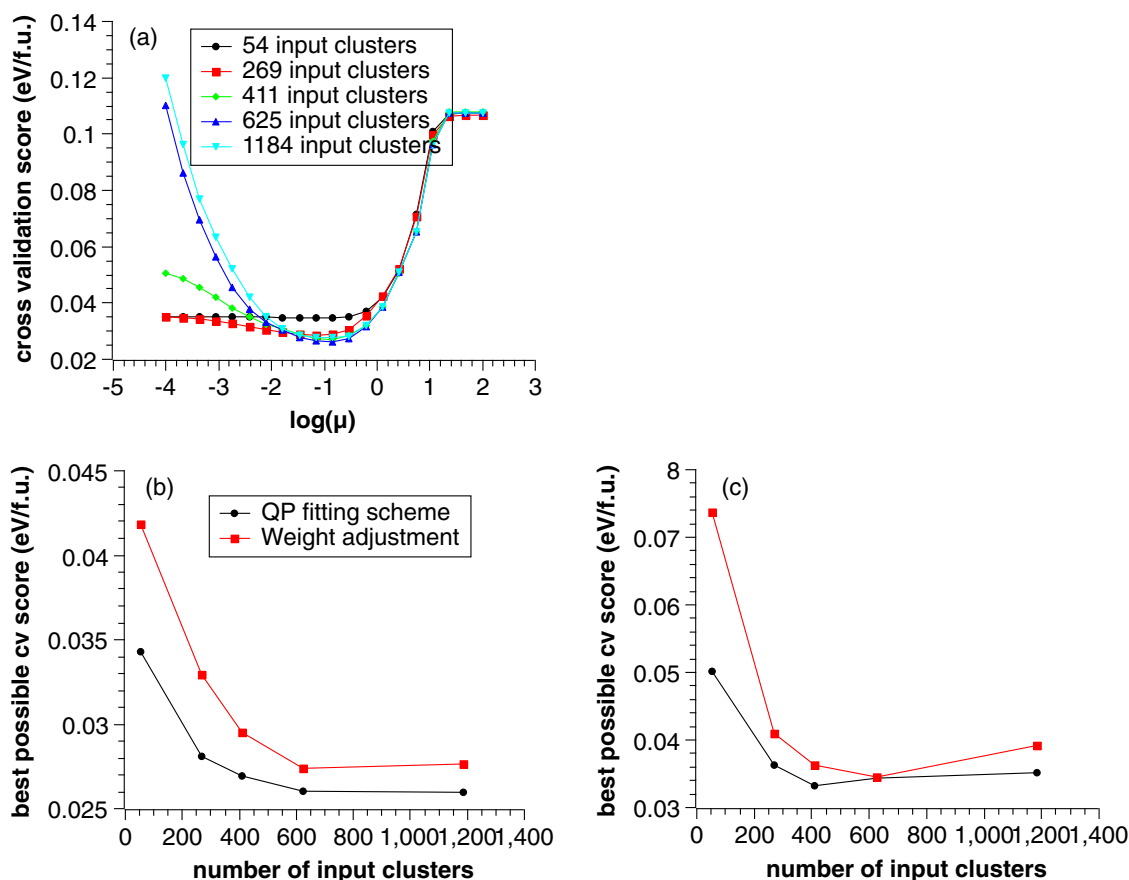
$$cv_\mu = \sqrt{\sum_{i=1 \dots N} e_{i,\mu}^2 / N}.$$

Using this definition, the optimal  $\mu$  resulting in the model with greatest predictive power can be determined by plotting  $cv_\mu$  against  $\mu$  and selecting the value of  $\mu$  that minimizes the cv score.

The cross-validation score  $cv_\mu$  of the  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$  system as function of  $\mu$  is shown in Fig. 3a for various different numbers of input clusters. Note that the number of input clusters to draw from is determined by a maximum interaction order (e.g., triplets) and a radial cutoff. Across all five curves, the cv score initially decreases and then increases with increasing  $\mu$ . We consider the concept of

bias variance tradeoff<sup>48</sup> to understand this behavior: The input DFT energies may conceptually be understood as the sum of an ideal cluster expansion and a certain degree of noise  $\varepsilon$ , i.e.,  $E_{\text{DFT}} = \Pi(\sigma)J + \varepsilon$ . Here, the noise could originate from numerical errors in the DFT energies. For small values of  $\mu$ , the CE fit uses all available degrees of freedom (i.e., all ECIs) to incorporate the noise  $\varepsilon$  into the CE model, resulting in severe overfitting. As the value of  $\mu$  increases, the number of non-zero ECIs decreases and the effect of noise, i.e., the variance in fitting, becomes less severe. However, when  $\mu$  becomes too large, the bias that ECIs should tend to 0 becomes dominating over the data itself, resulting in severe underfitting and thus increasing cv scores. As a consequence, the cv score has a pronounced minimum allowing to determine the optimal  $\mu$  corresponding to the best tradeoff between the variance and bias during fitting.

As seen in Fig. 3a, for small values of  $\mu$  where  $\log(\mu) < -2$  the cv score increases dramatically with the number of input clusters indicating overfitting as the result of insufficient regularization. As the sparseness parameter increases above  $\log(\mu) > -2$ , the cv score becomes less sensitive with respect to the number of input clusters, indicating that the regularization is effective and that most non-essential ECIs are fitted to zero regardless of the number of input clusters. The optimal cv scores are found for  $-1 \leq \log(\mu) \leq 0$  and are plotted in Fig. 3b for different numbers of input clusters (labeled “QP methodology”). As seen in the figure, the optimal cv score decreases from 0.0345 eV/f.u. (formula unit) for 54 input clusters to 0.0261 eV/f.u. for 625 input clusters. The cv score stabilizes at 625 input clusters and barely changes for 1184 input clusters (0.0260 eV/f.u.). Hence, we conclude that 625 input clusters and a sparseness parameter  $\mu = 0.144$  result in a CE model



**Fig. 3** **a** Cross-validation (cv) score  $cv_\mu$  plotted as function of the sparseness parameter  $\mu$  for the example of the  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$  system for various numbers of input clusters. The optimal  $\mu$  minimizes the cv score. **b**, **c** Optimal cv score for different numbers of input clusters as determined using the QP methodology and the conventional weight adjustment scheme for **(b)**  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$  and **(c)**  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$

with optimal predictive power for the  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$  system. The corresponding analysis for the  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$  system is shown in Fig. 3c, and the optimal parameters,  $\mu = 0.144$  and 411 input clusters, yield a cv score of 0.0331 eV/f.u.

In-sample ground-state preservation and comparison with conventional weight adjustment

Per construction, the QP form Eq. (11) guarantees that the CE fit preserves the ground states of the reference data set. Conventionally, such in-sample ground-state preservation is often achieved by assigning weights to the reference configurations to manually bias the fit. In the following, we compare the performance of the QP methodology with the conventional weight adjustment technique to further assess the utility of our approach. Before we detail the weight adjustment method, we briefly consider how configuration weights can be included in the QP approach in practice. For this purpose, we define a diagonal weight matrix  $\mathbf{W}$  whose diagonal entries  $w_{i,i}$  correspond to the weight of the  $i^{\text{th}}$  input configuration. With this definition,  $\mathbf{W}$  can be incorporated into Eq. (5) to achieve multiplying weights to the in-sample fitting error:

$$\min_{\mathbf{J}} \|\mathbf{W}(\mathbf{E}_{\text{DFT},S} - \mathbf{\Pi}_S \mathbf{J})\|_2^2 + \mu \|\mathbf{J}\|_1. \quad (12)$$

Note that large  $w_{i,i}$  result in a strong bias of the fitting error for the  $i^{\text{th}}$  input configuration to be 0. The concrete weight adjustment procedure that we employed in this work is as follows:

- (1) Initialize all weights to be 1.
- (2) Perform QP to construct a CE model.
- (3) Check if the CE model preserves in-sample ground states. If it does, the ground-state preserving fit is completed. If it does not, we define the set  $\mathbf{T}$  of all DFT and CE ground-state configurations

$$\mathbf{T} \equiv \{i \in S : S_i \text{ is DFT Ground state or CE Ground State}\} \quad (13)$$

Further, we define the maximum CE hull as  $err_{\text{hull}} \equiv \max_{j \in \mathbf{T}} |\mathbf{E}_{\text{CE},j} - \mathbf{E}_{\text{DFT},j}|$  and introduce a weight-increment set  $\mathbf{T}' \equiv \{i \in \mathbf{T} : |\mathbf{E}_{\text{CE},i} - \mathbf{E}_{\text{DFT},i}| > 0.5err_{\text{hull}}\}$ . For each configuration  $i \in \mathbf{T}'$ ,  $w_{i,i}$  is increased  $\sqrt[4]{2} \approx 1.19$  times. The procedure is continued with step (2).

This weight adjustment scheme guarantees that in-sample ground states are preserved, since it iteratively converges the CE hull to the DFT hull and corrects spurious ground-state configurations.

A comparison of the optimal cv scores obtained for different numbers of input clusters using both methods is shown in Figs. 3b, c. The cv score is, once again, used as a standard measure for the predictive power of the fits. In case of the  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$  system, the predictive power of the QP fit is consistently better than the fit obtained using the weight adjustment technique, and the improvement of the cv score is generally found to be around 2 meV/f.u. or 10%. For the  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$  system, the cv score of the QP CE fit also improves about 3 meV/f.u. or 10% of the CE fit from weight adjustment, except for 625 input clusters for which both methods give equivalent results. However, considering all numbers of input clusters, the overall best cv score for the QP method is 1.5 meV/f.u. or 5% better. While absolute energy errors on the order of a few meV/f.u. are close to the inherent error of density functional theory, similar errors in the relative energies of different configurations may add up and thereby give rise to qualitatively different phase diagrams.

Generally, we observed that the weight adjustment method biases some configurations by more than a factor of one thousand ( $w_{i,i} > 10^3$ ), resulting in overfitting of those particular configurations, whereas the QP scheme shows no evidence of such a partial overfitting.

In summary, we conclude that the QP methodology of this work has significant advantages over conventional weight adjustment

for the preservation of in-sample ground states. However, as will be demonstrated in the following section, the superiority of the QP approach becomes truly evident when out-of-sample configurations are considered.

Out-of-sample ground-state preservation

Suppose that we are certain that the set of reference configurations available for the construction of the CE model comprises all physical ground states of the system. This situation could occur after an extensive exploration of the configurational space or when the DFT data agrees exceptionally well with experiment. With such confidence in the reference data set, we would like to guarantee that the fitted CE model not only reproduces the ground states of the reference data, but also does not possess any additional ground states that are not already present in the reference data. We call this property out-of-sample ground-state preservation. In the following, we describe an iterative procedure for constructing CE models that guarantee out-of-sample ground-state preservation up to a given number of periodic sites. We will further show that this procedure is generally a useful strategy to construct CE models even when, initially, it is not known whether ground states outside of the reference set exist.

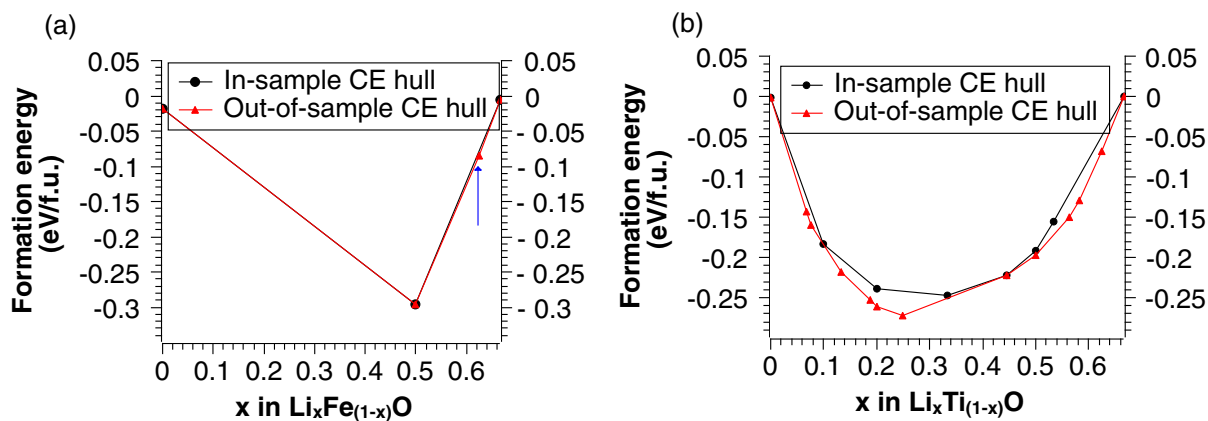
The QP formulation established in Eq. (11) provides ground-state preservation within the set of input data. However, out of sample ground-state preservation is not guaranteed. In principle, if the true configuration polytope,<sup>49</sup>  $P$ , is known for a set of possible ECIs, i.e.,  $\sigma \in P$  can be added and solved within a QP, one could add the following constraint:

$$\mathbf{\Pi}(\sigma) \mathbf{J} \geq \sum_{i \in H} x_i(\sigma) \mathbf{\Pi}(\sigma) \mathbf{J} \quad \forall \sigma \in P \quad (14)$$

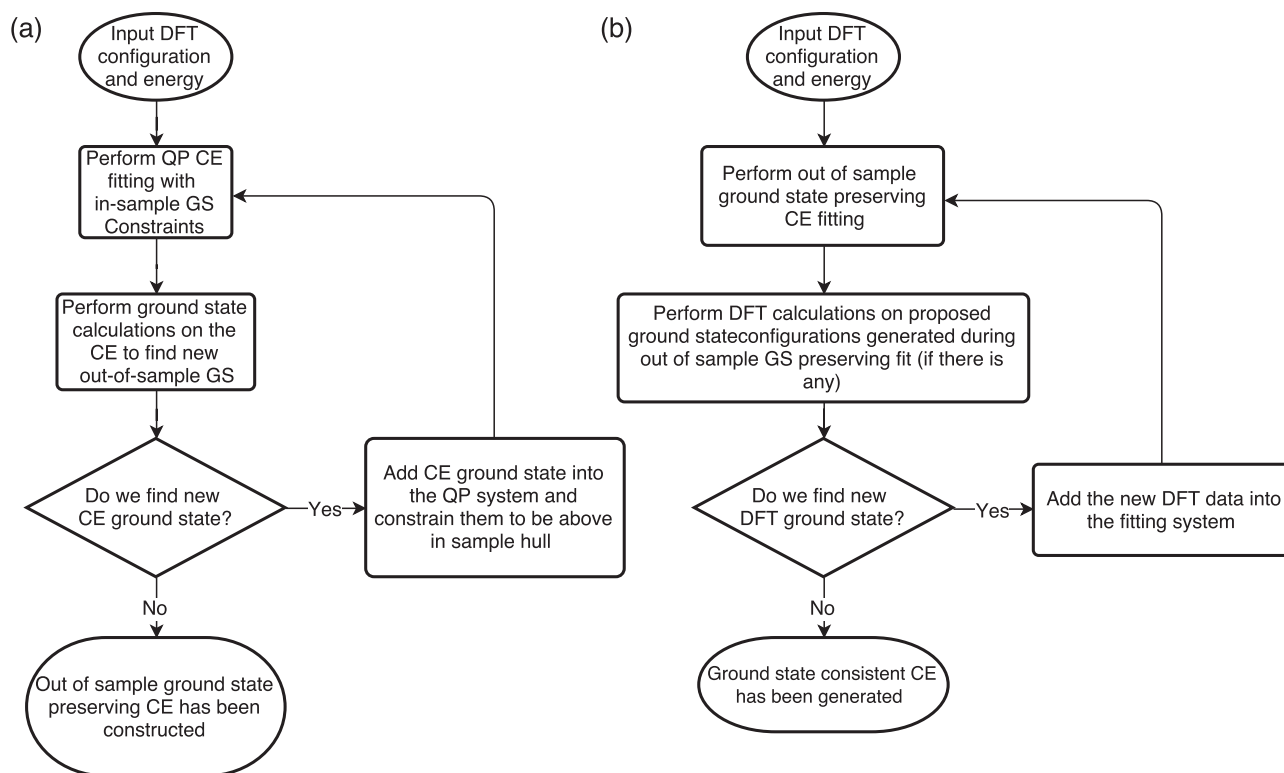
to Eq. (11) and the corresponding optimization problem will result in a globally ground-state preserved CE fit. In practice, however, solving the configurational polytope for an arbitrary CE is an undecidable problem.<sup>50</sup> Although this does not necessarily mean that finding a ground-state preserving fit is globally undecidable as well, this fact hints at the intrinsic difficulty of the out-of-sample ground-state preservation problem.

Instead of determining *a priori* constraints that guarantee out-of-sample preservation, we first examine a CE fit with in-sample ground-state preservation obtained from the QP methodology with optimal parameters (sparseness, number of input clusters) and determine all ground states of the CE model up to a defined system size using the methodology of reference.<sup>47</sup> The ground-state hull defined by the input configurations is denoted as the in-sample hull whereas we refer to the hull that is based on all identified ground states as the out-of-sample hull. A comparison of the in-sample and out-of-sample hulls for  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$  and  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$  for supercell sizes with up to 16 sites is shown in Fig. 4. For  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$ , one extra ground-state at  $x = 5/8$  is identified that is predicted to be 6 meV below the in-sample hull. Even though the distance between the in-sample and out-of-sample hulls is small (6 meV), this CE would produce a qualitatively wrong phase diagram due to the spurious ground-state at  $x = 5/8$ . For  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$ , the discrepancy between the in-sample CE hull and the out-of-sample CE hull is even more severe, as shown in Fig. 4b.

In the following we would like to arrive at a scheme to construct a CE that does not lead to additional ground states, i.e., out-of-sample ground-state preservation. Such scheme is useful in efficient determination of new ground-state configurations and self-consistent CE. Instead of determining the true configurational polytope of Eq. (14), we arrive at a CE model with out-of-sample ground-state preservation iteratively by determining the global ground states of preliminary CE models (as above) to identify those configuration  $\sigma \in P$  for which Eq. (14) is not satisfied. Afterwards, without additional DFT calculations, the constraint corresponding to these configurations  $\sigma$  are added to the QP form



**Fig. 4** Comparison of the convex hulls based only on the reference input data (in-sample CE hull) and based on all ground-state configurations of the cluster expansion model (out-of-sample CE hull) for (a)  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$  and (b)  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$ . The  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$  out-of-sample hull is below the in-sample hull with a maximal distance of 6 meV. An arrow is added to emphasize the out of sample ground-state at  $x = 0.625$  for  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$



**Fig. 5** **a** Flow chart of the iterative procedure for constructing out-of-sample ground-state preserving cluster expansion models. **b** Flow chart of a combined DFT-CE configurational sampling resulting in the construction of ground-state consistent cluster expansion models

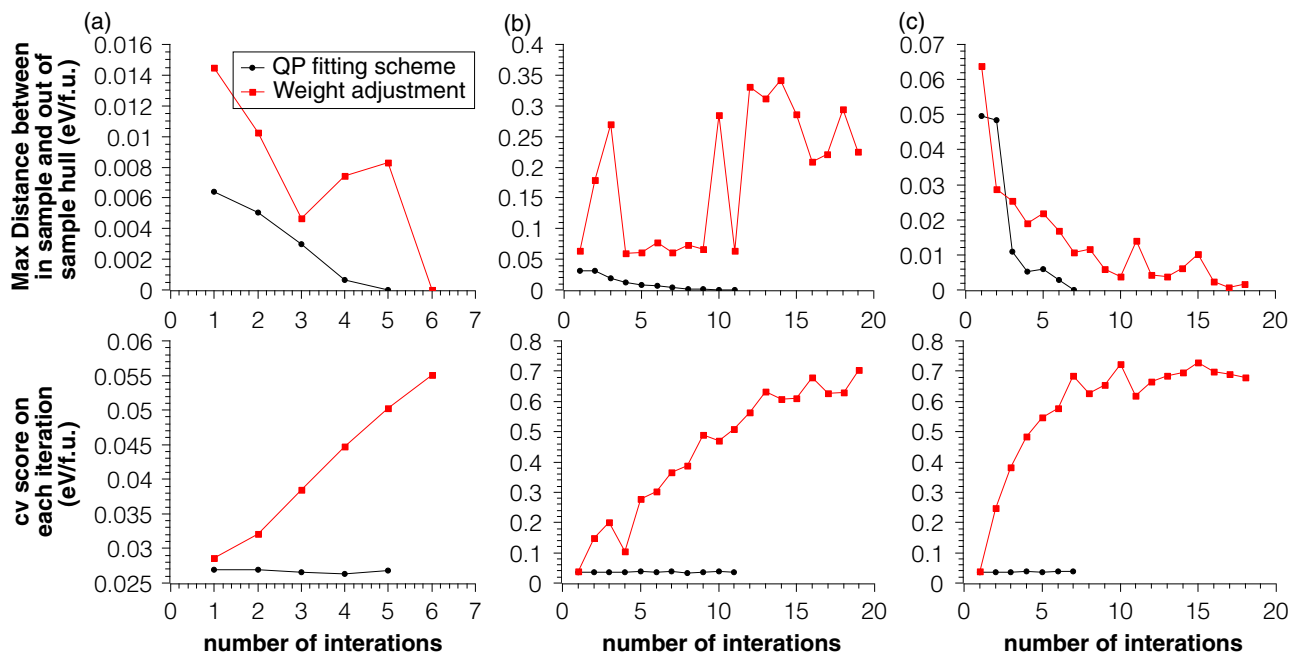
as in Eq. (14). By iteratively calculating the ground-state hulls and adding further constraints, global ground-state preservation up to a large supercell size can be achieved. The procedure is illustrated in Fig. 5a.

To demonstrate the convergence of this iterative refinement, we applied the procedure to the two model systems for supercell size of up to 16 sites. The weight adjustment procedure described above is used for comparison. Small initial weights,  $10^{-4}$ , and energies of about 1 meV above the hull are assigned to the predicted new ground states. The results are shown in Fig. 6a for  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$  and Fig. 6b for  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$ . For reference, the figure also shows the results of an iterative refinement using the weight adjustment method. The maximum distance between the in-sample hull and the out-of-sample hull is plotted in the upper

panel as a measure of the difference between the two hulls as the iteration progresses. The corresponding cv score is plotted in the lower panel as a measure of the predictive power of the CE fit.

As seen in Fig. 6, for both systems,  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$  and  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$ , the maximum distance (defined as the difference of energy under the same  $x$ ) between the out-of-sample and in-sample hulls decreases monotonously to 0 with the QP methodology. The iterative weight adjustment also converges for  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$ , though the distance between the hulls fluctuates and does not decay monotonously. For  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$  the weight adjustment method does not converge. More importantly, the cv scores of the QP fits are nearly constant throughout the iterations, whereas the cv score continuously increases for the weight adjustment algorithms. This means that, using the QP methodology, out-of-sample ground-state





**Fig. 6** Results of the iterative construction of cluster expansion models with out-of-sample ground-state preservation for (a)  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$ , (b)  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$  based on the exact ground states up to a supercell size of 16 sites and (c)  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$  based on the simulated annealing with a supercell size of  $8 \times 8 \times 8$  (512 sites). The upper panels show the maximal distance between the in-sample and out-of-sample hulls, and the lower panels show the corresponding cross-validation scores. For comparison, the corresponding results using iterative adjustment of configuration weights is shown as red lines and squares

preservation can be achieved without sacrificing the predictive power of the CE fit. On the other hand, the weight adjustment technique that is often used for CE construction is not guaranteed to converge and tends to achieve ground-state preservation at the cost of predictive power (increasing cv score). We therefore conclude that the QP methodology developed in the present work allows for the systematic construction of CE models with in-sample and out-of-sample ground-state preservation.

The results above are based on an exact ground-state search for system sizes of up to 16 sites, however, for the purpose of phase diagram calculations via Monte Carlo simulations much larger supercell sizes may be required. To construct CE models that are in practice ground-state preserving even for sufficiently large system sizes, the exact ground-state search may be replaced by simulated annealing simulations, which allow to determine plausible ground states for larger supercell size (but cannot provide proof that all ground states have been identified, see ref. <sup>47</sup> for a more detailed discussion). We repeated the iterative procedure of Fig. 5a for  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$  for supercell sizes with up to 512 sites using simulated annealing, and the results are depicted in Fig. 6c.

As shown in Fig. 6a for smaller cell sizes, using the QP methodology, the distance between the in-sample and out-of-sample CE hulls decreases monotonously to 0 within 7 iterations, and the cv score remains nearly constant. As before, for the iterative weight adjustment algorithm does not achieve complete convergence even after 18 iterations and gives rise to a dramatic increase of the cv score. This final example demonstrates again that the QP methodology is a robust scheme to obtain ground-state preserving CE fits even for large system sizes that are suitable for realistic Monte Carlo simulations.

Finally, we point out that the iterative procedure for out-of-sample ground-state preservation is not only useful, when the ground states of the system are known *a priori*. Instead, the procedure may also serve as a means for the sampling of the configurational space to generate additional reference data. For this purpose, the configurations that were identified as “spurious” ground states may be evaluated with the reference

method (i.e., DFT) to confirm whether any unknown ground-state has been discovered. By construction, this approach also provides a good stopping criterion for the cluster expansion fit when no additional ground states have been identified. This procedure is illustrated in Fig. 5b. If DFT calculations for all prospective new ground states are carried out and none of them turns out to be an actual ground-state, the out-of-sample ground-state preserving fit has the correct assumption and the resulting CE fit is a valid fit with consistently low cv errors. No further iteration is necessary, and the CE fit is finalized. On the other hand, if additional DFT ground states are found within the proposed set, then the out of sample ground-state preserving fit would have to be re-started.

To summarize, in this article, we presented a robust and efficient procedure to obtain ground-state preserving cluster expansion models. The method is formulated in terms of quadratic programming and compressive sensing and is mathematically rigorous. We demonstrated the robustness of the approach by application to the phase diagrams of  $\text{Li}_x\text{Fe}_{(1-x)}\text{O}$  and  $\text{Li}_x\text{Ti}_{(1-x)}\text{O}$  that are challenging to describe with conventional cluster expansion techniques. We further showed that out-of-sample ground-state preservation can be achieved up to large supercell sizes. These properties make the presented quadratic programming approach an attractive tool for the fit of general constraint lattice models and point the way towards the fully automated construction of cluster expansion models for materials simulations.

#### Data availability statement

The data that support the findings of this study have been deposited in Open Science Framework with the identifiers (DOI: 10.17605/OSF.IO/6DEHY).<sup>51</sup>

#### ACKNOWLEDGEMENTS

This work was supported primarily by the US Department of Energy (DOE) under Contract No. DE-FG02-96ER45571.

## AUTHOR CONTRIBUTIONS

W.H. and G.C. jointly proposed this project. W.H. is the lead on the project. W.H, A.U., Z.R., Z.D. and G.C. contributed to the theoretical and mathematical development of the quadratic programming method. W.H. and A.U. carried out the DFT calculations. W.H., A.U., Z.R., Z.D. and C.L. contributed to the implementation of the method. All authors contributed to the manuscript writing.

## ADDITIONAL INFORMATION

**Competing interests:** The authors declare no competing financial interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## REFERENCES

1. Fischer, C. C., Tibbetts, K. J., Morgan, D. & Ceder, G. Predicting crystal structure by merging data mining with quantum mechanics. *Nat. Mater.* **5**, 641–646 (2006).
2. Yang, K., Setyawan, W., Wang, S., Buongiorno Nardelli, M. & Curtarolo, S. A search model for topological insulators with high-throughput robustness descriptors. *Nat. Mater.* **11**, 614–619 (2012).
3. Rong, Z. *et al.* Materials design rules for multivalent ion mobility in intercalation structures. *Chem. Mater.* **27**, 6016–6021 (2015).
4. Liu, M. *et al.* Spinel compounds as multivalent battery cathodes: a systematic evaluation based on ab initio calculations. *Energy Environ. Sci.* **8**, 964–974 (2015).
5. Jain, A., Shin, Y. & Persson, K. A. Computational predictions of energy materials using density functional theory. *Nat. Rev. Mater.* **1**, 15004 (2016).
6. Sanchez, J. M., Ducastelle, F. & Gratias, D. Generalized cluster description of multicomponent systems. *Phys. A Stat. Mech. Appl.* **128**, 334–350 (1984).
7. van de Walle, A. & Ceder, G. Automating first-principles phase diagram calculations. *J. Phase Equilib.* **23**, 348–359 (2002).
8. De Fontaine, D. Configurational thermodynamics of solid solutions. *Solid State Phys.* **34**, 73–274 (1979).
9. De Fontaine, D. Cluster approach to order-disorder transformations in alloys. *Solid State Phys.* **47**, 33–176 (1994).
10. Artrith, N. & Urban, A. An implementation of artificial neural-network potentials for atomistic materials simulations: performance for TiO<sub>2</sub>. *Comput. Mater. Sci.* **114**, 135–150 (2016).
11. Artrith, N., Hiller, B. & Behler, J. Neural network potentials for metals and oxides—First applications to copper clusters at zinc oxide. *Phys. Status Solid.* **250**, 1191–1203 (2013).
12. Nahas, S., Ghosh, B., Bhowmick, S. & Agarwal, A. First-principles cluster expansion study of functionalization of black phosphorene via fluorination and oxidation. *Phys. Rev. B* **93**, 165413 (2016).
13. Predith, A., Ceder, G., Wolverton, C., Persson, K. & Mueller, T. Ab initio prediction of ordered ground-state structures in ZrO<sub>2</sub>-Y<sub>2</sub>O<sub>3</sub>. *Phys. Rev. B* **77**, 144104 (2008).
14. Ising, E. Beitrag zur Theorie des Ferromagnetismus. *Z. Physik.* **31**, 253–258 (1925).
15. Casola, F. *et al.* Direct observation of impurity-induced magnetism in a spin-(1/2) antiferromagnetic Heisenberg two-leg spin ladder. *Phys. Rev. Lett.* **105**, 067203 (2010).
16. Herder, L. M., Bray, J. M. & Schneider, W. F. Comparison of cluster expansion fitting algorithms for interactions at surfaces. *Surf. Sci.* **640**, 104–111 (2015).
17. Nelson, L. J., Hart, G. L. W., Zhou, F. & Ozoliņš, V. Compressive sensing as a paradigm for building physics models. *Phys. Rev. B* **87**, 035125 (2013).
18. Ceder, G. & Van der Ven, A. Phase diagrams of lithium transition metal oxides: investigations from first principles. *Electrochim. Acta* **45**, 131–150 (1999).
19. Candès, E. J. & Wakin, M. B. An introduction to compressive sampling. *IEEE Signal Process. Mag.* **25**, 21–30 (2008).
20. Andersen, E. D., Roos, C. & Terlaky, T. On implementing a primal-dual interior-point method for conic quadratic optimization. *Math. Program.* **95**, 249–277 (2003).
21. Ceder, G. A derivation of the Ising model for the computation of phase diagrams. *Comput. Mater. Sci.* **1**, 144–150 (1993).
22. Ceder, G. *Alloy Theory and its Applications to Long-Period Superstructure Ordering in Metallic Alloys and High-Temperature Superconductors* (California Univ., 1991).
23. Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **44**, 1–12 (2004).
24. Candès, E. J., Romberg, J. & Tao, T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory* **52**, 489–509 (2006).

25. Garbulsky, G. D. & Ceder, G. Linear-programming method for obtaining effective cluster interactions in alloys from total-energy calculations: application to the fcc Pd-V system. *Phys. Rev. B Condens. Matter.* **51**, 67–72 (1995).
26. Kohan, A., Tepeš, P., Ceder, G. & Wolverton, C. Computation of alloy phase diagrams at low temperatures. *Comput. Mater. Sci.* **9**, 389–396 (1998).
27. Winston, W. L., Venkataramanan, M. & Goldberg, J. B. *Introduction to Mathematical Programming* Vol. 1 (Thomson/Brooks/Cole Duxbury; Pacific Grove, 2003).
28. Gill, P. E. & Wong, E. Methods for convex and general quadratic programming. *Math Program Comput* **7**, 71–112 (2014).
29. Kim, S.-J., Koh, K., Lustig, M. & Boyd, S. in *2007 IEEE International Conference on Image Processing III-117–III-120* (IEEE, 2007).
30. Urban, A., Seo, D.-H. & Ceder, G. Computational understanding of Li-ion batteries. *Npj Comput. Mater.* **2**, 16002 (2016).
31. Goodenough, J. B. & Park, K. S. The Li-ion rechargeable battery: a perspective. *J. Am. Chem. Soc.* **135**, 1167–1176 (2013).
32. Hong, J., Gwon, H., Jung, S.-K., Ku, K. & Kang, K. Review—lithium-excess layered cathodes for lithium rechargeable batteries. *J. Electrochem. Soc.* **162**, A2447–A2467 (2015).
33. Rozier, P. & Tarascon, J. M. Review—Li-Rich layered oxide cathodes for next-generation Li-ion batteries: chances and challenges. *J. Electrochem. Soc.* **162**, A2490–A2499 (2015).
34. Urban, A., Lee, J. & Ceder, G. The configurational space of rocksalt-type oxides for high-capacity lithium battery electrodes. *Adv. Energy Mater.* **4**, 1400478-n/a. doi:10.1002/aenm.201400478 (2014).
35. Lee, J. *et al.* Unlocking the potential of cation-disordered oxides for rechargeable lithium batteries. *Science* **343**, 519–522 (2014).
36. Mizushima, K., Jones, P., Wiseman, P. & Goodenough, J. LiCoO<sub>2</sub> (0 < x < 1): a new cathode material for batteries of high energy density. *Mater. Res. Bull.* **15**, 783–789 (1980).
37. Glazier, S. L., Li, J., Zhou, J., Bond, T. & Dahn, J. R. Characterization of disordered Li (1 + x)Ti<sub>2</sub>xFe (1–3 x)O<sub>2</sub> as positive electrode materials in Li-ion batteries using percolation theory. *Chem. Mater.* **27**, 7751–7756 (2015).
38. Lee, J. *et al.* A new class of high capacity cation-disordered oxides for rechargeable lithium batteries: Li–Ni–Ti–Mo oxides. *Energy Environ. Sci.* **8**, 3255–3265 (2015).
39. Bongers, P. *Structure and magnetic properties of several complex oxides of the transition elements.* (University of Leiden thesis, 1957).
40. Lecerf, A. in *ANNALES DE CHIMIE FRANCE*. 513–8.
41. Hoffmann, A. Crystal chemistry of lithium Ferrite. *Naturwissenschaften* **26**, 431 (1938).
42. Posnjak, E. & Barth, T. F. A new type of crystal fine-structure: lithium ferrite (Li<sub>2</sub>O·Fe<sub>2</sub>O<sub>3</sub>). *Phys. Rev.* **38**, 2234 (1931).
43. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
44. Anisimov, V. V., Zaanen, J. & Andersen, O. K. Band theory and Mott insulators: Hubbard U instead of Stoner I. *Phys. Rev. B Condens. Matter.* **44**, 943–954 (1991).
45. Jain, A. *et al.* A high-throughput infrastructure for density functional theory calculations. *Comput. Mater. Sci.* **50**, 2295–2310 (2011).
46. Hart, G. L. & Forcade, R. W. Algorithm for generating derivative structures. *Phys. Rev. B* **77**, 224115 (2008).
47. Huang, W. *et al.* Finding and proving the exact ground state of a generalized Ising model by convex optimization and MAX-SAT. *Phys. Rev. B* **94**, 134424 (2016).
48. Bishop, C. M. Pattern recognition. *Mach. Learn.* **128**, 147–152 (2006).
49. Ducastelle, F. *Order and Phase Stability in Alloys* (North-Holland, 1991).
50. Huang, W. *et al.* Constructing and proving the ground state of a generalized Ising model by the cluster tree optimization algorithm. *arXiv* **1606**, 07429 (2016).
51. Huang, W. *et al.* Data for Construction of Ground-State Preserving Sparse Lattice Models for Predictive Materials Simulations. doi:10.17605/OSF.IO/6DEHY (2017).



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017