CLEASE: A versatile and user-friendly implementation of Cluster Expansion method

Chang, Jinho; Kleiven, David; Melander, Marko; Akola, Jaakko; García Lastra, Juan Maria; Vegge, Tejs

Published in:
Npj Computational Materials

Publication date:
2018

Document Version
Early version, also known as pre-print

Link back to DTU Orbit

Citation (APA):
CLEASE: A versatile and user-friendly implementation of Cluster Expansion method

Jin Hyun Chang,1, a) David Kleiven,2 Marko Melander,3 Jaakko Akola,2, b) Juan Maria Garcia Lastra,1 and Tejs Vegge1

1) Department of Energy Conversion and Storage, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
2) Department of Physics, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway
3) Department of Chemistry, University of Jyväskylä, Jyväskylä, Finland

(Dated: 31 October 2018)

Materials exhibiting a substitutional disorder such as multicomponent alloys and mixed metal oxides/oxyfluorides are of great importance in many scientific and technological sectors. Disordered materials constitute an overwhelmingly large configurational space, which makes it practically impossible to be explored manually using first-principles calculations such as density functional theory (DFT) due to the high computational costs. Consequently, the use of methods such as cluster expansion (CE) is vital in enhancing our understanding of the disordered materials. CE dramatically reduces the computational cost by mapping the first-principles calculation results on to a much faster Hamiltonian. In this work, we present our implementation of the CE method, which is integrated as a part of the Atomic Simulation Environment (ASE) open-source package. The versatile and user-friendly code automates the complex set up and construction procedure of CE while giving the users the flexibility to tweak the settings and to import their own structures and previous calculation results. Recent advancements such as regularization techniques from machine learning are implemented in the developed code. The code allows the users to construct CE on any bulk lattice structure, which makes it useful for a wide range of applications involving complex materials. We demonstrate the capabilities of our implementation by analyzing the two example materials with varying complexities: a binary metal alloy and a disordered lithium chromium oxyfluoride.

Keywords: Cluster Expansion; Monte Carlo; disordered materials; battery material; alloys

I. INTRODUCTION

A class of materials with a substitutional disorder such as multicomponent alloys and mixed metal oxides/oxyfluorides is said to have a configurational problem. The vast configurational space of these materials makes it practically impossible to explore directly using first-principles calculations such as density functional theory (DFT). A quantitative method capable of establishing the relationship between the structure and property of materials is therefore essential. Cluster Expansion (CE)1–6 is a method that has been used successfully in the past few decades to parameterize and express the configurational dependence of physical properties. The most widely parameterized physical property is energy computed using first-principles methods, but CE can also be used to parameterize other quantities such as band gap7,8 and density of states9.

Despite its success and usefulness in predicting physical properties of crystalline materials, CE remains as a niche tool used in a small subfield within the computational materials science, primarily used by specialists. On the other hand, the research fields in which CE is becoming relevant is on the rise; one such example is the use of disordered materials for battery applications10–14. The objective of our work is to make cluster expansion more accessible for a broad range of computational scientists who do not necessarily possess expertise in cluster expansion. Our approach to achieving such a goal is to implement CE as a part of a widely used, open-source Atomic Simulation Environment (ASE) package15. Henceforth, we refer to our implementation as CLEASE, which stands for CLuster Expansion in Atomic Simulation Environment.

Having CE as a part of a widely used package with interfaces to a multitude of open-source and commercial atomic-scale simulation codes accompanies several practical benefits: (1) a large existing user base does not need to install or learn a new program as the CE module is a part of ASE and inherits its syntax and code style, and (2) all of the atomic-scale simulation codes supported by ASE are also automatically supported by the implemented module. Therefore, the implementation presented in this article appeals to a significant portion of computational materials science community as a versatile and easy-to-learn package, thereby lowering the barrier to incorporate cluster expansion as a part of their research methods to accelerate computational materials prediction and design.

The rest of the paper is organized as follows. In Results, we describe the implementation of the CLEASE code. In Discussion, two application examples with different levels of complexities, namely a binary metal alloy

a)Electronic mail: jchang@dtu.dk
b)Laboratory of Physics, Tampere University of Technology, FI-33101 Tampere, Finland
and a lithium metal oxyfluoride, are presented. The computational settings and technical details for the examples are provided in Methods. As the objective of this work is to make cluster expansion accessible to non-specialists, a brief overview of cluster expansion formalism and other important concepts are provided in Supplementary Information in order to aid the readers who are not familiar with the cluster expansion method.

II. RESULTS

In cluster expansion, a configuration (or a structure), $\sigma$, is decomposed into a sum of clusters as shown in Fig. 1. The symmetrically equivalent clusters are classified as the same cluster, and the collection of all symmetrically equivalent clusters are denoted with an $\alpha$. Each cluster type, $\alpha$, has a correlation function, $\phi_\alpha$, which represents an average value of the product of a basis function assigned to each element consisting the cluster. The target physical quantity to be described by CE, $q(\sigma)$, is then written as

$$q(\sigma) = J_0 + \sum_\alpha J_\alpha \phi_\alpha,$$  

(1)

where $J_0$ is the effective cluster interaction (ECI) of an empty cluster and $J_\alpha$ is the ECI per atom for cluster $\alpha$, which corresponds to the clusters of size one and higher. Construction of a CE model ultimately boils down to selecting clusters that are relevant for mapping configuration, $\sigma$, to the target physical quantity, $q(\sigma)$, and determining ECI values of the selected clusters.

CLEASE utilizes the existing classes and methods of ASE to perform necessary manipulations and analyses for carrying out CE. Among many adopted features, the most noteworthy are the use of

- An **Atoms** object to represent an atomic configuration ($\sigma$),

- A built-in database to store settings, atomic configurations of the training set and values of the correlation functions ($\phi_\alpha(\sigma)$), and

- A **Calculator** class to determine the physical quantity $q(\sigma)$ of a new configuration based on its correlation functions and their ECI values.

Although it may not be trivial, CE needs to keep track of vacancies, which can easily be achieved by treating a vacancy as a regular atom with its atomic symbol set to ‘X’ or atomic number set to zero. The inheritance of the existing features of ASE allows CLEASE to be fully integrated to ASE where the users can incorporate CE as a part of their research without losing the continuity with the rest of their workflow.

A simple flowchart illustrating the procedure for constructing CE using CLEASE is shown in Fig. 2. The CLEASE workflow can be divided in to three main components: definition of CE settings, generation of training structures and evaluation of CE convergence. CLEASE is written in Python programming language and takes an object-oriented approach where each component has its own class. The modular design approach not only enables easy implementation of new features but also makes the code flexible to use and intuitive to follow the CE construction and evaluation procedure shown in Fig. 2. A more detailed description of main components of the procedure is provided below.

A. Definition of Cluster Expansion Settings

The most fundamental component is to define which underlying crystal structure to use. ASE offers two functions to generate a crystal structure: **bulk** and **crystal**. The **bulk** function provides a simple way of generating common types of crystal structures by specifying the name of the crystal structure and its lattice constant value(s). The supported crystal structures are simple cubic, face-centered cubic, body-centered cubic, hexagonal close packed, diamond, zinc blende, rocksalt, cesium chloride, fluorite and wurtzite structures. For more complicated crystal structures, a **crystal** function is used to generate a crystal structure by providing its space group, lattice parameters and scaled coordinates of the unique atomic sites. The definitions of the cluster expansion settings are specified using **CEBulk** and **CECrystal** classes, which respectively calls **bulk** and **crystal** functions to generate an **Atoms** object with the user-specified crystal structure.

The maximum size of the supercell on which the DFT calculations are performed is also defined along with the definition of the underlying crystal structure. The maximum supercell size is specified using a **supercell_factor** parameter, which is an integer corresponding to the product of the expansion coefficients $i$, $j$ and $k$ of $i \times j \times k$ supercell of the unit cell generated using **bulk** or **crystal** function. For example, a **supercell_factor** of 4 includes $1 \times 1 \times 1$, $1 \times 1 \times 2$, $1 \times 1 \times 3$, $1 \times 1 \times 4$ and $1 \times 2 \times 2$ cells and all of the possible permutations of $i$, $j$ and $k$ resulting in a unique cell (i.e., only one of $i \times j \times k$ and $i \times j \times k$ cells is kept if one can be mapped on to another by rotation and reflection). The use of supercells with varying sizes and shapes enables the exploration of a larger configurational space without adding extra computational burden compared to using one fixed supercell size and shape (e.g., only using $1 \times 2 \times 2$ supercell). A set of training structures for CE are later generated iteratively from the pool of possible structures that are realizable using these supercells. To reduce the required computational resources, the structures using smaller supercells are generated (and calculated) first, followed by the larger supercells. The users also have a flexibility to select the supercell size using an optional **size** parameter, which is a $3 \times 1$ vector specifying the values of $i$, $j$
FIG. 1. A simplified illustration of the decomposition of a body-centered cubic lattice.

and $k$.

Theoretically, an infinite number of clusters can be generated for a given system. The number of clusters is limited to a finite size in practice, and CLEASE takes an approach to generate all possible clusters that are under the truncation threshold (i.e., a maximum number of atoms in clusters and maximum diameter) specified by the user. A whole or subset of the generated clusters is selected during the convergence evaluation process. By default, up to four-body clusters (i.e., empty, one-, two-, three- and four-body clusters) with their diameters smaller than the maximum diameter $16$ that can be represented by the largest supercell are generated. The users have an option to define their own threshold settings both at the beginning of the CE procedure and at a later stage of the CE iteration cycles. CLEASE also offers a feature to visualize the generated clusters in order to assist the user to develop an intuition on the generated clusters.

Within the CE formalism, there does not exist a unique set of definitions for basis functions; they are considered valid if all of the basis functions are orthogonal. Consequently, several definitions are used in practice. The two most widely used definitions are the original definitions by Sanchez et al.\textsuperscript{1} and the one later developed by van de Walle\textsuperscript{17}, which is used in the Alloy Theoretic Automated Toolkit (ATAT)\textsuperscript{17,18}. The two definitions are equally valid, and both are implemented in CLEASE.

CLEASE offers an option to ignore a set of symmetrically inequivalent atomic sites if they are always occupied by one element type for all possible configurations. The contributions of these atoms are not explicitly included in the cluster expansion and are automatically included in the constant term ($J_0$) in Eq. 1. For example, lithium metal oxides (LiMO$_2$) with first-row transition metals (M = \{Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu\}) have a rocksalt structure\textsuperscript{13,19} with an exception of LiMnO$_2$, which is orthorhombic\textsuperscript{20}. The rocksalt structure consists of two face-centered cubic sublattices. For the case of rocksalt LiMO$_2$, one sublattice is occupied by lithium and other metal atoms while the other is occupied by oxygen atoms. The configurational space of such systems can be reduced to a cation sublattice consisting only of two element types (the oxygen sublattice is ignored), which significantly reduces the complexity of the CE model. As such, an optional Boolean argument is present in CLEASE to enable/disable the reduction of configurational space by ignoring the such atoms if they exist in the system.

B. Generation of Training Structures

CLEASE uses \texttt{NewStructures} class to generate training structures, which provides three different methods
Perform the task. The first and most trivial method is to generate a set of random structures. This method serves to generate an initial pool consisting of a small number of structures. The random generation method is used in the first iteration cycle of CE construction as shown in Fig. 2. An initial cluster expansion is capable of making a first set of predictions albeit with a low accuracy. It is noted that all of the generated training structures, along with their correlation function values, are stored in a database file.

Once the initial CE is constructed, the user is given three different choices for introducing an additional set of training structures. The first and most straightforward option is to keep generating random structures. Although trivial, generating random structures is claimed to be the best strategy when compressive sensing is used to select clusters. The second method is to generate ground-state and other low-energy structures based on current cluster expansion (i.e., based on the pool of structures already calculated), which have the enthalpies of formation on or near the convex hull. The inclusion of ground-state and near-ground-state structures serves an important purpose of increasing the accuracy in predicting the correct ground states. A global minimization technique can be used to generate (near) ground-state configurations, and CLEASE uses a simulated annealing technique.

The last method of generating the training set is referred to as a “probe structure” method. The probe structure method introduces a new structure that minimizes the mean variance of the predicted physical quantity \( q(\sigma) \). The mean variance of the predicted quantity \( q \) is written as:

\[
\text{Var}[\hat{q}_i] = \frac{1}{N_{\text{config}}} \sum_{i=1}^{N_{\text{config}}} (X_i(X^TX)^{-1}X_i^T - \mu(X^TX)^{-1}\mu^T)^2
\]

(2)

where \( \epsilon^2 \) is the variance of the error in the training set, \( \Sigma \) is the covariance matrix of the correlation functions of the training set and \( \mu \) is a vector of the mean correlation functions of the structures in the training set. The probe structure is the one that reduces the value of \( \Lambda \) the most when introduced to the training set, which is found using the simulated annealing procedure.

The newly generated structures are compared with the existing structures in the training set in order to avoid introducing duplicate structures. We adopted the structure comparison algorithm developed by Lonie and Zurek to identify equivalent structures. It is desirable to have the new structure compared against the existing structures in the training set as efficiently as possible. As a first step, the structures that have different chemical composition than the newly generated structure are filtered, and the new structure is compared only with the remaining structures. Once the candidate transformations for mapping the new structure onto one structure in the database are identified using the algorithm suggested by Lonie and Zurek, we note that exactly the same transformations can be used for the remaining structures in the database. Therefore, the structure comparison algorithm implemented in ASE is optimized for the case where one structure is to be compared against many.

In addition to the aforementioned methods of generating the training structures, CLEASE also offers a built-in function to import structures to the database. The import function also has an option to specify the calculated \( q \) value, which allows users to easily import the previously calculated results.

### C. Evaluation of Cluster Expansion Convergence

An evaluation process to determine the convergence of CE includes a selection of clusters, a determination of their ECI values and an assessment of the leave-one-out cross-validation (LOOCV) score using the selected clusters and their ECI values. An entire evaluation process
is performed using an Evaluate class.

The simplest way to determine the ECI values of the generated clusters is by using ordinary least squares (OLS) to minimize residual sum of squared errors (RSS). It is highly likely that the ECI values found using OLS are overfitted. Therefore, both $\ell_1$ and $\ell_2$ regularization methods are implemented, and it is highly recommended to use a regularization methods to select clusters and evaluate their ECI values.

A default option in the Evaluate class is to include all of the clusters generated using the truncation conditions specified in CEBulk or CECrystal class, and either the entire or a subset of these clusters are selected for fitting depending on the method used. The Evaluate class provides additional options in which the users can select a subset of the generated clusters to perform any of OLS, $\ell_1$ and $\ell_2$ regularization. The first option is by manually specifying which clusters to include, while the second option is to provide a stricter truncation conditions than the ones set in the CEBulk or CECrystal class. The feature to freely select a subset of a large pool of clusters along with the use of OLS, $\ell_1$ and $\ell_2$ regularization methods allows the users to easily experiment with various settings to understand how the system behaves and to optimize the ECI values for achieving the lowest LOOVCV score.

To further assist the evaluation process, the Evaluate class contains two built-in methods that automatically determine the LOOVCV when a regularization method is used. The first method, plot_fit, determines and stores the selected clusters and their ECI values for a value of regularization parameter ($\lambda$) specified by the user. It also plots the fit of all data points in the training set to their calculated values and presents the LOOVCV score of the specified $\lambda$ value. Since the most cumbersome task in determining the convergence of CE is finding the optimal $\lambda$ value that yields the lowest LOOVCV score, another method, plot_CV, is also implemented. It takes a range and number of $\lambda$ values to evaluate as inputs and returns the best $\lambda$ value in the specified range along with its LOOVCV score. The plot_CV method also plots the LOOVCV score as a function of $\lambda$ and provides an option to store the results in a log file such that the users can add more $\lambda$ values to the list at a later stage without having to re-evaluate the same $\lambda$ values in the process.

D. Metropolis Monte Carlo and Simulated Annealing

The user can perform statistical sampling of the system on a larger simulation cell once the cluster expansion is constructed. The final selection of cluster and their ECI values can be stored and passed to other classes to conduct statistical analyses. A separate Calculator class for cluster expansion is implemented in ASE. The Clease calculator class takes a list of clusters and their ECI values as inputs, and the users can select what type of trial moves are allowed. The sampling in the canonical ensemble allows the swapping two atoms with different constraint conditions (i.e., swap any two atoms, swap any two atoms in the same basis, swap two nearest neighbors, swap two nearest neighbors in the same basis) while the semi-grand canonical ensemble allows changing the type of occupying element at a random site.

The evaluation of the physical quantity $q(\sigma)$ is performed using Eq. 1, which is a fast because the Clease calculator keeps track of the changes in the Atoms object to update the correlation functions. When the physical quantity being modeled is energy, a trial move of the standard Metropolis algorithm has an acceptance probability

$$P_{\text{acc}} = \min \left\{ 1, \exp \left( \frac{-(E_{\text{new}} - E_{\text{old}})}{k_B T} \right) \right\},$$

where $E_{\text{new}}$ and $E_{\text{old}}$ are the energy of the new and old configuration, respectively. $k_B$ is the Boltzmann constant and $T$ is temperature in Kelvin. As the Clease calculator keeps track of the change in the Atoms object after each move, updating the correlation functions is restricted to the contributions of one and two atoms for the semi-grand canonical ensemble and canonical ensemble, respectively.

III. DISCUSSION

Here, we present two example systems to illustrate the capabilities of the CLEASe code. The first example illustrates the investigation of a Au–Cu binary alloy. The second example shows the cluster expansion on a more complex Li$_2$CrO$_2$F system consisting of four types of elements and vacancy.

A. Au–Cu Alloy

The binary Au–Cu alloy system is studied at temperatures ranging from 100 K to 800 K over the entire composition range. The resulting values obtained for both $\ell_1$ and $\ell_2$ regularization are shown in Fig. 3. The ECI value of the empty cluster, which is found to be $-3.49$ eV/atom for both cases, is not included in Fig. 3 for better visibility. The LOOVCV score for the $\ell_1$- and $\ell_2$-regularized fit were 6.05 meV/atom and 6.00 meV/atom, respectively. The two regularization schemes yield essentially the same CV score, despite the fact that the $\ell_1$-regularized fit has only 8 clusters while the $\ell_2$-regularized fit has 17. The two regularization schemes give a similar qualitative trend although they have given different number of clusters and their ECI values. For instance, the magnitude of the one-body term is the largest while the signs of the two-body terms (pair interaction) are the same (e.g., the nearest-neighbor term is positive while the second and fourth nearest-neighbor terms are negative).

A qualitative information on the thermodynamic behavior of the system can be extract by inspecting the
FIG. 3. ECIs obtained via a) $\ell_1$ regularization and b) $\ell_2$ regularization. The diameters of the clusters normalized by the nearest neighbor distance are shown on the $x$-axis.

ECI values for simple binary system. Based on the fact that the energetically favorable configurations have DFT energies that are more negative than less favorable ones and that the two site variables are $+1$ or $-1$, one can infer that a positive ECI value of the pair interaction term means that a pair consisting of two different elements is energetically preferred at a low temperature. Based on the ECI values in Fig. 3, the ECIs of the nearest-neighbor and second nearest-neighbor pairs are negative and positive, respectively. This indicates that the Au–Cu system energetically favors the strong mixing of the constituting elements such that the... –Au–Cu–Au–Cu–... pattern is likely to emerge, which is in a good agreement with experimental and computational observations.

It is experimentally determined that Au–Cu alloys have three ordered phases at low temperatures: AuCu$_3$, AuCu and Au$_2$Cu. The formation enthalpy, free energy of formation and configurational entropy are obtained through the Metropolis Monte Carlo simulation and are shown in Fig. 4. As the CE is trained with fully relaxed structures (zero pressure), the formation enthalpy is determined using

$$\Delta U = U - xU_{Au} - (1 - x)U_{Cu},$$  

where $U$ is the internal energy of the configuration, $x$ is the gold concentration, $U_{Au}$ is the internal energy of pure gold and $U_{Cu}$ is the internal energy of pure copper. Similarly, the free energy of formation is obtained by subtracting the weighted average of the free energy for the pure phases. The configurational entropy is given by the difference between the internal energy and the free energy, divided by the temperature at which the Monte Carlo is sampled. The three ordered phases (AuCu$_3$, AuCu and Au$_2$Cu) are found on the convex hull of the free energy of formation in Fig. 4b. Furthermore, the entropy of the ordered phases form local minima as shown in Fig. 4c. As the temperature increases, the free energy becomes a smooth convex curve with a minimum at 50% composition, and the system is in a random phase with no short-range order.

An accurate estimate of the order/disorder transition temperature can be found by tracking the evolution of an order parameter. As the system evolves, we track the average fraction $f_{\text{diff}}$ of sites having a different element than the same site in the ground state. Furthermore, we normalize this number by the expected fraction if different sites $f_{\text{diff,rand}}$ in the random phase, and use $\eta = 1 - f_{\text{diff}}/f_{\text{diff,rand}}$ as an order parameter for detecting the phase transition as shown in Fig. 5. The order/disorder transition is predicted to occur near 450 K, which is in a close agreement with the previous CE calculation.

B. Lithium Chromium Oxyfluoride

One of the recent focus areas of lithium-ion battery research is the development of high-capacity cathode materials. Lithium metal oxyfluorides (Li$_2$MO$_2$F, M = {V, Cr, Mn, Ti, Ni, ...}) is a family of materials that is at the forefront of the current research. The challenges for studying Li$_2$MO$_2$F is in the vast size of the configurational space, which exhibit not only the cation disorder commonly found in lithium metal oxides but also anion disorder which is also present due to the mixed O/F composition. The fact that the underlying crystal structure of Li$_2$MO$_2$F can vary at different lithiation levels adds the complexity to investigate their properties. It is, however, known that the most predominant crystal structure is of disordered rocksalt type, particularly at high-lithiation levels. We therefore show an example CE study of Li$_2$CrO$_2$F in a rocksalt configuration. It is important to point out that the presented test model does not account for the possible occupancy of the tetrahedral sites by Li, which is known to occur upon delithiation. More specifically, each tetrahedral site are adjacent to four octahedral sites, and Li from one of the four octahedral sites migrate to the tetrahedral site if the remaining three are vacant (referred to as...
FIG. 4. Thermodynamic quantities for the Au–Cu computed for temperatures ranging from 100 K to 800 K over the entire composition range. a) Formation enthalpy. b) Free energy of formation. c) Entropy. The insets show the three ordered phases: AuCu$_3$, AuCu and Au$_3$Cu.

Although the omission of the tetrahedral site occupancy makes the model incomplete, it provides useful insight on how complicated systems with a vast configurational space can be studied using CLEASe, which is the aim of this example.

The Monte Carlo annealing study reveals that Li$_2$CrO$_2$F (i.e., fully lithiated compound) takes a layered structure at room temperature (293 K) as shown in Fig. 6a. The layer structure shows a ...–Li–F–...

Li–O–Cr–O... pattern, which is similar to a ...–Li–O–M–O... layered pattern observed in lithium metal oxides$^{13,35,36}$. The layered structure is lost immediately upon delithiation, which leads to disordered structures as shown in Fig. 6b and c. The emergence of disordered structures agrees well with the previous experimental observations$^{31,33}$, and it is important to model the disordered atomic arrangement it has a direct link to the Li transport mechanism (e.g., a presence of zero-transition-metal pathways$^{13,19,34}$).

An interesting feature of Li$_x$CrO$_2$F is the phase-separation of Li and vacancies, which leads to a segregation of vacancies that can be seen even at high lithiation levels such as $x = 1.8$ and $x = 1.6$ as shown in Fig. 6b and c, respectively. The vacancy segregation is an important feature to understand since it affects the transport properties of Li and promotes leakage of O and F anions. In fact, the oxygen loss are known to occur in lithium metal oxides$^{37–39}$ and also in lithium metal oxyfluorides$^{40}$ although its extent is suppressed. The oxygen loss leads to poor capacity retention as it is an irreversible process. The phase-separation of Li and vacancy indicates that oxygen loss is likely to take place for Li$_2$CrO$_2$F, which aligns with its poor capacity retention of only 64% after 60 cycles$^{31}$ and irreversible change of the local structure$^{31}$.

It is pointed out that while the presented model correctly predicts the disordered structure and poor cyclability of Li$_2$CrO$_2$F, the vacancy segregation shown in Fig. 6b and c needs to be interpreted with care. The aforementioned omission of Li occupancy in interstitial tetrahedral sites is likely to occur, especially due to the vacancy segregation. The missing migration of Li to tetrahedral site may suppress the further clustering of vacancies. With the missing tetrahedral site occupation mechanism, current model predicts the vacancy cluster to keep growing as more Li is taken out, which means...
FIG. 6. Final structure of Li_{1−x}CrO\_2F after the Monte Carlo run at 293 K where x is a) 2.0, b) 1.8 and c) 1.6. The Li atoms are shown in green, the Cr atoms are shown in blue, the oxygen atoms are shown in red and the F atoms are shown in white.

that its voltage remains relatively constant as Li_{1−x}CrO\_2F is phase-separated into CrO\_2F and Li_2CrO\_2F. The experimentally observed voltage curve\(^{31,33}\) is more complex than what the test model can capture.

IV. METHODS

A. Density Functional Theory Calculations

All of the calculations are performed with the Vienna Ab initio Simulation Package (VASP)\(^{41}\) using the projector augmented-wave (PAW) method\(^{42}\). The generalized gradient approximation as parametrized by Perdew, Burke and Ernzerhof\(^{43}\) is used as the exchange-correlation functional. It is important to have a consistent and accurate dataset (i.e., DFT calculations with high energy cutoff and \(k\)-point mesh density) in order to minimize the numerical noise introduced to the CE training. The plane-wave cutoff of 500 eV is used, and both the cell and atomic positions are fully relaxed such that all the forces are smaller than 0.02 eV/Å. A rotationally invariant Hubbard \(U\) correction\(^{44,45}\) is applied to the \(d\) orbital of Cr with the \(U\) value of 3.7 eV. The calculations are performed with supercells containing up to 18 and 54 atoms for Au–Cu alloy and Li_{2}CrO\_2F systems, respectively. Integrations over the Brillouin zone were carried out using the Monkhorst-Pack scheme\(^{46}\) with a grid with a maximal interval of 0.04 Å\(^{-1}\).

B. Cluster Expansion Model

The CE model for Au–Cu alloy and Li_{1−x}CrO\_2F are trained using 34 and 390 DFT calculations, respectively. CE model is trained for the entire composition range of Au–Cu alloy (from pure Au to pure Cu) and rock-salt Li_{1−x}CrO\_2F with \(x\) ranges from 0 to 2. Up to three-body clusters are used in both cases, where the maximum diameter is set to 5.5 Å for Au–Cu alloy and 4.5 Å for Li_{1−x}CrO\_2F. \(\ell_1\) and \(\ell_2\) regularization schemes with the regularization parameter ranging from \(10^{-7}\) to \(10^{2}\) are assessed to find the optimal setting that leads to the lowest LOOCV score. For Au–Cu alloy, \(\ell_1\) and \(\ell_2\) regularization schemes yield nearly identical LOOCV score of 6.05 meV/atom and 6.00 meV/atom, respectively. On the other hand, \(\ell_1\) regularization performed better on Li_{1−x}CrO\_2F with the LOOCV score of 26.5 meV/atom\(^{47}\).

C. Metropolis Monte Carlo Simulations

For Au–Cu alloy, Metropolis Monte Carlo simulations are carried out using a supercell consisting of 1,000 atoms. At each temperature (i.e., 100 K to 800 K with a 100 K increment), the system is equilibrated with 10,000 MC steps, and an average energy is collected through additional 100,000 moves. A \(9 \times 9 \times 9\) cell consisting of 1,458 atoms is used for Li_{1−x}CrO\_2F. The temperature is gradually lowered from 10,000 K, and the structures are equilibrated at each temperature via 14,580 steps (10 times the number of atoms in the cell) to ensure that the system is equilibrated before sampling. The system is
then sampled for more than one million steps. Such procedures ensure that the systems are sufficiently sampled at each composition at different temperatures.

V. ACKNOWLEDGEMENT

Authors thank valuable discussions with Dr. Juhani Teeriniemi and Prof. Kari Laasonen.

VI. FUNDING

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 711792 (FETOPEN project LiRichFCC).


16. The maximum cluster diameter is defined to be strictly smaller than half the smallest distance between periodic images of all the constituting atoms in the cell.


Chemistry 55, 3789–3796 (2016).


36. A. Van der Ven and G. Ceder, “Ordering in Li\(_x\)(Ni\(_{0.5}\)Mn\(_{0.5}\))O\(_2\) and its relation to charge capacity and electrochemical behavior in rechargeable lithium batteries,” Electrochemistry Communications 6, 1045–1050 (2004).


47. Although the LOOCV seems larger compared to the case Au–Cu, it should be taken into account that the cohesive energy of metallic alloys is in general much smaller than those of oxides/oxyfluorides.
### I. CLUSTER EXPANSION FORMALISM

The core concept of the cluster expansion is to express the scalar physical quantity of a material, \( q(\sigma) \), to its configuration, \( \sigma \), where a crystalline system is represented with a fixed underlying grid of atomic sites. In such a representation, any configuration with the same underlying topology can be completely specified by the atomic occupation of each atomic site. For the case of a crystalline material with \( N \) atomic sites, any configuration can be specified by an \( N \)-dimensional vector \( \sigma = \{s_1, s_2, \ldots, s_N\} \), where \( s_i \) is a site variable that specifies which type of atom occupies the atomic site \( i \) (also referred to as an occupation variable\(^{S1–S3} \) or pseudospin\(^{S4–S8} \)). It is noted that the terms configuration and structure are often used interchangeably.

For the case of multinary systems consisting of \( M \) different atomic species, \( s_i \) takes one of \( M \) distinct values. The original formulation of Sanchez et al.\(^{S9} \) specifies the \( s_i \) to take any values from \( \pm m, \pm (m-1), \ldots, \pm 1 \) for \( M = 2m^{S10} \). Other choices of \( s_i \) are also commonly used such as values ranging from 0 to \( M-1 \) by van de Walle\(^{S3} \) and from 1 to \( M \) by Mueller and Ceder\(^{S11} \). Based on the original formalism by Sanchez et al., single-site basis functions are determined through an orthogonality condition

\[
\frac{1}{M} \sum_{s_i=-m}^{m} \Theta_n(s_i) \Theta_{n'}(s_i) = \delta_{nn'},
\]

where \( \Theta_n(s_i) \) is the \( n \)th single-site basis function (e.g., Chebyshev polynomials) for \( i \)th site and \( \delta_{nn'} \) is a Kronecker delta.

The configuration is decomposed into a sum of clusters. Each cluster has a set of associated cluster functions, which are defined as

\[
\Phi_n(s) = \prod_i \Theta_{n_i}(s_i),
\]

where \( n \) and \( s \) are vectors specifying the order of the single-site basis function and the site variables in the cluster, respectively. \( n_i \) and \( s_i \) specify the \( i \)th element of the respective vectors. The use of orthogonal basis functions guarantees that the cluster functions defined in Eq. S2 are also orthogonal. The symmetrically equivalent clusters are classified as the same cluster, and the collection of all symmetrically equivalent clusters are denoted with an \( \alpha \).

The average value of the cluster functions in cluster \( \alpha \) is referred to as a correlation function, \( \phi_{\alpha} \). The physical quantity, \( q(\sigma) \), normalized with the number of atomic sites \( N \) is then expressed as

\[
q(\sigma) = \sum_{\alpha} m_{\alpha} J_{\alpha} \phi_{\alpha},
\]

where \( m_{\alpha} \) is the multiplicity factor indicating the number of cluster \( \alpha \) per atom and \( J_{\alpha} \) is the effective cluster interaction (ECI) per occurrence, which needs to be determined. It is noted that the cluster \( \alpha \) includes the cluster of size zero, which have \( m_{\alpha} \phi_{\alpha} = 1 \). Alternatively, Eq. S3 can be written in a more explicitly form,

\[
q(\sigma) = J_0 + \sum_{\alpha} m_{\alpha} J_{\alpha} \phi_{\alpha},
\]

where \( J_0 \) is the ECI of an empty cluster while \( \alpha \) in this case corresponds to the clusters of size one and higher. It is often more practical and convenient to express the ECI per atom rather than per occurrence\(^{S12} \), in which case \( m_{\alpha} \) and \( J_{\alpha} \) are combined into one term, \( \bar{J}_{\alpha} = m_{\alpha} J_{\alpha} \) and Eq. S3 becomes

\[
q(\sigma) = \sum_{\alpha} \bar{J}_{\alpha} \phi_{\alpha}.
\]

CLEASE uses the ECI per atom (\( \bar{J}_{\alpha} \)), but interested users can determine the value of \( J_{\alpha} \) based on the values of \( m_{\alpha} \) and \( \bar{J}_{\alpha} \).

Theoretically, there is an infinite number of terms in Eq. S5, and the resulting expression can represent any scalar function \( q(\sigma) \) given that appropriate ECI values are found. In practice, sufficient accuracy is often reached with clusters with small number of atoms (e.g., one-, two- and three-body clusters) that are relatively compact in size (e.g., 5 to 7 Å in diameter).
II. CLUSTER SELECTION & DETERMINATION OF ECI VALUES

A crucial element of CE approach is to select relevant clusters from a theoretically infinite number of possible clusters. Many multicomponent systems yield thousands of clusters even when the expansion is limited to relatively compact size and small number of atoms, and they are vastly truncated since only a small fraction of them is needed to achieve the required accuracy. Determining the optimal set of clusters that minimizes the number of clusters without losing its predictive power has been a topic of keen interest in the past decade, and the cluster selection based on genetic algorithms was considered to be the most robust method.

More recently, the use of compressive sensing was proposed to efficiently select the clusters and determine their ECIs in one shot. The compressive sensing is based on $\ell_1$ norm (a special case of $\ell_p$ norm where $p=1$), which is defined as

$$||x||_p = \left(\sum_i |x_i|^p\right)^{1/p},$$

(S6)

where $x$ is a vector quantity. It is noted that cluster expansion defined in Eq. S5 is in the same form as a linear regression model in statistics and machine learning. Therefore, one can treat CE as a linear regression problem and apply regularization techniques based not only on $\ell_1$ norm but also on any other $p$ values, although $\ell_1$ and $\ell_2$ norms are most commonly used.

The use of regularization techniques for CE can be illustrated by expressing Eq. S5 in a matrix form,

$$q = X\omega.$$  

(S7)

$X$ is a matrix containing the correlation functions of the training data where each element in row $i$ and column $\alpha$ is defined as

$$X_{i\alpha} = \phi_\alpha(\sigma_i).$$

(S8)

$q$ is a column vector in which the $i$th element is the physical quantity $q$ of the configuration $\sigma_i$ and $\omega$ is a column vector in which $\alpha$th element is $\tilde{J}_\alpha$.

The simplest way of determining $\omega$ is by using ordinary least squares (OLS) method, which minimizes the residual sum of squared errors (RSS). RSS is defined as

$$\text{RSS} = ||X\omega - q||_2^2,$$

(S9)

and the minimization of the RSS has a unique solution $\tilde{\omega}$ where

$$\tilde{\omega} = \arg\min_\omega ||X\omega - q||_2^2 = (X^TX)^{-1}X^Tq.$$  

(S10)

The OLS has two major drawbacks. The first is the requirement on which the number of configurations in the training set needs to be larger than the number of clusters being considered. The matrix $X^TX$ becomes singular in such a case, and the limitations imposed by the first requirement become more severe for systems consisting of many element types since even strict expansion conditions (i.e., small number of atoms per cluster and compact size) can lead to a large number of clusters. The second drawback is the susceptibility to possible overfitting, which refers to the conditions in which the ECI values are over-tuned to accurately represent $q(\sigma)$ of the training set at a cost of losing its predictive power for the new configurations that are not included in the training set. The overfitting also makes the model prone to noise present in the training data because the model attempts to meticulously fit the model to the training data including the noise therein.

Regularization is an efficient technique to address the aforementioned drawbacks of OLS by adding a regularization term to Eq. S10. The most common regularization scheme are $\ell_1$ and $\ell_2$ regularization, which respectively uses $\ell_1$ and $\ell_2$ norm as a regularization term. For $\ell_1$ regularization, the solution becomes

$$\tilde{\omega} = \arg\min_\omega ||X\omega - q||_2^2 + \lambda||\omega||_1,$$

(S11)

where $\lambda$ is a regularization parameter that controls the weight given to the regularization term. The main benefit of $\ell_1$ regularization is its promotion of sparsity. In context of CE, the sparsity means a selection of a fewer number of clusters, or many clusters with their ECI values set to zero. It is noted that there is no unique analytical solution
for Eq. S11, and it needs to be solved iteratively. Unlike $\ell_1$ regularization, $\ell_2$ regularization has a unique analytical solution which is expressed as

$$\hat{\omega} = \arg \min_{\omega} \|X\omega - q\|_2^2 + \|\omega\|_2^2 = (X^TX + \lambda I)^{-1}X^Tq.$$  \hfill (S12)

However, $\ell_2$ regularization does not promote sparsity, and the resulting solution is likely to contain more clusters than necessary.

Regardless of the fitting technique used, the predictive power of the expansion needs to be assessed to determine its accuracy and reliability. Cross-validation (CV) is a technique used for assessing the prediction accuracy of the model. A leave-one-out (LOO) scheme is most commonly used in CE community, and the LOOCV score is defined as

$$\text{LOOCV} = \frac{1}{N_{\text{config}}} \sum_{i=1}^{N_{\text{config}}} (\hat{q}_i - q_i)^2 \frac{1}{2},$$  \hfill (S13)

where $N_{\text{config}}$ is the number of configurations in the training set, $\hat{q}_i$ is the physical quantity of a structure $i$ predicted by CE using $N_{\text{config}} - 1$ structures without a structure $i$ and $q_i$ is the calculated physical quantity of structure $i$. While OLS has only one (likely overfitted) solution, $\ell_1$ and $\ell_2$ regularization schemes have a solution for each $\lambda$ value. The solution — a selection of clusters and their ECI values — that yields the lowest LOOCV score is chosen.

III. THERMODYNAMICS IN LATTICE MODELS

The true benefit of CE is in its ability to predict the expanded scalar quantity $q(\sigma)$ based on trained data. An accurate prediction can be made if the CV score of the expanded $q(\sigma)$ is sufficiently low, and the prediction speed is very fast on modern computer architecture since it only involves executions of only a small number of simple numerical calculations specified in Eq. S5. Such a speed boost allows one to conduct types of analyses that require substantial statistical sampling.

In contrast to zero temperature studies where the system occupies the state with lowest energy, configurations with the lowest free energy are occupied at finite temperature. The free energy $G$ is given by

$$G = -\ln \frac{Z}{\beta},$$  \hfill (S14)

where $\beta = 1/k_BT$ and $Z$ is the partition function. $k_B$ is the Boltzmann constant and $T$ is temperature in Kelvin. In order to calculate the free energy, we utilize the exact differentials

$$d(\beta G) = -\frac{\partial \ln Z}{\partial \beta} d\beta = Ud\beta$$  \hfill (S15)

where $U$ is the average internal energy. The free energy can be obtained by a thermodynamic integration from a reference temperature $\beta_{\text{ref}}$ where $G$ is known, which can be written as

$$\beta G = (\beta G)_{\text{ref}} + \int_{\beta_{\text{ref}}}^{\beta} d\beta' U(\beta').$$  \hfill (S16)

Important information of the materials under study such as the stability of ordered/disordered phases can be determined by comparing the free energy of the material at a given composition with respect to the free energies in the pure phases of its constituents.

---

2. Y. S. Meng and M. E. Arroyo-de Dompablo, “First principles computational materials design for energy storage materials in lithium ion batteries,” Energy & Environmental Science 2, 589 (2009).
For the case where there is an odd number of element types, an additional value of 0 should be included in the possible values of $s_i$, and the relation between $M$ and $m$ becomes $M = 2m - 1$. 


DFT energies are obtained for fully relaxed structures without any external forces or pressure. Thus, the resulting thermodynamic quantities are effectively obtained in the NPT ensemble (fixed number of particles, fixed pressure and fixed temperature). However, the energy predicted by CE is only valid for the volume leading to the minimum energy of a particular atomic arrangement, and the volume fluctuations are neglected.
