Downloaded from orbit.dtu.dk on: May 02, 2024

DTU Library

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

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

Published in:
Journal of Physics Condensed Matter

Link to article, DOI:
10.1088/1361-648X/ab1bbc

Publication date:
2019

Document Version
Peer reviewed version

Link back to DTU Orbit

Citation (APA):
Chang, J., Kleiven, D., Melander, M., Akola, J., García Lastra, J. M., & Vegge, T. (2019). CLEASE: A versatile and user-friendly implementation of Cluster Expansion method. *Journal of Physics Condensed Matter*, *31*(32), Article 325901. https://doi.org/10.1088/1361-648X/ab1bbc

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Journal of Physics: Condensed Matter

ACCEPTED MANUSCRIPT • OPEN ACCESS

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

To cite this article before publication: Jin Hyun Chang *et al* 2019 *J. Phys.: Condens. Matter* in press https://doi.org/10.1088/1361-648X/ab1bbc

Manuscript version: Accepted Manuscript

Accepted Manuscript is "the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an 'Accepted Manuscript' watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors"

This Accepted Manuscript is **© 2019 IOP Publishing Ltd**.

As the Version of Record of this article is going to be / has been published on a gold open access basis under a CC BY 3.0 licence, this Accepted Manuscript is available for reuse under a CC BY 3.0 licence immediately.

Everyone is permitted to use all or part of the original content in this article, provided that they adhere to all the terms of the licence https://creativecommons.org/licences/by/3.0

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions may be required. All third party content is fully copyright protected and is not published on a gold open access basis under a CC BY licence, unless that is specifically stated in the figure caption in the Version of Record.

View the article online for updates and enhancements.

This content was downloaded from IP address 192.38.90.17 on 10/05/2019 at 12:52

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

Jin Hyun Chang[1], David Kleiven[2], Marko Melander[3], Jaakko Akola[2,4], Juan Maria Garcia Lastra[1] and Tejs Vegge[1]

[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

[4] Computational Physics Laboratory, Tampere University, P.O. Box 692, FI-33014 Tampere, Finland

E-mail: `jchang@dtu.dk`

Abstract. 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 Hamiltonian which is much faster to evaluate. 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.

Submitted to: *J. Phys.: Condens. Matter*

Accepted Manuscript

1. Introduction

Computational modeling of materials with a substitutional disorder such as multicomponent alloys and mixed metal oxides is said to have a *configurational* problem [1, 2, 3, 4, 5]. 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, 5, 6, 7, 8, 9] 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 gap [10, 11] and density of states [12].

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 applications [13, 14, 15, 16, 17]. 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) package [18]. 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. In addition, CLEASE utilizes the database management feature implemented in ASE, which provides an efficient way to store, maintain and share both DFT and CE results. 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. A brief overview of cluster expansion formalism and other important concepts are provided in section 2 in order to aid the readers who are not familiar with the cluster expansion method. The implementation of CLEASE is described in section 3. Section 4 contains two application examples with different levels of complexities, namely a binary metal alloy and a lithium metal oxyfluoride. The computational settings and technical details for the examples are provided in section 5.

Accepted Manuscript

2. Theory

2.1. Cluster Expansion Formalism

The core concept of the cluster expansion is to express the scalar physical quantity of a material, $q(\boldsymbol{\sigma})$, to its configuration, $\boldsymbol{\sigma}$, 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 $\boldsymbol{\sigma} = \{s_1, s_2, \ldots, s_N\}$, 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 [19, 4, 20] or pseudospin [1, 10, 21, 22, 23]). 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. [6] specifies the s_i to take any values from $\pm m$, $\pm(m-1)$, ..., ± 1 for $M = 2m$ (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 $M = 2m - 1$). Other choices of s_i are also commonly used such as values ranging from 0 to $M - 1$ by van de Walle [20] and from 1 to M by Mueller and Ceder [24]. Based on the original formalism by Sanchez et al., single-site basis functions are determined through an orthogonality condition

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

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

The configuration is decomposed into a sum of clusters as shown in figure 1. Each cluster has a set of associated cluster functions, which are defined as

$$\Phi_{\boldsymbol{n}}(\boldsymbol{s}) = \prod_i \Theta_{n_i}(s_i), \tag{2}$$

where $\boldsymbol{n}$ and $\boldsymbol{s}$ 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 ith element of the respective vectors. The use of orthogonal basis functions guarantees that the cluster functions defined in (2) 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 α.

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

$$q(\boldsymbol{\sigma}) = \sum_\alpha m_\alpha J_\alpha \phi_\alpha, \tag{3}$$

Accepted Manuscript

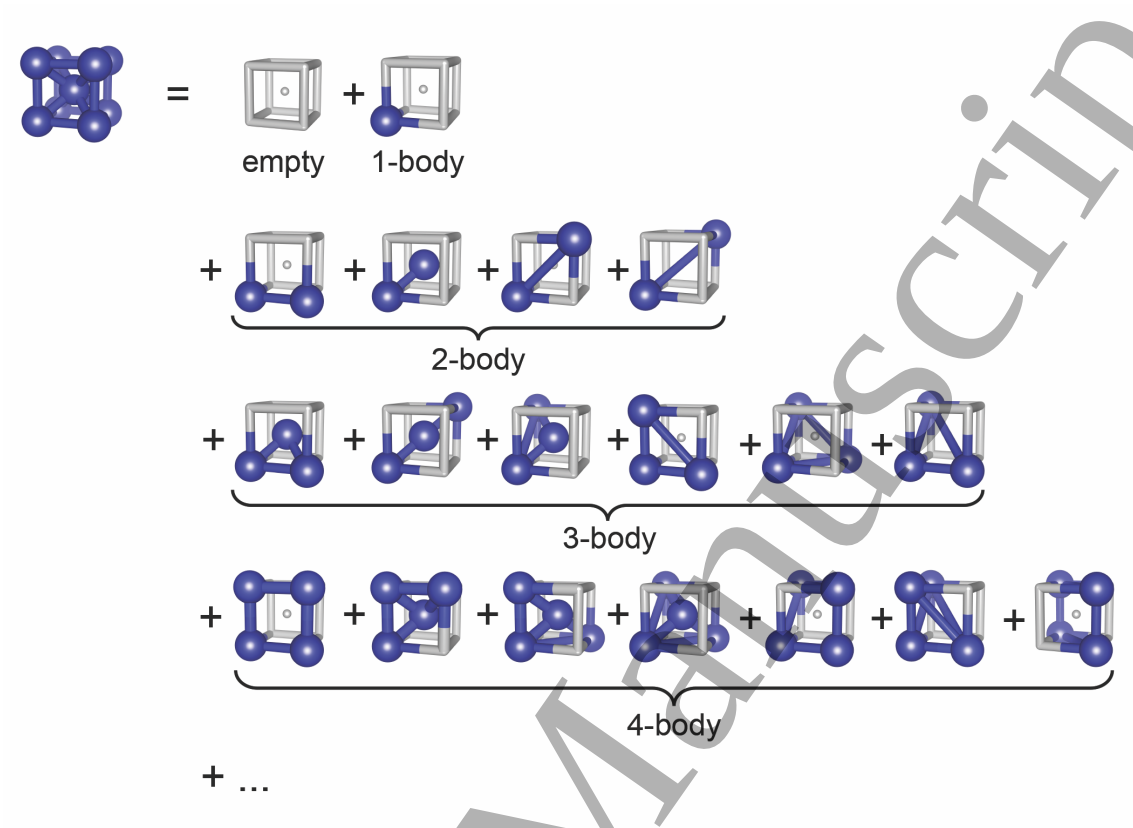

Figure 1. A simplified illustration of the decomposition of a body-centered cubic lattice.

where m_α is the multiplicity factor indicating the number of cluster α per atom and J_α is the effective cluster interaction (ECI) per occurrence, which needs to be determined. It is noted that the cluster α includes the cluster of size zero, which have $m_\alpha \phi_\alpha = 1$. Alternatively, (3) can be written in a more explicitly form,

$$q(\boldsymbol{\sigma}) = J_0 + \sum_\alpha m_\alpha J_\alpha \phi_\alpha, \tag{4}$$

where J_0 is the ECI of an empty cluster while α 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 [5], in which case m_α and J_α are combined into one term, $\widetilde{J}_\alpha = m_\alpha J_\alpha$ and (3) becomes

$$q(\boldsymbol{\sigma}) = \sum_\alpha \widetilde{J}_\alpha \phi_\alpha. \tag{5}$$

CLEASE uses the ECI per atom ($\widetilde{J}_\alpha$), but interested users can determine the value of J_α based on the values of m_α and $\widetilde{J}_\alpha$.

Theoretically, there is an infinite number of terms in (5) for an infinite crystal, and the resulting expression can represent any scalar function $q(\boldsymbol{\sigma})$ given that appropriate ECI values are found. In practice, sufficient accuracy is often reached with clusters

Accepted Manuscript

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).

2.2. 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 [19, 25, 26, 27, 3], and the cluster selection based on genetic algorithms [25, 26, 3] was considered to be the most robust method.

More recently, the use of compressive sensing [22] was proposed to efficiently select the clusters and determine their ECIs in one shot. The compressive sensing is based on ℓ_1 norm (a special case of ℓ_p norm where $p = 1$), which is defined as

$$||\mathbf{x}||_p = \left(\sum_i |x_i|^p \right)^{1/p}, \tag{6}$$

where $\mathbf{x}$ is a vector quantity. It is noted that cluster expansion defined in (5) 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 ℓ_1 norm but also on any other p values, although ℓ_1 and ℓ_2 norms are most commonly used.

The use of regularization techniques for CE can be illustrated by expressing (5) in a matrix form,

$$\mathbf{q} = \mathbf{X}\boldsymbol{\omega}. \tag{7}$$

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

$$\mathbf{X}_{i\alpha} = \phi_\alpha(\boldsymbol{\sigma}_i). \tag{8}$$

$\mathbf{q}$ is a column vector in which the ith element is the physical quantity q of the configuration $\boldsymbol{\sigma}_i$ and $\boldsymbol{\omega}$ is a column vector in which αth element is $\widetilde{J}_\alpha$.

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

$$\text{RSS} = ||\mathbf{X}\boldsymbol{\omega} - \mathbf{q}||_2^2, \tag{9}$$

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

$$\begin{aligned} \hat{\boldsymbol{\omega}} &= \arg\min_{\boldsymbol{\omega}} ||\mathbf{X}\boldsymbol{\omega} - \mathbf{q}||_2^2 \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{q}. \end{aligned} \tag{10}$$

Accepted Manuscript

The OLS has two major drawbacks [22]. 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 $\mathbf{X}^T\mathbf{X}$ 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 $q(\boldsymbol{\sigma})$ 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 (10). The most common regularization scheme are ℓ_1 and ℓ_2 regularization, which respectively uses ℓ_1 and ℓ_2 norm as a regularization term. For ℓ_1 regularization, the solution becomes

$$\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} ||\mathbf{X}\boldsymbol{\omega} - \mathbf{q}||_2^2 + \lambda||\omega||_1, \tag{11}$$

where λ is a regularization parameter that controls the weight given to the regularization term. The main benefit of ℓ_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 (11), and it needs to be solved iteratively. Unlike ℓ_1 regularization, ℓ_2 regularization has a unique analytical solution which is expressed as

$$\begin{aligned}\hat{\boldsymbol{\omega}} &= \arg\min_{\boldsymbol{\omega}} ||\mathbf{X}\boldsymbol{\omega} - \mathbf{q}||_2^2 + ||\boldsymbol{\omega}||_2^2 \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{q}.\end{aligned} \tag{12}$$

However, ℓ_2 regularization does not promote sparsity, and the resulting solution is likely to contain more clusters than necessary. It is noted that Bayesian compressive sensing [21] scheme is introduced for cluster expansion, which effectively eliminates the parameter λ in ℓ_1 and ℓ_2 regularization schemes while promoting sparsity.

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

$$\text{LOOCV} = \left(\frac{1}{N_{\text{config}}}\sum_{i=1}^{N_{\text{config}}}(\hat{q}_i - q_i)^2\right)^{1/2}, \tag{13}$$

where N_{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

Accepted Manuscript

i and q_i is the calculated physical quantity of structure i. While OLS has only one (likely overfitted) solution, ℓ_1 and ℓ_2 regularization schemes have a solution for each λ value. The solution — a selection of clusters and their ECI values — that yields the lowest LOOCV score is chosen. Although LOO is the most common cross validation scheme in cluster expansion community, k-fold CV is one of the most common schemes used in machine learning community. In a k-fold CV scheme, the pool of configurations are randomly partitioned into k parts of equal size. The structures in $k-1$ parts are used as training data while the remaining one part is used as a validation set, and the cross validation is repeated k times.

2.3. Thermodynamics in Lattice Models

The true benefit of CE is in its ability to predict the expanded scalar quantity $q(\boldsymbol{\sigma})$ based on trained data. An accurate prediction can be made if the CV score of the expanded $q(\boldsymbol{\sigma})$ 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 (5). 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, an ensemble of configurations with the lowest free energy are occupied at finite temperature. The free energy G is given by [28]

$$G = -\frac{\ln Z}{\beta}, \tag{14}$$

where $\beta = 1/k_B T$ and Z is the partition function. k_B is the Boltzmann constant and T is temperature in Kelvin. It is noted that the 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. The free energy can be calculated by utilizing the exact differential

$$\begin{aligned} \mathrm{d}(\beta G) &= -\frac{\partial \ln Z}{\partial \beta}\mathrm{d}\beta \\ &= U\mathrm{d}\beta \end{aligned} \tag{15}$$

where U is the average internal energy. The free energy can be obtained by a thermodynamic integration from a reference temperature β_{ref} where G is known, which can be written as [29]

$$\beta G = (\beta G)_{\mathrm{ref}} + \int_{\beta_{\mathrm{ref}}}^{\beta} \mathrm{d}\beta' U(\beta'). \tag{16}$$

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

Accepted Manuscript

material at a given composition with respect to the free energies in the pure phases of its constituents.

3. Implementation

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 ($\boldsymbol{\sigma}$),
- a built-in database to efficiently store, maintain and share settings, atomic configurations of the training set, values of the correlation functions ($\phi_\alpha(\boldsymbol{\sigma})$) and DFT energies,
- Python programming language and modular design to remove the strict input file/format requirements and to enable easy implementation of new features, and
- a `Calculator` class to determine the physical quantity $q(\boldsymbol{\sigma})$ of a new configuration based on its correlation functions and their ECI values.

It is noted that the evaluation of correlation functions of a new configuration and the determination of physical quantity, $q(\boldsymbol{\sigma})$, based on ECI values can be a slow process using Python programming language. It is especially true for carrying out Monte Carlo simulations after the CE model training is complete. CLEASE includes an optional external module written in C++ programming language that can be installed to accelerate the critical and repetitive calculations, but the usage of the code remains unchanged even when the external module is installed (i.e., CLEASE automatically determines if the C++ add-on is installed, and uses the C++ version if it is present).

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. The existing users of ASE do not have to install or learn a new CE program nor select a particular DFT package that a CE code supports. In addition to the benefits of integrating CE as a part of ASE, highlights of the features that makes CLEASE versatile and user-friendly include:

- a multicomponent cluster expansion that goes beyond binary systems,
- a support for several types of single-site basis functions (e.g., basis functions by Sanchez et al. [6], Van de Walle [20] and Zhang and Sluiter [5]) for a comparison and compatibility with other CE codes,
- many methods for selecting clusters and determining ECI values such as OLS, ℓ_1 and ℓ_2 regularization schemes, Bayesian compressive sensing and genetic algorithm, and
- both leave-one-out and k-fold cross validation schemes.

A simple flowchart illustrating the procedure for constructing CE using CLEASE is shown in figure 2. The CLEASE workflow can be divided in to three main components:

Accepted Manuscript

definition of CE settings, generation of training structures and evaluation of CE convergence. CLEASE 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 figure 2. A more detailed description of main components of the procedure is provided below.

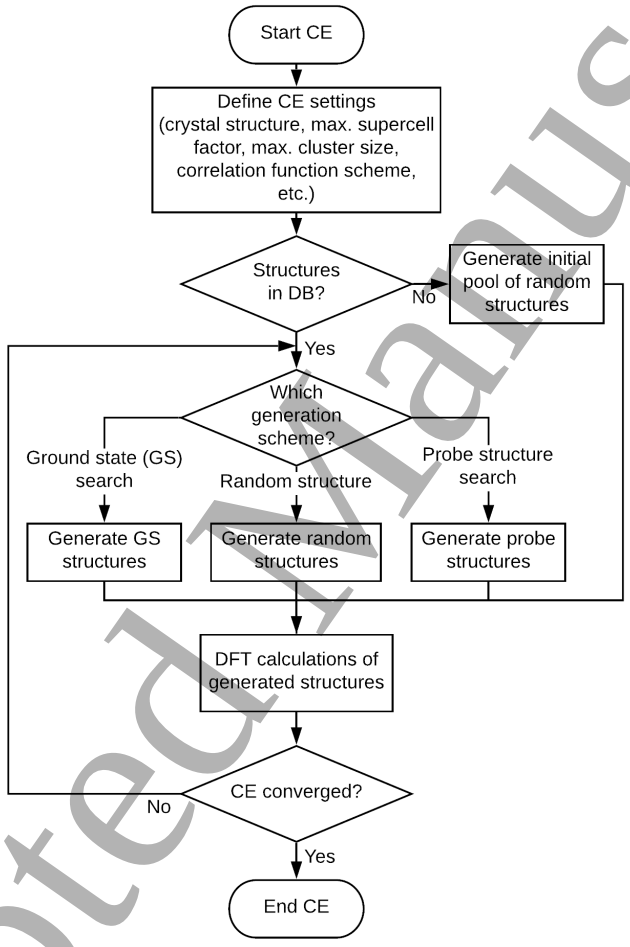

Figure 2. A flowchart of constructing and evaluating CE using CLEASE.

3.1. 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 crystal structures supported by the `bulk` function 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

Accepted Manuscript

`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 (of the primitive cell) 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 absolute values of expansion coefficients (integer weights of a general linear combinations of the unit cell vectors). In other words, if a unit cell has three vectors $\vec{a}$, $\vec{b}$ and $\vec{c}$, the configurations in the training set on which the DFT calculations are performed have cell vectors $\vec{a'}$, $\vec{b'}$ and $\vec{c'}$, which are defined as

$$\begin{aligned} \vec{a'} &= i_1\vec{a} + j_1\vec{b} + k_1\vec{c} \\ \vec{b'} &= i_2\vec{a} + j_2\vec{b} + k_2\vec{c} \\ \vec{c'} &= i_3\vec{a} + j_3\vec{b} + k_3\vec{c} \end{aligned} \tag{17}$$

with integer coefficients i_x, j_x and k_x, where $x \in \{1, 2, 3\}$. The `supercell_factor` is then defined as

$$\texttt{supercell_factor} \geq \prod_{x=1}^{3}\prod_{y=1}^{3}\prod_{z=1}^{3} |i_x|\,|j_y|\,|k_z|\,, \tag{18}$$

and all of the cells used in the training set should have coefficients satisfying the condition in (18). Only unique cell shapes are included in the pool by omitting the cells that can be mapped on to the existing cells in the pool 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. 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 3×3 matrix (or a nested list in Python) specifying the values of the integer coefficients in (17).

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 up to 5 Å are generated. The users have an option to define their own threshold settings both at the beginning of the CE procedure and

Accepted Manuscript

at a later stage of the CE iteration cycles. CLEASE also offers `view_clusters` method in `CEBulk` and `CECrystal` classes 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 single-site basis functions; the single-site basis functions are considered valid if they form a complete set. Consequently, several definitions are used in practice. The two most widely used definitions are the original definitions by Sanchez et al. [6] and the one later developed by van de Walle [20], which is used in the Alloy Theoretic Automated Toolkit (ATAT) [20, 30]. 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 (4). 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 lattice structure [16, 31] with an exception of $LiMnO_2$, which is orthorhombic [32]. The rocksalt lattice structure consists of two face-centered cubic sublattices. For the case of the cation-disordered rocksalt lattice $LiMO_2$, one sublattice is occupied by lithium and other metal atoms while the other is occupied by oxygen atoms. The complexity of the CE model of such systems can be reduced to a cation sublattice consisting only of two element types (the oxygen sublattice is ignored). As such, an optional Boolean argument is present in CLEASE to enable/disable the reduction of the complexity of the model by ignoring the such atoms if they exist in the system. The reduction of the model complexity leads to a reduction in computational cost as it requires a smaller number of interaction terms to calculate.

A range of compositions (or concentration) of the system to be studied is specified using a `Concentration` class. First, the constituting elements of the system are categorized into the basis which they belong. For example, $LiVO_2$ in a rocksalt lattice structure is expressed using two lists: ['Li', 'V'] and ['O']. It is noted that CE needs to keep track of the location of vacancies when they are present in the system. The location of vacancies are tracked by treating a vacancy as a regular atom with its atomic symbol set to 'X' or atomic number set to zero. The $LiVO_2$ with Li vacancies is then expressed using ['Li', 'V', 'X'] and ['O'].

The range of each element (including vacancies) can be specified in one of the two convenient methods built in to the `Concentration` class. The simplest method is to specify the concentration range of each constituting element by calling `set_conc_ranges` method in `Concentration` class. For the cases where concentrations of two or more elements depend on one another, one can specify concentration range using `set_conc_formula_unit` method where the relationships between the concentrations of two or more elements can be expressed in a list of strings. For the example of $LiVO_2$ with Li vacancies, a list of strings that specifies relationship between the number of Li atoms and the number of vacancies, ["Li<x>V<1>X<1-x>", "O<2>"],

Accepted Manuscript

is passed as an argument to the `set_conc_formula_unit` method. Another argument specifying the range of the concentration variable, e.g., {'x': (0, 1)}, is also passed to the `set_conc_formula_unit` method in order to specify the concentration range of Li and Li vacancies. The concentration ranges specified by either `set_conc_ranges` or `set_conc_formula_unit` methods are internally interpreted in the `Concentration` class as a list of linear equations that specify (1) the relationships of the concentrations of constituting elements and (2) their upper/lower bounds. The advanced users can alternatively specify the coefficients of the linear equations used in the `Concentration` class if a greater flexibility is needed in specifying the concentration ranges.

3.2. Generation of Training Structures

CLEASE uses `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 figure 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 [22]. 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) [33], which have the enthalpies of formation either on or near the convex hull [34]. 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. Simulated annealing takes an initial atomic configuration and cures it at a sequence of decreasing temperatures in order to let the structure evolve towards the ground-state arrangement. In CLEASE, the user can generate ground-state structure by invoking `generate_gs_structure` method in `NewStructures` class while specifying the initial and final temperatures (intermediate temperatures are interpolated in a logarithmic scheme), number of temperatures and the number of site swaps per temperature.

The last method of generating the training set is referred to as a "probe structure" method [35, 23]. The probe structure method introduces a new structure that minimizes the mean variance of the predicted physical quantity $q(\boldsymbol{\sigma})$. The mean variance of the

Accepted Manuscript

predicted quantity q is written as [35]

$$\begin{aligned}\mathrm{Var}[\hat{q}_i] &= \frac{1}{N_{\mathrm{config}}} \sum_{i=1}^{N_{\mathrm{config}}} [\mathbf{X}_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_i^T]e^2 \\ &= \{\mathrm{tr}[(\mathbf{X}^T\mathbf{X})^{-1}\Sigma] + \mu(\mathbf{X}^T\mathbf{X})^{-1}\mu^T\}e^2 \\ &= \Lambda \cdot e^2,\end{aligned} \tag{19}$$

where e^2 is the variance of the error in the training set, Σ is the covariance matrix of the correlation functions of the training set and μ is a vector of the mean correlation functions of the structures in the training set. The probe structure is the one that reduce the value of Λ the most when introduced to the training set, which is found using the simulated annealing procedure. Similar to generating ground-state structures, the probe structures are generated by invoking `generate_probe_structure` method in `NewStructures` class.

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 [36] 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.

3.3. 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 LOO or k-fold CV 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 OLS to minimize residual sum of squared errors (RSS). It is highly likely that the ECI values found using OLS are overfitted. Therefore, Bayesian compressive sensing and ℓ_1 and ℓ_2 regularization methods are implemented, and it is highly recommended to use a regularization methods to select clusters and evaluate their ECI values.

Accepted Manuscript

A default option in the `Evaluate` class is to include all of the clusters generated using the cluster 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, Baysian compressive sensing and ℓ_1 and ℓ_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 first option allows the `Evaluate` class to be used in conjunction with other cluster selection methods such as genetic algorithm. For example, a user can optionally use genetic algorithm (included in CLEASE as a separate `GAFit` class) to pre-screen a large cluster pool and subsequently pass a subset of clusters to the `Evaluate` class. The feature to freely select a subset of a large pool of clusters along with the use of OLS, Bayesian compressive sensing and ℓ_1 and ℓ_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 LOOCV score.

To further assist the evaluation process, the `Evaluate` class contains two built-in methods that automatically determine the LOOCV 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 (λ) specified by the user. It also plots the fit of all data points in the training set to their calculated values and presents the LOO/k-fold CV score of the specified λ value. Since the most cumbersome task in determining the convergence of CE is finding the optimal λ value that yields the lowest CV score, another method, `plot_CV`, is also implemented. It takes a range and number of λ values to evaluate as inputs and returns the best λ value in the specified range along with its LOO/k-fold CV score. The `plot_CV` method also plots LOO/k-fold CV score as a function of λ and provides an option to store the results in a log file such that the users can add more λ values to the list at a later stage without having to re-evaluate the same λ values in the process.

3.4. 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.

Accepted Manuscript

The evaluation of the physical quantity $q(\boldsymbol{\sigma})$ is performed using (5), 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 [37]

$$P_{\text{acc}} = \min\left\{1, \exp\left(\frac{-(E_{\text{new}} - E_{\text{old}})}{k_B T}\right)\right\}, \tag{20}$$

where E_{new} and E_{old} are the energy of the new and old configuration, respectively. 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.

4. Examples

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_2CrO_2F system consisting of four types of elements and vacancy. All of the interactions of cluster expansions are computed from DFT calculations of energies, and the computational settings used for generating the results shown in this section are specified in section 5.

4.1. 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 ℓ_1 and ℓ_2 regularization are shown in figure 3. The ECI value of the empty cluster is found to be -3.49 eV/atom for both cases, and the ECI value of the one-body cluster is 0.27 eV/atom and 0.13 eV/atom for ℓ_1 and ℓ_2 regularization, respectively. The ECI values of empty and one-body clusters are not included in figure 3 for better visibility. The LOOCV score for the ℓ_1- and ℓ_2-regularized fit were 4.49 meV/atom and 4.67 meV/atom, respectively. The ℓ_1 regularization scheme yields a slightly lower CV score despite having a smaller number of clusters (ℓ_1-regularized fit has 20 clusters while the ℓ_2-regularized fit has 34 clusters).

A qualitative information on the thermodynamic behavior of the system can be extracted by inspecting the 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. It can be seen in figure 3 that the ECIs of the nearest-neighbor and second nearest-neighbor pairs are positive and positive, respectively. The signs indicate that the Au–Cu system energetically favors the strong mixing of the constituting elements such that the alternating patterns found in $L1_0$- and $L1_2$-type ordered structures are likely to

Accepted Manuscript

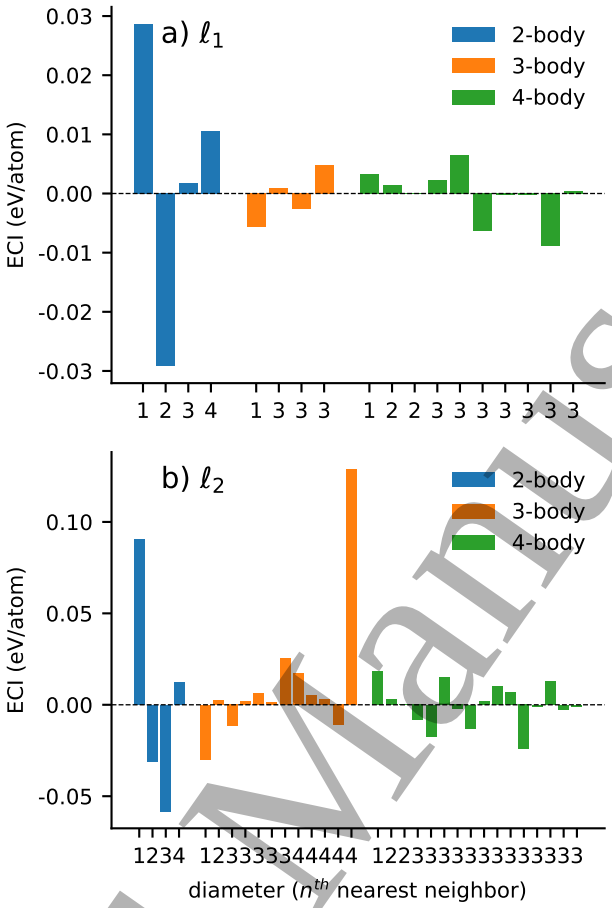

Figure 3. ECIs obtained via a) ℓ_1 regularization and b) ℓ_2 regularization.

emerge, which is in a good agreement with experimental and computational observations [33, 38, 39, 40, 41, 42, 43, 44, 45].

It is experimentally determined that Au–Cu alloys have three ordered phases at low temperatures [43, 44, 45]: $AuCu_3$, AuCu and Au_3Cu. Furthermore, the transition temperatures for $AuCu_3$, AuCu and Au_3Cu are reported to be 663 K, 683 K and ∼490 K, respectively, and they are often used as reference values for assessing the computational models [33, 38, 39]. The formation energy, free energy of formation and configurational entropy are obtained through Metropolis Monte Carlo simulations and are shown in figure 4. As the CE is trained with fully relaxed structures (zero pressure), the formation energy is determined using

$$\Delta U = U - xU_{\mathrm{Au}} - (1-x)U_{\mathrm{Cu}}, \tag{21}$$

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

Accepted Manuscript

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_3Cu) are found on the convex hull of the free energy of formation in figure 4b. Furthermore, the entropy of the ordered phases form local minima as shown in figure 4c. For comparison, the entropy of an ideal mixture, S_{ideal}, defined as [46, 47]

$$S_{\text{ideal}} = k_B \ln \frac{N!}{n_{\text{Au}}! n_{\text{Cu}}!} \approx -k_B \left[\frac{n_{\text{Cu}}}{N} \ln \frac{n_{\text{Cu}}}{N} + \frac{n_{\text{Au}}}{N} \ln \frac{n_{\text{Au}}}{N} \right], \tag{22}$$

where N, n_{Au} and n_{Cu} are the number of atomic sites, the number of sites occupied by Au and Cu atoms, respectively. The entropy of an ideal mixture is included as a gray line in figure 4c. The entropy curve resembles that of an ideal mixture at the high-temperature limit at 800 K. The curve starts to deviate from that of the ideal mixture as the temperature is lowered mainly because the entropy drops sharply at concentrations at which the ordered phases (e.g., $L1_0$ and $L1_2$) are energetically preferred. As the temperature increases, the free energy becomes a smooth convex curve with a minimum at around 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. The average fraction of sites having a different element than the same site in the ground state, f_{diff}, is tracked as the system evolves. f_{diff} is normalized by the expected fraction of different sites in a random phase, $f_{\text{diff,rnd}}$, and an order parameter, η, is defined as

$$\eta = 1 - f_{\text{diff}}/f_{\text{diff,rnd}}. \tag{23}$$

The order parameter is is used for detecting the phase transition as shown in figure 5. The computationally predicted order/disorder transition temperature of $AuCu_3$, AuCu and Au_3Cu are around 600 K, 665 K and 385 K, respectively, which are in a good agreement with the experimental reference values [43, 44, 45].

One of the most common way to describe the characteristics of a binary alloy is by constructing a phase diagram. A phase diagram can be generated computationally using a semi-grand canonical MC where a grand potential is obtained via thermodynamic integration in (16) at fixed chemical potentials. The integration starts from the low temperature limit for the ordered phases and from the high temperature limit for disordered phases where the free energy per atom is given by $k_B T \ln 2$. The phase boundary between two phases is identified by locating the intersection point between the grand potential in the two co-existing phases. The phase diagram generated via semi-grand canonical MC is shown in figure 6. The phase diagram closely resembles the phase diagrams constructed from the experimental measurements [43, 44, 45] and is also in a qualitatively agreement with the phase diagrams constructed from computational results [33, 38, 48].

Accepted Manuscript

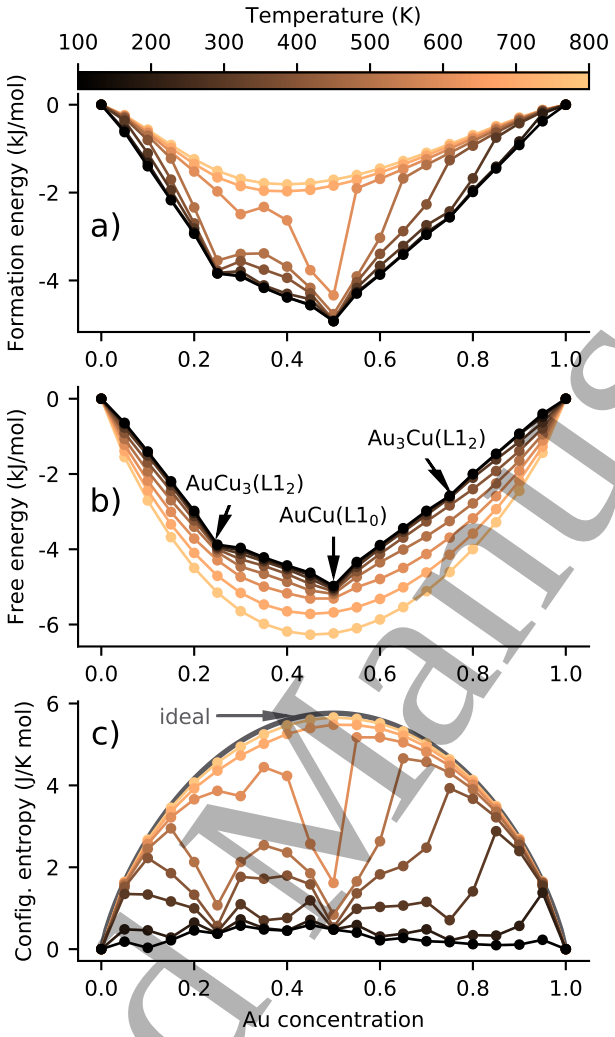

Figure 4. Thermodynamic quantities for the Au–Cu computed for temperatures ranging from 100 K to 800 K over the entire composition range. a) Formation energy. b) Free energy of formation. c) Configurational entropy.

4.2. 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_2MO_2F, 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_2MO_2F is in the vast size of the configurational space, which exhibit not only the cation disorder commonly found in lithium metal oxides [14, 16] but also anion disorder which is also present due to the mixed O/F composition [49, 50]. The fact that the underlying crystal structure of Li_2MO_2F can vary at different lithiation levels [51] adds the complexity to investigate their properties. It is, however, known that the most predominant crystal structure is of disordered rocksalt type [52], particularly at high-lithiation levels. We therefore show an example CE study of Li_2CrO_2F in a rocksalt lattice configuration.

Accepted Manuscript

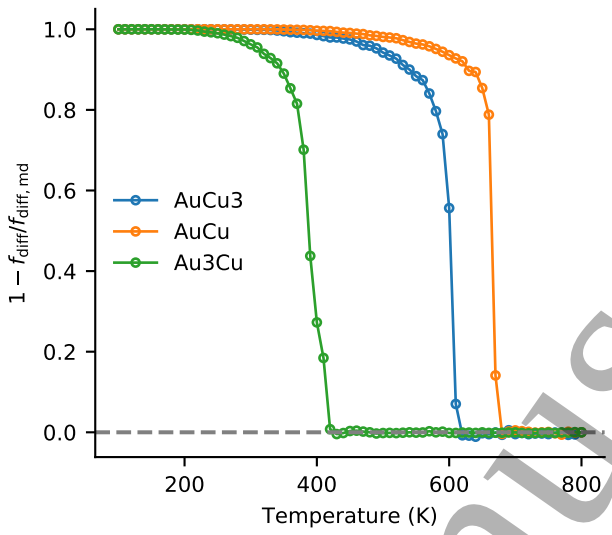

Figure 5. Order parameter as a function of temperature (1 in an ordered phase and 0 in a random phase).

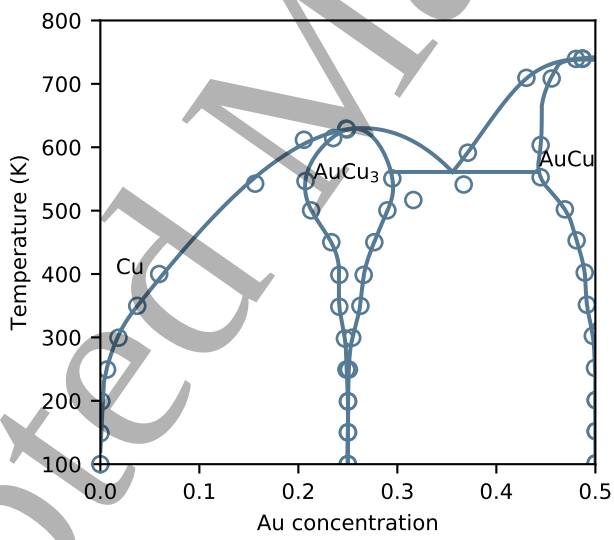

Figure 6. Phase diagram of Au_xCu_{1-x} where $0 \leq x \leq 0.5$. Circles are computed phase boundary points and lines are spline fits of the computed boundary points.

The Monte Carlo annealing study reveals that Li_2CrO_2F (i.e., fully lithiated compound) takes a layered structure at room temperature (293 K) as shown in figure 7a. 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 [16, 53, 54]. The layered structure is lost upon delithiation, which leads to disordered structures as shown in figure 7b. The emergence of disordered structures agrees well with the previous experimental observations [50, 52], and it is important to model the disordered atomic arrangement as it has a direct link to the Li transport mechanism (e.g., a presence of

Accepted Manuscript

zero-transition-metal pathways [16, 31, 55]).

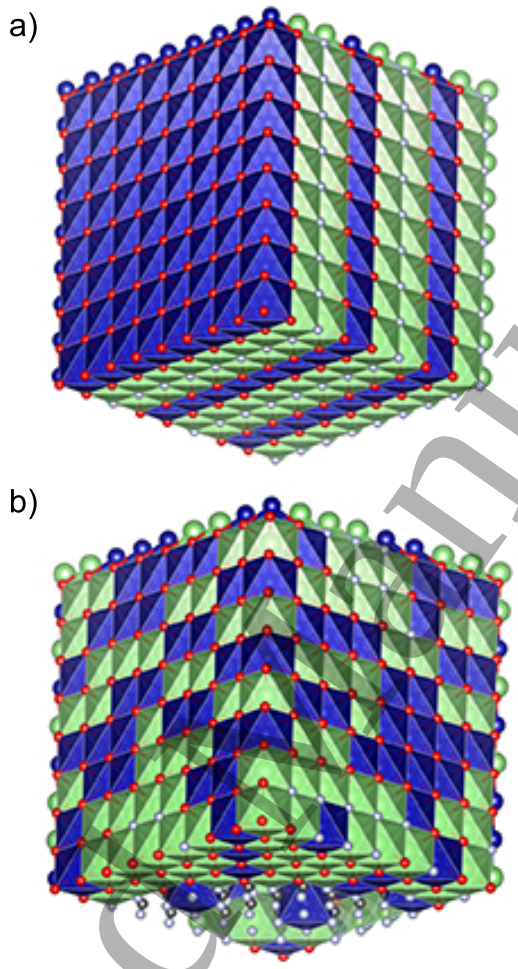

Figure 7. A snapshot of Li_xCrO_2F during the Monte Carlo run at 293 K where x is a) 2.0, b) 1.5. The Li atoms are shown in green, the Cr atoms are shown in blue, the oxygen atoms are shown in red and the F atoms are shown in white.

Thermodynamics quantities of Li_xCrO_2F can be extracted with the same procedure described for the Au–Cu system. One of the most crucial thermodynamic parameters for characterizing cathode materials for Li-ion batteries is the free energy as it is directly linked to the operating voltage of the cell. The operating voltage of Li_xCrO_2F is defined as

$$\begin{aligned} \text{voltage} &= -\frac{\mu_{\text{Li}}^{\text{cathode}} - \mu_{\text{Li}}^{\text{anode}}}{e} \\ &= -\frac{\frac{\mathrm{d}G_{\text{Li}_x\text{Cr}_2\text{F}}}{\mathrm{d}x} - \mu_{\text{Li}}^{\text{anode}}}{e}, \end{aligned} \tag{24}$$

where μ_{Li} is the chemical potential in eV per Li atom, e is an electron charge and $G_{\text{Li}_x\text{Cr}_2\text{F}}$ is the free energy of Li_xCrO_2F in eV per formula unit. Li metal is used as an anode and thus, $\mu_{\text{Li}}^{\text{anode}}$ is constant.

Accepted Manuscript

The free energy of Li_xCrO_2F and its voltage profile at 293 K are shown in figure 8. The free energy in figure 8a has three parts: free energy values computed from MC simulations, a smooth curve fitted to the computed values (using Redlich-Kister polynomials [56]) and a convex hull of the fitted curve. The curve fit is used for generating the voltage plot because the derivative of the free energy used for calculating the voltage values are susceptible to small noise that are present in the MC simulation results. Furthermore, a range in which the free energy curve is above the convex hull represents the region where a phase transition occurs: the cathode forms a mixture of two phases at which the fitted curve and the convex hull intersect. The voltage profile in figure 8b is generating using (24) where the values on the convex hull are used for $G_{\mathrm{Li}_x\mathrm{Cr}_2\mathrm{F}}$. The voltage profile in figure 8b is in a good agreement with those observed experimentally [50, 52].

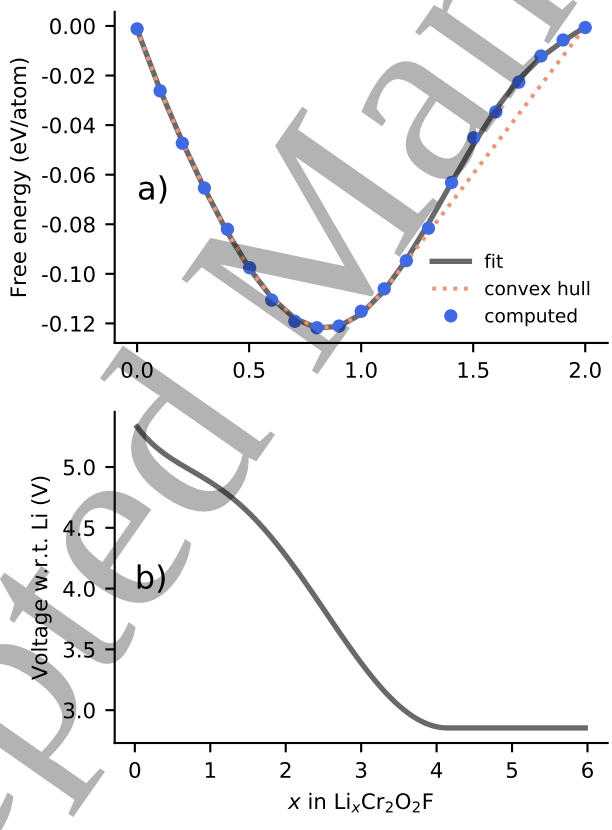

Figure 8. Free energy of formation for Li_xCrO_2F and its voltage with respect to Li metal at 293 K.

Accepted Manuscript

5. Methods

5.1. Density Functional Theory Calculations

All of the calculations are performed with the Vienna Ab initio Simulation Package (VASP) [57, 58, 59, 60] using the projector augmented-wave (PAW) method [61]. The generalized gradient approximation as parametrized by Perdew, Burke and Ernzerhof [62] 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 [63, 64] 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_2CrO_2F systems, respectively. Integrations over the Brillouin zone were carried out using the Monkhorst-Pack scheme [65] with a grid with a maximal interval of 0.04 $Å^{-1}$.

5.2. Cluster Expansion Model

The CE model for Au–Cu alloy and Li_xCrO_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 Li_xCrO_2F on a rocksalt lattice with x ranges from 0 to 2. Up to four-body clusters with the maximum diameter of 6.0 Å are generated for Au–Cu alloy. Up to four-body clusters are generated for Li_xCrO_2F with the maximum diameter of 7.0 for two- and three-body clusters and 4.5 Å for four-body clusters. ℓ_1 and ℓ_2 regularization schemes with the regularization parameter ranging from 10^{-7} to 10^2 are assessed at various maximum radii to find the optimal setting that leads to the lowest LOOCV score. For the Au–Cu alloy, ℓ_1 regularization with the maximum diameter of 6.0 Å, 5.0 Å and 5.0 Å for 2-, 3- and 4-body clusters, respectively, yields the lowest LOOCV score of 4.49 meV/atom. The minimum LOOCV score achieved using ℓ_2 regularization scheme is 4.67 meV/atom when the maximum diameter is set to 6.0 Å, 6.0 Å and 5.0 Å for 2-, 3- and 4-body clusters, respectively. Similarly, ℓ_1 regularization performed better than ℓ_2 regularization on Li_xCrO_2F with the lowest LOOCV score of 21.38 meV/atom (maximum diameter set to 7.0 Å, 7.0 Å and 4.5 Å for 2-, 3- and 4-body clusters, respectively). It is noted that although the LOOCV of Li_xCrO_2F seems larger compared to that of Au–Cu, it should be taken into account that the cohesive energy of metallic alloys are in general much smaller than those of oxyfluorides.

5.3. Metropolis Monte Carlo Simulations

For Au–Cu alloy, Metropolis Monte Carlo simulations are carried out using a $10\times10\times10$ supercell consisting of 1,000 atoms for determining thermodynamic quantities. The

Accepted Manuscript

system is equilibrated with 100 sweeps, and an average energy is collected through an additional 2,000 sweeps at each temperature for determining the thermodynamic quantities. A $30 \times 30 \times 30$ supercell consisting of 27,000 atoms is used to determine the transition temperatures and to construct a phase diagram. The transition temperatures are determined by equilibrating the systems with 100 sweeps, followed by sampling the order parameter via an additional 1,000 sweeps. A phase diagram is generated by performing a semi-grand canonical MC, where the system is equilibrated using 100 sweeps, followed an additional 1,000 sweeps to obtain an average semi-grand canonical energy at each temperature at a fixed chemical potential. A $9 \times 9 \times 9$ cell consisting of 1,458 atoms is used for Li_xCrO_2F. The temperature is gradually lowered from 10,000 K, and the structures are equilibrated at each temperature via 100 sweeps to ensure that the system is equilibrated before sampling. The average energy is then sampled via 1,000 sweeps at each temperature.

6. Conclusions

We present the implementation of CLEASE, which fully integrates the cluster expansion method to ASE package. The aim of the developed code is to make cluster expansion more accessible to non-specialists and to incorporate modern machine learning techniques to cluster expansion method in one comprehensive and versatile package. The use of the popular Python programming language and implementing the code as a part of widely used ASE package lowers the barrier for the newcomers to the field to easily learn and use CE as a part of their research methods. By automatically generating clusters and calculating the correlation functions of both semi-automatically generated and user-supplied structures, it minimizes both the possible introduction of user errors and complicated process of constructing/evaluating the cluster expansion. The capability of CLEASE is presented with two example usage cases with a different level of system complexity. The examples demonstrate that CE can correctly predict the material behavior that require statistical sampling on a large simulation cell.

7. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 711792 (FETOPEN project LiRichFCC). Authors thank valuable discussions with Dr. Juhani Teeriniemi and Prof. Kari Laasonen. JMGL acknowledges support from the Villum Foundation's Young Investigator Programme (4th round, project: *In silico design of efficient materials for next generation batteries.* Grant number: 10096).

[1] Fontaine D D 1994 Cluster Approach to Order-Disorder Transformations in Alloys *Solid State Physics* vol 47 ed Ehrenreich H and Turnbull D (Academic Press) pp 33–176

[2] Zunger A, Wang L G, Hart G L W and Sanati M 2002 *Modelling and Simulation in Materials Science and Engineering* **10** 685–706

Accepted Manuscript

[3] Lerch D, Wieckhorst O, Hart G L W, Forcade R W and Müller S 2009 *Modelling and Simulation in Materials Science and Engineering* **17** 055003

[4] Meng Y S and Arroyo-de Dompablo M E 2009 *Energy & Environmental Science* **2** 589

[5] Zhang X and Sluiter M H F 2016 *Journal of Phase Equilibria and Diffusion* **37** 44–52

[6] Sanchez J M, Ducastelle F and Gratias D 1984 *Physica A: Statistical Mechanics and its Applications* **128** 334–350

[7] Zunger A 1994 First-Principles Statistical Mechanics of Semiconductor Alloys and Intermetallic Compounds *Statics and Dynamics of Alloy Phase Transformations* ed Turchi P E A and Gonis A (Boston, MA: Springer US) pp 361–419

[8] Asta M, Ozolins V and Woodward C 2001 *JOM* **53** 16–19

[9] van de Walle A 2008 *Nature Materials* **7** 455–458

[10] Magri R and Zunger A 1991 *Physical Review B* **44** 8672–8684

[11] Franceschetti A and Zunger A 1999 *Nature* **402** 60–63

[12] Geng H Y, Sluiter M H F and Chen N X 2005 *The Journal of Chemical Physics* **122** 214706

[13] Wang R, Li X, Liu L, Lee J, Seo D H, Bo S H, Urban A and Ceder G 2015 *Electrochemistry Communications* **60** 70–73

[14] Abdellahi A, Urban A, Dacek S and Ceder G 2016 *Chemistry of Materials* **28** 3659–3665

[15] Abdellahi A, Urban A, Dacek S and Ceder G 2016 *Chemistry of Materials* **28** 5373–5383

[16] Urban A, Matts I, Abdellahi A and Ceder G 2016 *Advanced Energy Materials* **6** 1600488

[17] Kitchaev D A, Lun Z, Richards W D, Ji H, Clément R J, Balasubramanian M, Kwon D H, Dai K, Papp J K, Lei T, McCloskey B D, Yang W, Lee J and Ceder G 2018 *Energy & Environmental Science* **11** 2159–2171

[18] Hjorth Larsen A, Jørgen Mortensen J, Blomqvist J, Castelli I E, Christensen R, Dułak M, Friis J, Groves M N, Hammer B, Hargus C, Hermes E D, Jennings P C, Bjerre Jensen P, Kermode J, Kitchin J R, Leonhard Kolsbjerg E, Kubal J, Kaasbjerg K, Lysgaard S, Bergmann Maronsson J, Maxson T, Olsen T, Pastewka L, Peterson A, Rostgaard C, Schiøtz J, Schütt O, Strange M, Thygesen K S, Vegge T, Vilhelmsen L, Walter M, Zeng Z and Jacobsen K W 2017 *Journal of Physics: Condensed Matter* **29** 273002

[19] Zarkevich N A and Johnson D D 2004 *Physical Review Letters* **92** 255702

[20] van de Walle A 2009 *Calphad* **33** 266–278 (*Preprint* `0906.1608`)

[21] Nelson L J, Ozoliš V, Reese C S, Zhou F and Hart G L W 2013 *Physical Review B* **88** 155105

[22] Nelson L J, Hart G L W, Zhou F and Ozoliš V 2013 *Physical Review B* **87** 035125

[23] Seko A and Tanaka I 2014 *Journal of Physics: Condensed Matter* **26** 115403

[24] Mueller T and Ceder G 2010 *Physical Review B* **82** 184107

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

[26] Hart G L W, Blum V, Walorski M J and Zunger A 2005 *Nature Materials* **4** 391–394

[27] Díaz-Ortiz A, Dosch H and Drautz R 2007 *Journal of Physics Condensed Matter* **19**

[28] Andersen J O 2012 *Introduction to statistical mechanics* (Akademica Publishing)

[29] Tuckerman M E 2010 *Statistical mechanics: theory and molecular simulation* (Oxford University Press)

[30] van de Walle A, Asta M and Ceder G 2002 *Calphad* **26** 539–553

[31] Urban A, Lee J and Ceder G 2014 *Advanced Energy Materials* **4** 1400478

[32] Hewston T and Chamberland B 1987 *Journal of Physics and Chemistry of Solids* **48** 97–108

[33] van de Walle A and Ceder G 2002 *Journal of Phase Equilibria* **23** 348

[34] Urban A, Seo D H and Ceder G 2016 *npj Computational Materials* **2**

[35] Seko A, Koyama Y and Tanaka I 2009 *Physical Review B* **80** 165122

[36] Lonie D C and Zurek E 2012 *Computer Physics Communications* **183** 690–697

[37] Thijssen J 2007 *Computational Physics* (Cambridge University Press)

[38] Wei S H, Mbaye A A, Ferreira L G and Zunger A 1987 *Physical Review B* **36** 4163–4185

[39] Ozoliš V, Wolverton C and Zunger A 1998 *Physical Review B* **57** 6427–6443

[40] Ozoliš V, Wolverton C and Zunger A 1998 *Physical Review B* **58** R5897–R5900

Accepted Manuscript

[41] Wolverton C, Ozoliš V and Zunger A 1998 *Physical Review B* **57** 4332–4348

[42] Lysgaard S, Mýrdal J S G, Hansen H A and Vegge T 2015 *Physical Chemistry Chemical Physics* **17** 28270–28276

[43] Massalski T B, Murray J, Bennett L and Baker H 1986 *American Society for Metals, 1986,* 2224

[44] Hultgren R, Hawkins D T and Desai P D 1973 Selected values of the thermodynamic properties of binary alloys (American Society for Testing and Materials)

[45] Fedorov P P and Volkov S N 2016 *Russian Journal of Inorganic Chemistry* **61** 772–775

[46] Reuter K and Scheffler M 2003 *Phys. Rev. B* **68**(4) 045407

[47] Murty B S, Yeh J W and Ranganathan S 2014 *High-entropy alloys* (Oxford: Butterworth-Heinemann)

[48] Walle A V D and Asta M 2002 *Modelling and Simulation in Materials Science and Engineering* **10** 521–538

[49] Chen R, Ren S, Knapp M, Wang D, Witter R, Fichtner M and Hahn H 2015 *Advanced Energy Materials* **5** 1–7

[50] Chen R, Ren S, Mu X, Maawad E, Zander S, Hempelmann R and Hahn H 2016 *ChemElectroChem* **3** 892–895

[51] Cambaz M A, Vinayan B P, Clemens O, Munnangi A R, Chakravadhanula V S K, Kübel C and Fichtner M 2016 *Inorganic Chemistry* **55** 3789–3796

[52] Ren S, Chen R, Maawad E, Dolotko O, Guda A A, Shapovalov V, Wang D, Hahn H and Fichtner M 2015 *Advanced Science* **2** 1500128

[53] Mizushima K, Jones P, Wiseman P and Goodenough J 1981 *Solid State Ionics* **3-4** 171–174

[54] Van der Ven A and Ceder G 2004 *Electrochemistry Communications* **6** 1045–1050

[55] Lee J, Urban A, Li X, Su D, Hautier G and Ceder G 2014 *Science* **343** 519–522

[56] Redlich O and Kister A T 1948 *Industrial & Engineering Chemistry* **40** 345–348

[57] Kresse G and Hafner J 1993 *Phys. Rev. B* **47**(1) 558–561

[58] Kresse G and Hafner J 1994 *Phys. Rev. B* **49**(20) 14251–14269

[59] Kresse G and Furthmüller J 1996 *Computational Materials Science* **6** 15–50

[60] Kresse G and Furthmüller J 1996 *Phys. Rev. B* **54**(16) 11169–11186

[61] Blöchl P E 1994 *Physical Review B* **50** 17953–17979

[62] Perdew J P, Burke K and Ernzerhof M 1996 *Physical Review Letters* **77** 3865–3868

[63] Anisimov V I, Zaanen J and Andersen O K 1991 *Physical Review B* **44** 943–954

[64] Cococcioni M and de Gironcoli S 2005 *Physical Review B* **71** 035105

[65] Monkhorst H J and Pack J D 1976 *Physical Review B* **13** 5188–5192

Accepted Manuscript