A Novel Cluster Expansion Approach
for Finite and Infinite Systems of Arbitrary Shapes

Jesper Kristensen$^{1}$ and Nicholas J. Zabaras$^{2,*}$

$^{1}$School of Applied and Engineering Physics, 271 Clark Hall, Cornell University, Ithaca, New York 14853-3501, USA
$^{2}$School of Engineering, University of Warwick, CV4 7AL, UK

(Dated: July 30, 2014)

The cluster expansion enables fast alloy property computations especially useful for predicting alloy thermodynamics. It expands configurational alloy properties in basis functions called clusters with associated expansion coefficients called effective cluster interactions (ECI) which must be learned. The number of ECI increases when the symmetries of the system are reduced. Therefore, when applied to non-bulk low-symmetry systems, the cluster expansion faces difficulties. To improve the cluster expansion of low-symmetry structures, the coarse grained meta cluster expansion (MCE) was recently invented. However, it suffers from important practical issues. The cluster grouping in the MCE depends on both the system and the property being studied, in which sense it is not transferable. An MCE grouping is not unique either. Furthermore, it is not readily implemented in existing cluster expansion software built for bulk systems. Therefore, in this Letter, a new cluster expansion approach is introduced called the ghost lattice method (GLM). It is transferable, unique, and can readily be implemented in existing cluster expansion software. The GLM can cluster expand any geometry such as nanowires, surfaces, spheres, etc., no matter its shape complexity.

PACS numbers: 66.70.-f, 61.46.Km, 62.23.Hj

The cluster expansion [1] is a successful surrogate model for alloys, employed to speed up computation of expensive configurational alloy properties such as quantum mechanical energies [2], thermal conductivities [3], band gaps [4], etc. This is necessary, e.g., when computing phase diagrams [5] but also when optimizing alloy properties, requiring typically in the millions of property evaluations. In essence, it expands a configurational alloy property in a set of basis functions, called clusters, with associated expansion coefficients called effective cluster interactions (ECI). The clusters are groups of lattice sites. Each cluster has an associated spatial extent, which will be defined here as the largest distance between any two sites in the cluster [6]. The goal is to determine the unknown ECI by fitting to data expensive to obtain.

So far, the cluster expansion has been mainly used in bulk systems where it greatly enjoys the, typically, many symmetries of bulk structures. This is so, because the space group symmetries of the structure heavily reduce the number of unknown ECI [7], which, in turn, greatly aids the fitting process. In this Letter, however, the interest will be in cluster expanding non-bulk systems, which typically suffer from reduced symmetries.

To further understand the importance of symmetries in cluster expansions, consider the following hypothetical scenario. We want to cluster expand some function of configuration on a bulk face-centered cubic (fcc) lattice with a 5 Å lattice constant, and clusters with up till 4 points (4-pts) are to be included. This is a very common cluster expansion problem [6, 8–10]. The unknown ECI are learned by training the cluster expansion on some data set containing the property values of, say, 80–200 alloy structures. Assume, for the sake of the argument, that the more clusters that are included, the better the property is captured. For quantum mechanical energies, e.g., this assumption seems to hold well [6]. Therefore, it is of interest to include large sized clusters. Ref. [11] established general rules about the size-hierarchy of clusters. Assume that this cluster expanded, hypothetical, property follows these rules as well. In particular, we will include all 2-pt clusters of a given maximum spatial extent $L$. Then, we also include all 3-pt clusters within some lesser extent, chosen here as $L - 1$, and finally all 4-pt clusters with spatial extents less than $L - 2$ are included as well. In this sense, all clusters are bound by the length $L$ quantifying the maximum spatial extent of any included cluster. As $L$ increases, the number of clusters, generally speaking, increases as well. This, in turn, increases the number of unknown ECI.

In Fig. 1(a), we plot the number of unknown ECI versus $L$, first if we do not exploit the fcc lattice space group symmetries ($N_{\text{none}}$), and then together with the number of unknowns if we do use all the 48 space group symmetries of the bulk ($N_{\text{bulk}}$). The inset shows the ratio of the two curves from the main plot. Notice how this ratio increases rapidly with $L$. The fcc lattice is illustrated in Fig. 1(b) where red lines, with parenthesized numbers, between lattice sites, illustrate the largest 2-pt cluster for three different values of $L$. The parenthesized numbers can be matched with those above the abscissa in Fig. 1(a) to find the corresponding value of $L$. Notice that $L$ is not necessarily the length of the largest 2-pt cluster. Furthermore, no other clusters than the largest 2-pt cluster are illustrated in Fig. 1(b) for any given $L$. From Fig. 1(a) we learn that, e.g., when the 2-pt cluster has
choosing the largest 2-pt cluster included in the cluster expansion when black lines. The red lines with parenthesized numbers show as green spheres. The cubic cell structure is delineated with a 5 Å lattice constant and its lattice sites represented are matched with those in (b). (b) The fcc lattice considered $L_{\text{bulk}}$ sized numbers above the abscissa at $N_{\text{none}}$. The inset shows the ratio of the main plot curves versus the same $N$ symmetries in the system ($N_{\text{symmetries}}$). The parenthesis numbered above the abscissa at $L$ equal to 4, 5, and 7 Å are matched with those in (b). (b) The fcc lattice considered with a 5 Å lattice constant and its lattice sites represented as green spheres. The cubic cell structure is delineated with black lines. The red lines with parenthesized numbers show the largest 2-pt cluster included in the cluster expansion when choosing $L$ at three different values along the abscissa in (a).

![Graph](image)

**FIG. 1.** (Color) (a) Number of unknown ECI to be determined for a face-centered cubic (fcc) lattice, representing some hypothetical alloy, versus the maximum spatial extent $L$ of any included cluster in the cluster expansion, when not using any symmetries in the system ($N_{\text{none}}$) and when using all 48 bulk space group symmetries ($N_{\text{bulk}}$). The inset shows the ratio of the main plot curves versus the same $L$. The parenthesized numbers above the abscissa at $L$ equal to 4, 5, and 7 Å are matched with those in (b). (b) The fcc lattice considered with a 5 Å lattice constant and its lattice sites represented as green spheres. The cubic cell structure is delineated with black lines. The red lines with parenthesized numbers show the largest 2-pt cluster included in the cluster expansion when choosing $L$ at three different values along the abscissa in (a).

maximum spatial extent $L = 10$ Å we have $N_{\text{bulk}} = 118$ and $N_{\text{none}} = 3160$. These cluster sizes are very common which implies that, exploiting symmetries in the cluster expansion, critically lowers the number of unknowns. In Ref. [12], a similar plot is given for the body-centered cubic lattice, but only showing $N_{\text{bulk}}$.

The implications of the above discussion recently motivated the authors in Ref. [3] to invent a clever coarse grained meta cluster expansion, which will be referred to as simply the MCE from now on. The authors wanted to capture the thermal conductivity in silicon germanium (SiGe) nanowires versus configuration. To avoid having a large set of unknown ECI, they did not group the clusters by the few space group symmetry elements available in this non-bulk system, but instead grouped them based on physical intuition about the thermal conductivity.

From a practical point of view, the MCE has some serious issues though. First, it requires us to develop a new set of, more or less ad-hoc, cluster groupings for each new property being studied. Different researchers can come up with different groupings for the same problem: the grouping is not unique. Also, each new system, e.g., a surface instead of a nanowire, requires an entirely new set of groupings as well. The user must invent these and obtain expert knowledge about the problem at hand each time. In this sense, the MCE is not transferable, and it is very user-dependent. Secondly, it would be beneficial if the non-bulk systems could be handled with minor modifications to the current cluster expansion software packages. In particular, the alloy theoretic automated toolkit (ATAT) [5, 6, 13–15] is founded on using cluster expansions in bulk systems.

In response to these issues, the ghost lattice method (GLM) is introduced in this Letter. Similarly to the MCE, an ad-hoc cluster grouping is created in order to reduce the number of unknowns. However, the clusters will be grouped with a more practical perspective. In particular, we show how to avoid the issues of transferability and user-dependability, and we also demonstrate how the GLM can be readily implemented in existing cluster expansion software designed for bulk systems. Of course, crucial to the usefulness of the GLM is whether it can be used to capture the configurational property in question quantified in this Letter by its predicability. Both the GLM and the MCE, generally speaking, have ad-hoc ways of grouping the clusters, so it is not a priori clear how well they perform on any given problem. Therefore, after introducing the theoretical framework in the following paragraph, we apply the GLM to an application of predicting cohesive energies in a single-layer boron nitride (BN) material.

The GLM, practically speaking, extends the standard cluster expansion. To see how, we first lay out the theoretical framework of the standard cluster expansion, with further details in Refs. [14, 16]. An alloy is defined via a lattice, where each lattice site hosts one member from a set of possible components. Specifying which component occupies each lattice site gives the configuration represented by the vector $\mathbf{\sigma}$. The set of all possible configurations of the lattice forms the configuration space. Any quantity $q$, which depends on the alloy configuration, can be viewed as a function living in the space of functions over configuration space. Ref. [16] showed that this function space has an orthonormal complete basis with clusters as the basis elements. We can expand $q$ in this basis as:

$$q(\mathbf{\sigma}) = \sum_\alpha m_\alpha \langle J_\alpha (T_{\alpha \cdot \sigma}) \rangle_\alpha,$$

(1)

where $\alpha$ is a cluster represented as a vector. If $M_i$ different species can occupy site $i$, $\alpha_i$ can take values from zero to $M_i - 1$. If $\alpha_i = 0$, site $i$ is not contained in the cluster. The sum in Eq. (1) is over all symmetrically distinct clusters under the space group operations of the underlying empty lattice. Sites hosting different sets of species are to
be considered symmetrically distinct. The $J_{\alpha}$'s are the ECI and $m_{\alpha}$ is the number of clusters equivalent to $\alpha$. The average $\langle \cdot \rangle_{\alpha}$ is called the correlation function and is over all clusters $\alpha'$ equivalent to $\alpha$ by a space group operation. The quantity $\Gamma_{\alpha}(\sigma)$ is a cluster function defined as the following product over all lattice sites:

$$\Gamma_{\alpha}(\sigma) = \prod_{i} \gamma_{\alpha_{i},M_{i}}(\sigma_{i}),$$

where $\gamma_{0,M_{i}}(\sigma_{i}) = 1$, and

$$\frac{1}{M_{i}} \sum_{\sigma_{i}=0}^{M_{i}-1} \gamma_{\alpha_{i},M_{i}}(\sigma_{i}) \gamma_{\beta_{i},M_{i}}(\sigma_{i}) = \begin{cases} 1 & \text{if } \alpha_{i} = \beta_{i} \\ 0 & \text{otherwise} \end{cases},$$

which implies that the cluster functions are orthonormal: $\langle \Gamma_{\alpha}, \Gamma_{\beta} \rangle$ is one if $\alpha = \beta$ and zero otherwise. For a particular implementation of $\gamma_{\alpha_{i},M_{i}}(\cdot)$, see Ref. [14]. The cluster expansion, Eq. (1), formalizes the earlier symmetry discussion. We see how clusters, which are equivalent by a space group operation, are grouped together and assigned a single unknown expansion coefficient, namely $J_{\alpha}$. For a bulk fcc system, the result is $N_{\text{bulk}}$ unknowns instead of $N_{\text{none}}$ in Fig. 1 for any given $L$.

The discussion above is now extended, in a practical sense, with the GLM. The idea behind this new method is simple but powerful: a large box is created containing lattice sites on the same underlying lattice as the non-bulk structure in question (e.g., fcc). Then, a subset of sites inside this box is selected to represent the non-bulk structure. All sites part of the non-bulk structure behave in just the same way as if we had not created the large box in the first place, but the sites not part of the non-bulk structure are called ghost sites. By definition, if a cluster $\alpha$ contains a ghost site then its cluster function $\Gamma_{\alpha}(\cdot)$ is zeroed. Hence, it does not contribute to the cluster expanded property $q$ in Eq. (1). In Fig. 2 we consider an example. A single-layer SiGe slab on a simple cubic lattice is shown implemented in the GLM. The large box in Fig. 2(a) is on a simple cubic lattice as well, containing white ghost sites. The non-bulk structure is not visible inside this box until the cross-section, identified by the black surrounding square in both Fig. 2(a) and Fig. 2(b), is taken. This reveals the structure in Fig. 2(b). A black ellipse delineates a 3-pt cluster originating from an atom inside the structure, but which includes ghost sites. The atoms within the cluster are colored green, which is purely a visual construct. Since the cluster includes at least one ghost site, its cluster function is zero, and hence does not contribute to the cluster expanded property. Importantly, it is the large box (non-bulk system plus ghost sites), which is to be parsed by the cluster expansion software, and for all the software knows, this system is a bulk system. Thus, the ghost sites help disguise the system as a bulk system without affecting the cluster expanded property.

The next step is to choose the ad-hoc cluster groupings in the GLM. To this end, we simply use the bulk clusters, and group them by the bulk space group symmetries. Notice that this grouping is unique and thus not user-dependent. Also, it has the added feature that, if no ghost sites are present, the system is truly bulk, in which case the GLM naturally merges into the standard cluster expansion. Of course, the way GLM groups the clusters is not necessarily how the clusters should ideally be grouped in the non-bulk system, but in effect, the GLM just corresponds to a more practical way of grouping the clusters compared to the MCE, since it utilizes all the bulk features part of the cluster expansion software written for bulk systems. At the code level, one essentially just needs two things. First, to identify ghost sites, and second, to add a control statement making sure to zero a cluster function if a ghost site is part of the associated cluster.

In summary, there are at least three immediate benefits of the GLM. First, the user needs not re-invent groupings for each new property and/or system: the GLM is transferable. Secondly, the GLM provides a unique grouping of clusters for each system, so it is user-independent. Thirdly, the method fits naturally into software written to cluster expand bulk systems. In fact, in this work the GLM was implemented in ATAT in the way described in the previous paragraph.

Finally, we verify the implementation of the GLM in ATAT and demonstrate its usefulness on non-bulk structures by considering the cluster expansion of single-layer structures. These have recently received a lot of attention due to their highly useful electronic, optical, and mechanical novel behaviors as compared to the corresponding bulk structures [17, 18]. A data set of 100 single-layer planar BN structures, on a hexagonal lattice with parameters $(a, b, c) = (2.512 \, \text{Å}, 2.512 \, \text{Å}, 3.8355 \, \text{Å})$, were
created using ATAT. The visualization tool OVITO [19] was used to verify the geometries, which each had 32 non-ghost sites with a random BN configuration, and periodic boundary conditions in the plane containing the hexagonal base. A perspective view and a top view of one particular structure are shown in the insets of Fig. 3, where ghost sites are white, B atoms are red, and N atoms are blue. Clusters of types 2-pter, 3-pter, and 4-pter with respective maximum spatial extents of 7, 4, and 2 Å, were computed and used to compute the correlation functions. Then, the cohesive energy of each structure was calculated in LAMMPS [20] using the Tersoff potential optimized for B, N, and carbon [21]. The potential energy of each system was minimized via a conjugate gradient method before being outputted. Notice that ghost sites were removed before parsing the structure into LAMMPS. Then, 90 structures were selected as training data and the remaining 10 as test data. A simple least squares method, coupled with a repeated random sub-sampling validation scheme of 10 splits, was used to obtain the ECI on the training data and then used to test the GLM on both the training data and the test data. As a side note, quantifying predictive uncertainties is not in the aim of this work. The GLM performance on the data sets are shown in Fig. 3, where the predictions made by GLM of the cohesive energies are compared to the “ground truth” values from LAMMPS. The training error is 11 meV per atom and the test error is 13 meV per atom. The test error is comparable to the training error, which provides evidence that the GLM is implemented correctly and, importantly, that it is a useful method for non-bulk systems.

In conclusion, the MCE was recently introduced in Ref. [3] as a way of cluster expanding non-bulk low-symmetry systems. It groups the clusters by using physical intuition. This is necessary in order to reduce the number of unknown ECI to be learned. However, this Letter discussed some important practical issues with the MCE. To avoid these issues, a novel cluster expansion technique was introduced, called the GLM. The main drawback of the GLM is that its performance on any non-bulk system compared to the MCE is a-priori unknown. The present application provided evidence that the GLM can be useful in cluster expanding non-bulk structures. In future work, it will be interesting to apply the GLM to other non-bulk structures and ideally compare to the MCE as well. Finally, notice how the GLM can potentially open up for an exciting new research area not explored so far in the literature. Since the GLM can be applied to any geometry, on some fixed lattice, one can imagine optimizing not only the alloy configuration, but also the alloy shape. In particular, one could find not only the alloy configuration, which optimizes some particular property, but also determine whether the alloy should be shaped into, e.g., a cylinder, a sphere, or a cube, on some fixed underlying lattice.

NJZ as ‘Royal Society Wolfson Research Merit Award’ holder acknowledges support from the Royal Society and the Wolfson Foundation. NJZ also acknowledges strategic grant support from EPSRC to the University of Warwick for establishing the Warwick Centre for Predictive Modeling (grant EP/L027682/1). In addition, NJZ as Hans Fisher Senior Fellow acknowledges support of the Technische Universität München - Institute for Advanced Study, funded by the German Excellence Initiative and the European Union Seventh Framework Programme under grant agreement no 291763. The research of both authors while at Cornell was supported by an OSD/AFOSR MURI09 award on uncertainty quantification, the US Department of Energy, Office of Science, Advanced Scientific Computing Research and the Computational Mathematics program of the National Science Foundation (NSF) (award DMS-1214282). This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Finally, both authors acknowledge support from the Pacific Northwest National Laboratory Institutional Computing team.
Corresponding author; nzabaras@gmail.com; http://www.zabaras.com