APS physics

CHORUS

This is the accepted manuscript made available via CHORUS. The article has been published as:

Predicting low-thermal-conductivity Si-Ge nanowires with a modified cluster expansion method

Jesper Kristensen and Nicholas J. Zabaras

Phys. Rev. B **91**, 054105 — Published 12 February 2015

DOI: 10.1103/PhysRevB.91.054105

Predicting Low-Thermal-Conductivity Si-Ge Nanowires with a Modified Cluster Expansion Method

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] *Warwick Centre for Predictive Modelling, University of Warwick, CV4 7AL, UK*

(Dated: January 20, 2015)

We introduce the cluster expansion ghost lattice method, which extends the applicability of existing cluster expansion software, to cluster expand structures of arbitrary finite and infinite geometries in a fast, unique, and transferable way. The ghost site is introduced that zeroes the cluster function of any cluster which includes it. This enables the use of bulk clusters grouped by bulk symmetries in non-bulk systems and distinguishes the cluster expansion ghost lattice method from a regular ternary cluster expansion with an inactive vacuum atom type. Even though the method does not treat surface terms, it can be used as an efficient way of obtaining the bulk term in [Lerch, D. et al., 2009. Model Simul Mater Sc]. We use the method to learn the thermal conductivity of Si-Ge nanowires, oriented along the [111] direction on a diamond lattice, versus their configuration of Si and Ge atoms. Once learned, the ghost lattice method cluster expansion is shown to be able to predict the lowest-thermal-conductivity nanowire configuration, in agreement with the configuration recently found in [Chan, M. et al., 2010. Physical Review B].

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

I. INTRODUCTION

Alloy properties such as the quantum mechanical energy,[1,2] the thermal conductivity,[3] the band gap,[4] etc., are typically expensive to obtain in terms of time and other resources, even on a supercomputer. We can therefore only observe the property of a few alloy structures, on the order of thousands in the best case. But if our task at hand is to optimize some property, e.g., say we want the lowest thermal conductivity configuration, then hundreds of millions of evaluations are typically needed.

The computer code used to obtain the property $q(\cdot)$, such as VASP[5] or LAMMPS,[6] provides a data set of, what we assume to be, true properties for a set of inputs of our choosing. Based on these true properties, we can attempt to approximate the value of any output from the computer code by using a truncated cluster expansion[7] with basis functions called clusters, which are quick to compute. This brings the optimization task back into our reach. The clusters have associated expansion coefficients called *effective cluster interactions* (ECI) and the space group symmetry operations of the lattice groups together the clusters under the same expansion coefficient. The unknown ECI are determined by fitting to the expensive data. Once known, the ECI can provide an estimate of $q(\cdot)$ for, in principle, *any* input, many orders of magnitude faster than the time needed to compute the true property.

To perform a cluster expansion, the open source software package ATAT[8–13] has been available for many years. Non-bulk systems such as surfaces can be cluster expanded with ATAT,[14] but consider now using ATAT to cluster expand *any* geometry—even one fully void of periodicity and/or large in size. The computation of the clusters in ATAT, and of the associated correlation functions, the latter to be introduced later in this paper, for arbitrary structure shapes, easily becomes prohibitively expensive to compute for large unit cells. One of the main reasons is that the number of clusters to compute, as will be shown in Section III, grows extremely fast with the unit cell size.

Here we introduce the *cluster expansion ghost lattice method* (CE-GLM). The CE-GLM is an idea, which enables, via minor modifications, a cluster expansion software package to cluster expand structures of *arbitrary* shapes with the same computational effort as the corresponding bulk system. Its main strength is in the ease of implementation in existing cluster expansion code. What makes the method different than a regular cluster expansion with vacuum atom types, is in the way it handles large low-symmetry structures by using a practically convenient ad-hoc grouping of the clusters. The work of Ref. 3 provided evidence that *some* ad-hoc grouping of the clusters should be expected to be necessary for low-symmetry systems since the number of unknowns otherwise becomes insurmountable. We propose to make the ad-hoc choice unique and transferable.

In order to demonstrate that our choice of using bulk clusters in non-bulk low-symmetry systems can be useful, we set out to reproduce the work of Ref. 3, but by using the CE-GLM implemented in ATAT. If successful, this will provide a strong case for the usefulness of the CE-GLM method.

The paper is organized as follows. We lay down the theoretical groundwork for the cluster expansion in Section II followed by a discussion of how the CE-GLM extends it in Section III. Then, in Section IV A, we start out with an overview of the nanowire problem considered, followed by details on how the nanowire data set was obtained in Section IV B. In Section IV C we discuss how the

thermal conductivities of the nanowires were computed. The results of this work are provided in Section V before concluding in Section VI.

II. THE CLUSTER EXPANSION

Define an alloy via a lattice and consider a set of atoms which can occupy each lattice site. A single configuration, represented by a vector $\boldsymbol{\sigma}$, is given by specifying which atom sits where on the lattice sites. The set of all possible configurations of the lattice forms the configuration space. If we let $q(\cdot)$ be a configuration-dependent alloy property, Ref. 15 showed that we can expand this property in an orthonormal basis, which is exact if untruncated:

$$q(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\alpha}} m_{\boldsymbol{\alpha}} J_{\boldsymbol{\alpha}} \langle \Gamma_{\boldsymbol{\alpha}'}(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\alpha}}, \tag{1}$$

where $\boldsymbol{\alpha}$ is a cluster represented as a vector. If M_i different atoms can occupy lattice site i, α_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 inequivalent clusters under the space group operations of the underlying empty lattice. Sites hosting different sets of atoms are to be considered symmetrically distinct. The $J_{\boldsymbol{\alpha}}$'s are the ECI and $m_{\boldsymbol{\alpha}}$ is the number of clusters equivalent to $\boldsymbol{\alpha}$. The average $\langle \cdot \rangle_{\boldsymbol{\alpha}}$ is called the *correlation function* and is over all clusters $\boldsymbol{\alpha}'$ equivalent to $\boldsymbol{\alpha}$ by a space group operation. All such equivalent clusters are grouped into what can be called a *cluster orbit*. The quantity $\Gamma_{\boldsymbol{\alpha}}(\boldsymbol{\sigma})$ is a *cluster function* defined, for a single cluster, as the following product over all lattice sites in the system:

$$\Gamma_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) = \prod_i \gamma_{\alpha_i, M_i}(\sigma_i),$$

where

$$\gamma_{0, M_i}(\sigma_i) = 1 \tag{2}$$

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, i.e., $\langle \Gamma_{\boldsymbol{\alpha}}, \Gamma_{\boldsymbol{\beta}} \rangle$ is one if $\boldsymbol{\alpha} = \boldsymbol{\beta}$ and zero otherwise. The cluster expansion is thus an expansion over the average cluster function in all cluster orbits. Due to Eq. (2), a cluster can be thought of as a subset of the lattice sites and is called an n-point cluster if it contains n sites. Ref. 16 provides more details and discusses a particular implementation of $\gamma_{\alpha_i, M_i}(\cdot)$.

From Eq. (1), the cluster expansion of the thermal conductivity, the latter to be introduced later as κ, can be expressed in the linear regression form, with added Gaussian noise $\boldsymbol{\epsilon}$ as:

$$\boldsymbol{\kappa} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3}$$

where any dependence on $\boldsymbol{\sigma}$ is suppressed, the ith element of $\boldsymbol{\kappa}$ is κ of structure i with its associated noise in the ith position of $\boldsymbol{\epsilon}$, $\mathbf{X}$ is called the design matrix and contains, in the ith row and jth column, the correlation function of structure i for cluster orbit j, and the ECI is the vector $\boldsymbol{\beta}$. In this work, we will use the least squares method to obtain a fitted set of ECI denoted $\hat{\boldsymbol{\beta}}$ and as a side note, we will not be interested in a thorough estimate of the uncertainty in the thermal conductivity. We refer to the framework developed in Ref. 17 for this.

In summary, our task is to obtain $\boldsymbol{\kappa}$ and $\mathbf{X}$ for some data set that we will denote $\mathcal{D}$, and then to solve for the ECI $\hat{\boldsymbol{\beta}}$. We emphasize that, since the CE-GLM is employed in computing the correlation functions, and therefore used to compute $\mathbf{X}$, a good $\hat{\boldsymbol{\beta}}$ cannot be learned if the CE-GLM does not perform well.

III. THE CLUSTER EXPANSION GHOST LATTICE METHOD

The CE-GLM removes the computational overhead of cluster expanding non-bulk possibly low-symmetry systems compared to bulk structures. The cluster expansion of 2D sheets, nanowires, spheres, etc., takes the same time as cluster expanding a bulk structure equivalent in size—the equivalence quantified, e.g., by using the bulk structure which just exactly encloses the non-bulk structure. This is accomplished by introducing a new type of site in the system called a *ghost* site. The cluster function $\Gamma_{\cdot}(\boldsymbol{\sigma})$ of any cluster, which includes a ghost site, is zeroed.

In addition, for low-symmetry systems, the CE-GLM chooses to use the bulk clusters and group them by the bulk symmetries. We discuss the implication of this at the end of this section, but before that, we present how one employs the CE-GLM. A large box is created containing lattice sites on the same underlying lattice as the non-bulk structure in question. Then, a subset of lattice sites inside this box is selected to represent the non-bulk structure. The remaining sites are the *ghost sites*. We note that this is not a new idea in the cluster expansion framework; our modification is in how the ghost sites are kept in the system and thus enables the use of bulk clusters grouped by their bulk symmetries. In Fig. 1 we consider an example of a CE-GLM implementation for a hypothetical system, namely a single-layer Si-Ge slab on a simple cubic (sc) lattice. The large box in Fig. 1(a) is on an sc lattice as well, containing white ghost sites. As a side note, we can think of this entire box as the equivalent bulk structure. The non-bulk structure is not visible inside this box until the cross-section, identified by the black surrounding square in both Fig. 1(a) and Fig. 1(b), is taken. This reveals the structure in Fig. 1(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 in-

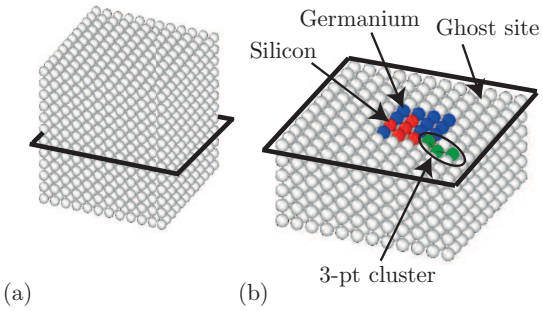

FIG. 1. (Color) Demonstration of the CE-GLM for some hypothetical single-layer slab of Si (Silicon; red) and Ge (Germanium; blue) on a simple cubic lattice. Ghost sites are white. The black square surrounding the boxes in (a) and (b) show where a cross-section is taken in (a) to reveal the slab in (b). A black ellipse in (b) delineates a 3-pt cluster containing ghost sites. The sites in the cluster are colored green, purely as a visual construct, to associate them with the cluster.

cludes at least one ghost site, its cluster function is zero, and hence does not contribute to the cluster expanded property.

The CE-GLM employs the bulk clusters and associated bulk symmetries in the cluster expansion even for non-bulk geometries. To further understand the importance of symmetries in cluster expansions, and why an ad-hoc grouping is expected to be necessary in most low-symmetry 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.[18–21] 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.[18] Therefore, it is of interest to include large sized clusters. Assume that the hypothetical property follows the size-hierarchy rules in Ref. 22. 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. 2(a), we plot the number of unknown ECI versus L, first if we do not exploit the fcc lattice space group symmetries (N_{none}), and then together with the number of unknowns if we do use all the 48 space group symmetries of the bulk (N_{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. 2(b) where red lines 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. 2(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. 2(b) for any given L. From Fig. 2(a) we learn that, e.g., when the 2-pt cluster has maximum spatial extent $L = 10$ Å we have $N_{\text{bulk}} = 118$ and $N_{\text{none}} = 3160$. Since these cluster sizes are very common it implies that, exploiting symmetries in the cluster expansion critically lowers the number of unknowns. In Ref. 23, a similar plot for N_{bulk} is given for the body-centered cubic lattice.

The choice of grouping the clusters by the bulk symmetries of the underlying lattice in the CE-GLM makes the

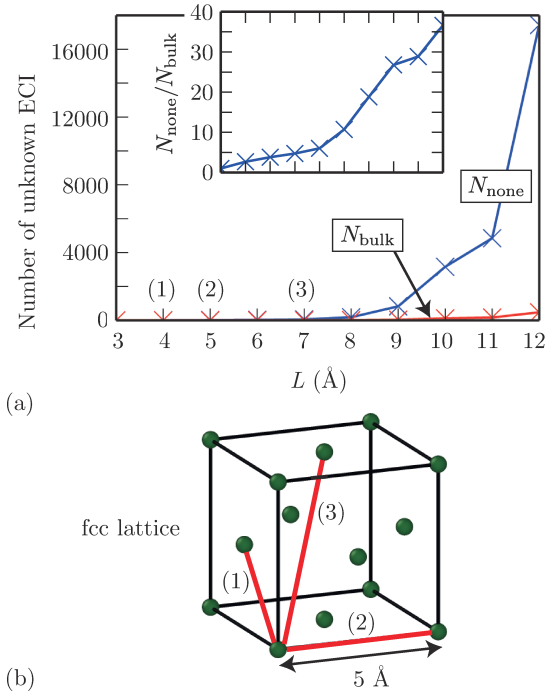

FIG. 2. (Color) (a) Number of unknown ECI to be determined for a face-centered cubic (fcc) lattice, representing some hypothetical alloy, versus the maximum spatial extent L of any included cluster in the cluster expansion, when not using any symmetries in the system (N_{none}) and when using all 48 bulk space group symmetries (N_{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 balls. The cubic cell structure is delineated with black lines. The red lines with parenthesized numbers show the maximum allowed length of any 2-pt cluster included in the cluster expansion when choosing L at three different values along the abscissa in (a).

clusters unique in the sense of being user-independent. Also, if one develops a cluster grouping, particular to a problem at hand, this will, almost never, be useful for other structure shapes thus requiring new groupings to be developed for each new property and/or geometry. This issue is dealt away with in the CE-GLM since the choice is to use the bulk clusters. It is in this sense that the CE-GLM is transferable. Also, the choice presented in this work has the added feature that, if no ghost sites are present, the system is truly bulk, in which case the CE-GLM naturally merges into the standard cluster expansion without being a special case. Finally, the choice of cluster-grouping-scheme for low-symmetry systems does not increase computational efforts when compared to the bulk because of the ghost sites.

Finally, it is important to point out the fact that we do not claim that the choice of bulk clusters will always work equally well on any geometry and configurational property. If it does not perform well, one may have to consider using a more elaborate grouping scheme. But the CE-GLM could provide a starting point for cluster expanding exotic geometries and the method might turn out to perform sufficiently well. In the context of the work presented in Refs. 24 and 25, the CE-GLM can also be used as a computationally efficient way to obtain the bulk contribution to the energy in Eq. (13) in Ref. 24. To do so, let us call the energy computed for the surface structure with the ghost region filled with ghost sites E_{glm}. Let then E_{ghost} be the energy contributions only from the clusters extending from the surface structure into the ghost region, but where this region is filled with the atoms that would have been there had the surface not formed in the first place (i.e., the ghost region does not contain ghost sites in this case). To get E_{ghost} one must implement the GLM in a way that can aggregate the contributions to the energy from just the clusters extending from the surface region into the ghost region. As an example, if the GLM has been implemented by identifying these clusters and ignoring their contribution to the total energy, as was the case in this work, one now simply needs to sum these contributions to form E_{ghost}. The bulk contribution then equates to $E_{\text{glm}} + E_{\text{ghost}}$.

IV. APPLICATION: PREDICTING LATTICE THERMAL CONDUCTIVITY OF SI-GE NANOWIRES

A. Nanowires as Thermoelectric Devices

We consider Si-Ge nanowires used as thermoelectric devices.[26,27] The efficiency of a thermoelectric device is quantified by the dimensionless figure of merit:[28]

$$ZT = \frac{S^2 \sigma_{\text{el}}}{\kappa} T_{\text{avg}}, \tag{4}$$

where S is the Seebeck coefficient quantifying the induced thermoelectric voltage in response to a temperature gradient across the device, σ_{el} is the electrical conductivity of the device, T_{avg} its average temperature, and κ its thermal conductivity. The thermal conductivity κ can be decomposed into contributions from the electrons (el) and the phonons/lattice (lat) writing $\kappa = \kappa_{\text{el}} + \kappa_{\text{lat}}$, but typically, and this will be assumed here as well, $\kappa_{\text{el}} \ll \kappa_{\text{lat}}$.[28,29] Therefore, from now on, we let $\kappa \approx \kappa_{\text{lat}}$. Thus we see from Eq. (4) that, keeping everything else constant, a minimal κ leads to a maximal ZT. The best thermoelectric devices currently have $ZT \approx 2$ at room temperature,[28,30,31] and with $ZT > 3$, we can begin replacing compressor-based refrigeration units.[32–34]

Ref. 35 discovered that Si nanowires possessed two orders of magnitude lower thermal conductivity compared to the bulk value of 150 W/m.K.[36] This brought a lot of interest into studying nanowires for thermoelectric applications. We are interested in the heat transferred along the nanowire axis, so we now let κ mean the axial lattice thermal conductivity. The thermal conductivity has been shown to depend on the configuration of the Si and Ge atoms,[37] and is thus well suited to be described by the cluster expansion.

In the present work, 50 nm long Si nanowires are considered, with roughened surfaces—roughened to help scattering phonons—which currently achieve a ZT around 1. A factor of, at least, 10 reduction in κ is, however, still needed. The authors in Ref. 3 used a clever meta cluster expansion (MCE) to capture κ. In essence, this is a cluster expansion technique which uses an ad-hoc grouping of the clusters based on physical intuition. Then, in conjunction with the learned MCE, a genetic algorithm was employed to find the configuration with lowest κ. They found that, a silicon nanowire with pure planes of germanium (PPG)—the planes spaced about 1 nm apart—was the lowest thermal conductivity configuration. While the method was very successful, extending it to other properties and/or system geometries requires a new set of cluster groupings which can be a complex task.

B. Creating the Nanowire Data Set

This section details how we obtained the nanowire data set $\mathcal{D}$. First, computing the correlation functions requires a set of cluster orbits. Given these cluster orbits, the data set generation is a two-step process. First, the nanowire must be created, and its thermal conductivity computed. Then, we must compute the correlation functions in Eq. (1) using the orbits. Carrying this out for all wires gives $\boldsymbol{\kappa}$ and $\mathbf{X}$, respectively, in Eq. (3). As discussed in the following section, we used molecular dynamics (MD), as implemented in the *large-scale atomic/molecular massively parallel simulator* (LAMMPS), to compute κ.

Interestingly, it was shown in Ref. 3 that, instead of modeling long (∼50 nm) nanowires with roughened surfaces, as we intend to do here, we can instead model short (∼2 nm) ones with perfect surfaces. This is com-

putationally easier, as we need not worry about modeling the surface roughness, and we have fewer atoms overall. Therefore, we adopt this approach here as well.

The CE-GLM requires us to create two representations of the same nanowire structure. One representation for use in ATAT-GLM to compute the correlation functions. We call this the *ATAT representation*. The other representation was generated for directly computing κ in LAMMPS without periodic images and ghost sites. We call this representation the *LAMMPS representation*. Creating both representations started out at a common origin as follows: a large cubic box containing a diamond lattice, with a lattice constant $a_0 = 5.431$ Å was created. A cylindrical region was selected along the [111] direction representing the wire. OVITO[38] was used along the way to verify the nanowire geometry. All sites inside the box, but not inside the nanowire or any of the periodic images along its axis, were selected as ghost sites. The non-ghost atoms, including images of the wire along its axis, to be shown shortly, were chosen as Si or Ge depending on the particular configuration of the wire. This is the ATAT representation of the nanowire. Notice that the number of periodic images needed here depends on the cluster with largest spatial extent. One periodic image was enough along each axial direction of the nanowire in our case, since the largest cluster orbit included was 1.2 nm and the wire had a length of roughly 2 nm. Further details about exactly which cluster orbits were included are presented later.

Fig. 3 shows the ATAT representation of the nanowire. At first, it can be difficult to see the atoms in the nanowire. This is because all the ghost sites are included—shown as smaller black atoms—and they engulf the nanowire, just as in Fig. 1(a). Notice the repetition of the wire by one image along its axis in both directions. The image atoms are colored slightly lighter than the atoms in the main wire. Of course, if we were not modeling the wire as being infinitely periodic along its axis, the images should not be there. Also, to be on the safe side, the distance from the wire to any point on the surface should be larger than the longest length of any cluster included for the same reasons, which can easily be achieved by making the cell large enough; there is little increased overhead in computing the correlation functions with an increase in the number of ghost sites. The entire geometry is bulk, so it is straightforward to parse with ATAT-GLM.

Next, to create the LAMMPS representation of the nanowire, all periodic images of the ATAT representation, and all the ghost sites, were removed. The nanowire axis was rotated to lie along the [100] direction in the box. The box dimension along the wire axis was changed to the periodic length of the wire closest to 2 nm, which is $2\sqrt{3}a_0 \approx 1.88$ nm along [111] in a diamond lattice before thermalization, to be discussed in the next section. The nanowire diameter was made 1.5 nm, also before thermalization, leaving it with 220 atoms. Please consult Fig. 4 for a visualization of the LAMMPS representation

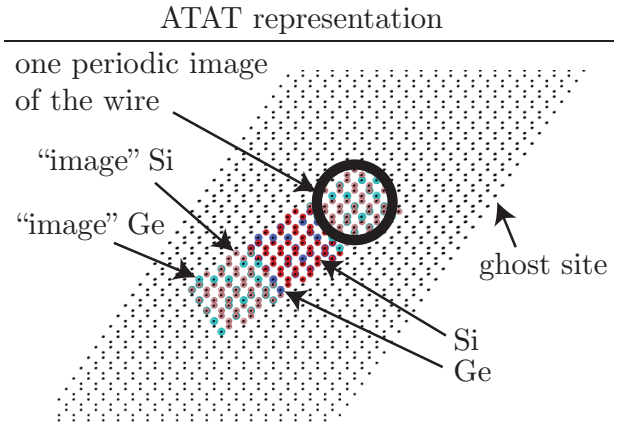

FIG. 3. (Color) The ATAT representation of an arbitrary Si (red atoms) nanowire from the data set $\mathcal{D}$ alloyed with 6.4 % Ge (blue atoms). Ghost sites are black atoms, made small so the wire becomes visible. The atoms in the two images of the wire along its axis are colored with a slightly lighter color to distinguish them from the main wire. The thick black circle identifies one of the two nanowire images. Ge atoms are made slightly larger than Si atoms for visual clarity.

of the nanowire. Fig. 4(a) and (b) show the end and side view, respectively, with the black box in each view being the LAMMPS simulation cell. In Fig. 4(b), notice how the wire becomes periodic along its axis when applying periodic boundary conditions along x in LAMMPS.

C. Computing the Nanowire Thermal Conductivities and the Design Matrix

To compute the thermal conductivities $\boldsymbol{\kappa}$ of the LAMMPS representations of the nanowires, the Green-Kubo method[39] was employed. Ref. 40 provides a thorough study, and detailed discussion, of this. We will refer to the LAMMPS representations of the nanowires as simply (nano)wires in this section. The wire axis is aligned along x with periodic boundary conditions, and fixed boundary conditions along dimensions transverse to the axis, so

$$\kappa = \frac{V_{\text{wire}}}{k_{\text{B}}T^2} \lim_{\tau_m \to \infty} \int_0^{\tau_m} \langle J_x(\tau) J_x(0) \rangle \mathrm{d}\tau, \tag{5}$$

where V_{wire} is the nanowire volume, T is its temperature, k_{B} is the Boltzmann constant, $J_x(\tau)$ is the x component of the heat flux vector $\boldsymbol{J}$ at time τ and $\langle J_x(\tau)J_x(0)\rangle$ is the heat flux autocorrelation function (HCACF) at lag τ. The heat flux vector $\boldsymbol{J}$ must be collected while the system is in a constant particle number N, system volume V, and energy E (NVE) ensemble following a thermalization according to the constant N, V, and temperature T (NVT) ensemble.

The MD time step was set to 1 fs and the heat flux

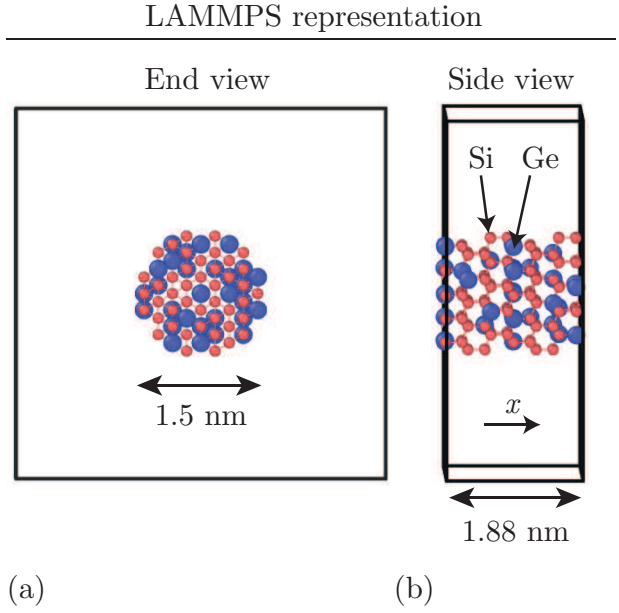

FIG. 4. (Color) The corresponding LAMMPS representation of the nanowire in Fig. 3. The black surrounding box is the LAMMPS simulation cell. (a) Shows the nanowire looking down its axis, and (b) illustrates the same wire from the side. The horizontal arrow inside the cell in (b) identifies the x-axis. Double headed arrows report sizes of the wire, and the smaller arrows in (b), over the wire, identify the atom types. Ge atoms are made slightly larger than Si atoms for visual clarity.

vector was defined as:

$$\boldsymbol{J} = \frac{1}{V_{\text{wire}}}\left[\sum_{i=1}^{N} e_i \boldsymbol{v}_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\left(\frac{\partial \phi_{ij}}{\partial \boldsymbol{r}_j}\cdot \boldsymbol{v}_j\right)\boldsymbol{r}_{ij}\right],$$

where the sums are over all N atoms in the nanowire, $\boldsymbol{r}_i$ is the position of atom i, $\boldsymbol{r}_{ij} \equiv \boldsymbol{r}_i - \boldsymbol{r}_j$, e_i and $\boldsymbol{v}_i$ are the total per-atom energy (kinetic plus potential) and velocity of atom i, respectively, and ϕ_{ij} is the interaction energy between atoms i and j modeled with the Tersoff scheme.[41] Note that $\langle\cdot\rangle$ in Eq. (5) is an ensemble average, which, in an MD simulation, becomes a time average, assuming, as we will, ergodicity is satisfied. The time τ_m should be much smaller than the total simulation time, but larger than the time required for the HCACF to decay to zero.[40]

To collect $\boldsymbol{J}$, the system needs to be thermalized in an NVT ensemble at $T = 300$ K. In order to do this, it is important to thermalize the nanowire surface, especially for wires of this small size having large surface to volume ratios. We found the following procedure successful. The atomic coordinates in the wire were first adjusted using a conjugate gradient method until finding a (possibly local) minimum in the potential energy. We then began an annealing process to thermalize the surface: an initial set of velocities were chosen for each atom from a mean zero Gaussian distribution with standard deviation scaled to 1000 K, and the system was run for 500 ps. Then, the temperature was gradually lowered 100 K at a time over 10 ps. At each temperature, we ran the system for 100 ps before decreasing by another 100 K, continuing in this way until reaching 300 K. After obtaining a room temperature NVT ensemble, the nanowire axis was pressurized to 1 bar using a constant N, pressure P, and $T = 300$ K ensemble. This was necessary to reduce strains from the size mismatch between the Si and Ge atoms. Indeed, we observed the pressure to be, typically, around 500 bar before pressurizing to 1 bar. After the NPT ensemble run, the system was switched back to the room temperature NVT ensemble and run for 1 ns.

Following this, the system was finally switched to an NVE ensemble, and run for 16 ns. The total energy stayed constant to within numerical accuracy expected with the Verlet integrator, and the temperature stayed constant, on average, by oscillating around 300 K with a 40 K amplitude. The axial heat flux J_x was output every fourth time step (for memory reasons), and the HCACF computed from this. The HCACF was then integrated. Determining when to stop the integration, i.e., choosing the best τ_m in Eq. (5), called τ_m^*, was done as follows. Forty moving averages of various window sizes ranging from 50 ps to 200 ps were computed. The time point where the standard deviation of all window sizes obtained a minimum was chosen as τ_m^*. In the context of Ref. 42, the above method is a way of finding the "convergence region" when convergence of the HCACF integral is clear, which it was for some wires, but also to find the "neck region" in case convergence is not clear. For more details on issues with the integration of the HCACF please see Ref. 42.

To verify our overall implementation, and to show explicitly how the choice of τ_m was carried out, please refer to Fig. 5. We ran a pure Si wire, Fig. 5(a), and the PPG wire, Fig. 5(b). Our approach gives 4.1 ± 0.4 W/m.K for the pure Si wire and 0.12 ± 0.03 W/m.K for the PPG wire. The errors are determined from re-running the wires with different initial random velocity distributions. In comparison, Ref. 3 obtains 4.1 ± 0.3 W/m.K for the pure Si wire and 0.23 ± 0.05 W/m.K for the PPG wire with the same method of determining the standard deviations. Discrepancies between these numbers are due to different MD software used, a different surface reconstruction technique, and different thermalization and pressurization timings and methods.

Finally, this is how the design matrix $\mathbf{X}$ was obtained. The ith row of the design matrix contains the cluster-orbit-averaged correlation functions of nanowire i. It is *a priori* unclear which cluster orbits are needed to describe thermal conductivity. In this work we used up till 5-pt cluster orbits with maximum spatial extents of the cluster orbits chosen as 12, 8, 6, and 5 Å for the 2, 3, 4, and 5-pt cluster orbits, respectively. Since the largest cluster orbit is 1.2 nm, one periodic nanowire image, which is 1.88 nm, is enough along the axis. For each structure in

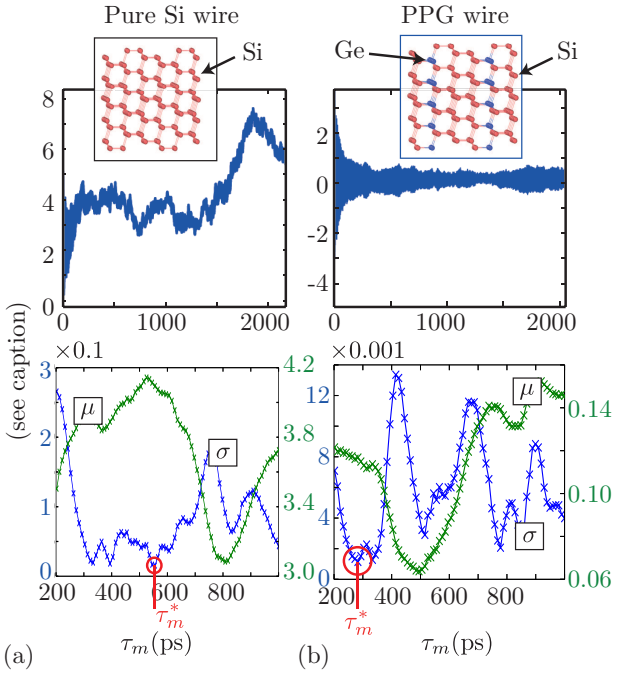

FIG. 5. (Color) Demonstration of the method used to obtain the best τ_m in Eq. (5) called τ_m^*. All ordinates are in units of W/m.K. (a) Shows the pure Si wire and (b) the PPG wire. Top plots are κ versus τ_m in Eq. (5). In the bottom plots the μ/σ graph is the mean/standard deviation of 40 different unweighted running averages of window sizes ranging from 50 to 200 ps applied to the top graphs. The μ/σ graph is measured on the right/left ordinate. The red circles with vertical lines crossing the abscissa mark the minima of the σ graphs and hence the times τ_m^*.

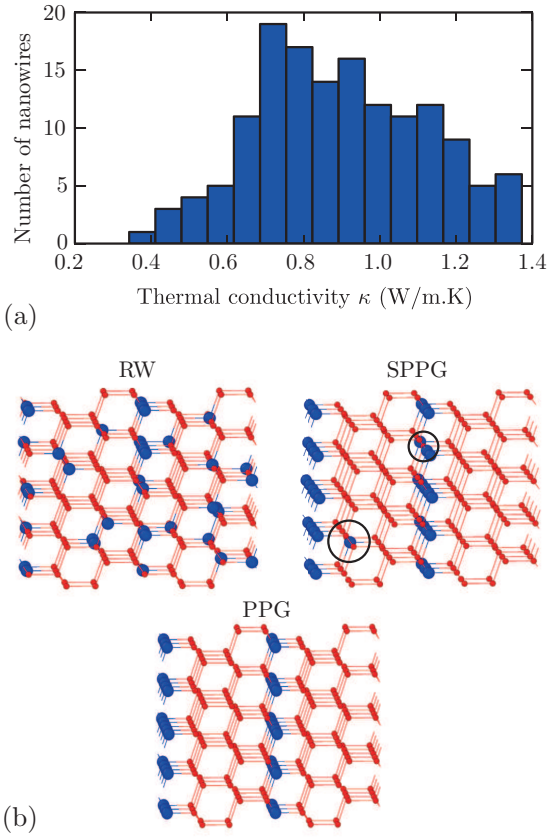

FIG. 6. (Color) (a) Nanowire data set of random Si-Ge configurations plotted in a histogram with the number of elements versus their thermal conductivity κ. (b) Each type of wire considered in the fitting and predictions of thermal conductivity: wires with random Si-Ge configurations (RW), the pure planes of Ge wire (PPG), and wires that are similar to the PPG wire (SPPG). Red/blue atoms are Si/Ge. For the SPPG wire, black circles show where two random atoms, one atom from one of the pure Ge planes, and one atom from the non-plane region, were substituted to perturb the PPG wire.

the data set, the ATAT representation of the nanowire was parsed by ATAT-GLM providing the nanowire correlation functions as output.

V. RESULTS

In Fig. 6(a) we show a histogram of $\mathcal{D}$ containing 145 wires each with a random Si-Ge configuration. The concentration of Ge was restricted to the range 3 % to 22 % as in Ref. 3, which could be due to, e.g., weight constraints, and the wires had an average thermal conductivity of 0.90 W/m.K compared to 1.1 W/m.K in Ref. 3. We will refer to a wire from $\mathcal{D}$ as a random wire (RW). Please see the top left of Fig. 6(b) for an illustration of this type of wire. For the purpose of learning $\boldsymbol{\beta}$ we split the RW data set into a training set (RW train), of 140 wires, and a test set (RW test) containing 5 wires. The ECI were learned on RW train, using least squares with repeated random sub-sampling validation using 10 splits,[43] and then initial predictions were made on RW test. The predictions on RW test serves as a preliminary test of the CE-GLM in the following sense. If the prediction errors are much larger than the training errors at this stage, the CE-GLM is not capturing the underlying physics and we should not trust further predictions. In Fig. 7(a) we first show the least squares fit on RW train. The average *root-mean-square* (RMS) training error of all random splits was 0.15 W/m.K.

In Fig. 7(b) we show the initial test of the RW train fit on RW test, and as can be seen, the predictions made on RW test have errors comparable to those on RW train. Indeed, the test error is 0.12 W/m.K. Note that the errors in general are relatively large. This was the same observation made in Ref. 3, and is due to inherent sizable noise in equilibrium MD simulations.

The purpose of this work is to show whether the CE-GLM method would have lead to the same conclusions as in Ref. 3. To this end, a data set of 18 nanowires was created, each wire being a perturbed configuration from the PPG wire. We refer to these as the similar-to-PPG (SPPG) wires. As an example, one perturbation is

obtained by substituting a random Ge atom from one of the pure Ge planes in the PPG wire with a random Si atom in the non-planes region of the wire. Each SPPG wire has its own number of atoms, ranging from 1 to 8, randomly substituted in this way. See Fig. 6(b) for an example SPPG wire with 1 atom substituted.

Notice that, importantly, the RW test wires, the SPPG wires, and the single PPG wire, are not used in training the least squares model. Only RW train is employed for this task. If the CE-GLM is useful here, we will predict that a large majority (if not all) of the SPPG wires have lower thermal conductivity than the RWs. In Fig. 7(b) we see that this is also the case. Furthermore, the RMS prediction error is 0.25 W/m.K, which is similar in magnitude to the RW data sets. Also, a larger error for the, more or less, ordered SPPG wires is to be expected since the predictions are made based on a fit to the *random* data set. In principle, this should not matter since we assume that the thermal conductivity is captured well in all of configuration space. However, in practice, since we only observe a limited amount of data, it will typically be the case that we enjoy an increased predictive capability in regions closer to the structures employed in the training process.

As a final additional test, we see whether CE-GLM predicts the PPG wire to have overall lowest thermal conductivity of all wires considered above. This is an important test. If it (almost) has the lowest conductivity, we would have been able to provide the same conclusion as Ref. 3 if coupling the cluster expanded thermal conductivity with an optimization routine such as the genetic algorithm or an adaptive sequential Monte Carlo method.[44] Remarkably, Fig. 7(b) shows that we do indeed make this important prediction. Note that, whether this is indeed the global minimum configuration is out the scope of this work to determine. These results provide evidence that the CE-GLM does capture the thermal conductivity of Si-Ge nanowires to, at least, the same degree as the MCE, and is therefore a useful method to apply in this problem.

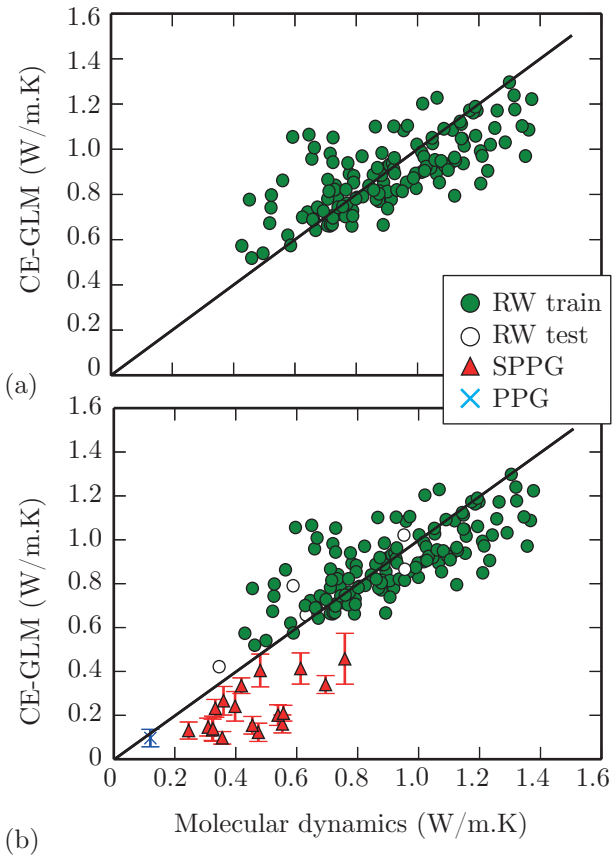

FIG. 7. (Color) Predictions of thermal conductivity from the cluster expansion ghost lattice method (CE-GLM) versus the results from molecular dynamics. In (a) the least squares fit on RW train is shown. Only one of the 10 train/test splits are shown (green/white circles). In (b) the CE-GLM is tested on three different test sets: RW test, the SPPG wires, and the PPG wire. Note that in (b), RW train (green circles) is the same as in (a). Error bars on the SPPG wires and the PPG wire stem from 10 different random splits of the RW data into train and test. For more information on the various nanowire data sets see, e.g., Fig. 6(b).

VI. CONCLUSION

In this work we introduced a modified cluster expansion method, called the cluster expansion ghost lattice method (CE-GLM), which is distinct to the regular ternary cluster expansion with vacuum atom types. The CE-GLM uses the bulk clusters and bulk symmetries for any geometry. The ghost site—zeroing cluster functions of clusters containing it—turns any structure into an effective bulk geometry and makes the CE-GLM computationally efficient when implemented in existing cluster expansion software packages—no matter the shape and/or size of the structure in question when compared to an equivalent bulk system. We discussed how to implement the method in the cluster expansion software package ATAT.

To test the usefulness of the CE-GLM the method was employed to predict the thermal conductivity of Si-Ge nanowires. When comparing to a similar work carried out recently we find great agreement: the lowest thermal-conductivity-nanowire configuration is that of an Si wire with planes of pure Ge spaced 1 nm apart. This establishes that the CE-GLM can be useful for cluster expanding non-bulk low-symmetry systems with minor modification to existing cluster expansion software.

The main drawback of the CE-GLM is that it does not provide a way to systematically check whether increasingly precise surface terms matter greatly in the problem at hand or not. In future work it will be interesting to employ the method on other even more exotic geometries and preferably compare to other cluster grouping methods.

ACKNOWLEDGMENTS

The work of NJZ at the University of Warwick is supported by the Royal Society through a Wolfson Research Merit Award and from EPSRC (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

[1] R. H. Taylor, S. Curtarolo, and G. L. Hart, Physical Review B **81**, 024112 (2010).

[2] J. Kristensen, I. Bilionis, and N. Zabaras, Phys. Rev. B **87**, 174112 (2013), URL http://link.aps.org/doi/10.1103/PhysRevB.87.174112.

[3] M. Chan, J. Reed, D. Donadio, T. Mueller, Y. Meng, G. Galli, and G. Ceder, Physical Review B **81**, 174303 (2010).

[4] Y. Kumagai, Y. Soda, F. Oba, A. Seko, and I. Tanaka, Phys. Rev. B **85**, 033203 (2012), URL http://link.aps.org/doi/10.1103/PhysRevB.85.033203.

[5] G. Kresse and J. Hafner, Physical Review B **47**, 558 (1993).

[6] S. Plimpton, Journal of computational physics **117**, 1 (1995).

[7] J. M. Sanchez, Phys. Rev. B **81**, 224202 (2010), URL http://link.aps.org/doi/10.1103/PhysRevB.81.224202.

[8] A. van de Walle, Calphad **33**, 266 (2009).

[9] A. van de Walle, M. Asta, and G. Ceder, Calphad **26**, 539 (2002).

[10] A. van de Walle and M. Asta, Model. Simul. Mater. Sc. **10**, 521 (2002).

[11] A. van de Walle, JOM - J. Min. Met. Mat. S. **65**, 1523 (2013).

[12] A. van de Walle, Nat. Mater. **7**, 455 (2008).

[13] E. Cockayne and A. van de Walle, Phys. Rev. B **81**, 012104 (2010).

[14] B. Han, A. Van der Ven, G. Ceder, and B.-J. Hwang, Physical Review B **72**, 205409 (2005).

[15] J. M. Sanchez, F. Ducastelle, and D. Gratias, Physica A: Statistical Mechanics and its Applications **128**, 334 (1984).

[16] A. van de Walle, Calphad **33**, 266 (2009).

[17] J. Kristensen and N. J. Zabaras, Computer Physics Communications **185**, 2885 (2014).

[18] A. Van de Walle, M. Asta, and G. Ceder, Calphad **26**, 539 (2002).

[19] G. D. Garbulsky and G. Ceder, Phys. Rev. B **51**, 67 (1995), URL http://link.aps.org/doi/10.1103/PhysRevB.51.67.

[20] L. J. Nelson, F. Zhou, G. L. Hart, and V. Ozolins, arXiv preprint arXiv:1208.0030 (2012).

[21] G. Ceder, G. D. Garbulsky, D. Avis, and K. Fukuda, Phys. Rev. B **49**, 1 (1994), URL http://link.aps.org/doi/10.1103/PhysRevB.49.1.

[22] N. A. Zarkevich and D. D. Johnson, Phys. Rev. Lett. **92**, 255702 (2004), URL http://link.aps.org/doi/10.1103/PhysRevLett.92.255702.

[23] G. L. Hart, V. Blum, M. J. Walorski, and A. Zunger, Nature materials **4**, 391 (2005).

[24] D. Lerch, O. Wieckhorst, G. L. W. Hart, R. W. Forcade, and S. Müller, Modelling and Simulation in Materials Science and Engineering **17**, 055003 (2009), ISSN 0965-0393, URL http://stacks.iop.org/0965-0393/17/i=5/a=055003?key=crossref.773502c73bbcd3147a3f9fa4d54d3b49.

[25] R. Drautz, H. Reichert, M. Fähnle, H. Dosch, and J. Sanchez, Physical review letters **1**, 1 (2001), URL http://journals.aps.org/prl/abstract/10.1103/PhysRevLett.87.236102.

[26] M. Amato, M. Palummo, R. Rurali, and S. Ossicini, Chemical Reviews **114**, 1371 (2014), URL http://pubs.acs.org/doi/abs/10.1021/cr400261y.

[27] M. Shelley and A. A. Mostofi, EPL (Europhysics Letters) **94**, 67001 (2011).

[28] D. Donadio and G. Galli, Physical review letters **102**, 195901 (2009).

[29] T. T. Vo, A. J. Williamson, V. Lordi, and G. Galli, Nano letters **8**, 1111 (2008).

[30] B. Qiu, H. Bao, G. Zhang, Y. Wu, and X. Ruan, Computational Materials Science **53**, 278 (2012).

[31] C. C. Yang and S. Li, ChemPhysChem **12**, 3614 (2011).

[32] D. M. Rowe, *Materials, Preparation, and Characterization in Thermoelectrics*, vol. 1 (CRC press, 2012).

[33] T. Harman, P. Taylor, M. Walsh, and B. LaForge, Science **297**, 2229 (2002).

[34] A. Majumdar, Science **303**, 777 (2004).

[35] S. G. Volz and G. Chen, Applied Physics Letters **75**, 2056 (1999).

[36] C. Glassbrenner and G. A. Slack, Physical Review **134**, A1058 (1964).

[37] B. Abeles, Phys. Rev. **131**, 1906 (1963), URL http://link.aps.org/doi/10.1103/PhysRev.131.1906.

[38] A. Stukowski, Modelling and Simulation in Materials Science and Engineering **18**, 015012 (2010), URL http://ovito.org/.

[39] D. Frenkel and B. Smit, *Understanding molecular simulation: from algorithms to applications*, vol. 1 (Academic press, 2001).

[40] P. K. Schelling, S. R. Phillpot, and P. Keblinski, Physical Review B **65**, 144306 (2002).

[41] J. Tersoff, Phys. Rev. B **39**, 5566 (1989).

[42] A. J. McGaughey and M. Kaviany, Advances in Heat Transfer **39**, 169 (2006).

[43] S. M. Weiss and C. A. Kulikowski, *Computer Systems That Learn: Classification And Prediction Methods From Statistics, Neural Nets, Machine Learning And Exp* (Morgan Kaufmann, 1990).

[44] P. Del Moral, A. Doucet, and A. Jasra, Statistics and Computing **22**, 1009 (2012).