Journal of Computational Physics 323 (2016) 17–44

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

Quantifying uncertainties in first-principles alloy thermodynamics using cluster expansions

Manuel Aldegunde [a,b], Nicholas Zabaras [a,b,*], Jesper Kristensen [c]

[a] *Warwick Centre for Predictive Modelling, University of Warwick, Coventry CV4 7AL, United Kingdom*
[b] *Department of Aerospace and Mechanical Engineering, University of Notre Dame, 365 Fitzpatrick Hall, Notre Dame, IN 46556, USA*
[c] *School of Applied and Engineering Physics, 271 Clark Hall, Cornell University, Ithaca, NY 14853-3501, USA*

ARTICLE INFO

Article history:
Received 1 March 2016
Received in revised form 23 June 2016
Accepted 6 July 2016
Available online 20 July 2016

Keywords:
Bayesian
Cluster expansion
Uncertainty quantification
Predictive modeling
Alloy modeling
Ground state line
Phase diagram
Phase transition

ABSTRACT

The cluster expansion is a popular surrogate model for alloy modeling to avoid costly quantum mechanical simulations. As its practical implementations require approximations, its use trades efficiency for accuracy. Furthermore, the coefficients of the model need to be determined from some known data set (training set). These two sources of error, if not quantified, decrease the confidence we can put in the results obtained from the surrogate model. This paper presents a framework for the determination of the cluster expansion coefficients using a Bayesian approach, which allows for the quantification of uncertainties in the predictions. In particular, a relevance vector machine is used to automatically select the most relevant terms of the model while retaining an analytical expression for the predictive distribution. This methodology is applied to two binary alloys, SiGe and MgLi, including the temperature dependence in their effective cluster interactions. The resulting cluster expansions are used to calculate the uncertainty in several thermodynamic quantities: ground state line, including the uncertainty in which structures are thermodynamically stable at 0 K, phase diagrams and phase transitions. The uncertainty in the ground state line is found to be of the order of meV/atom, showing that the cluster expansion is reliable to *ab initio* level accuracy even with limited data. We found that the uncertainty in the predicted phase transition temperature increases when including the temperature dependence of the effective cluster interactions. Also, the use of the bond stiffness versus bond length approximation to calculate temperature dependent properties from a reduced set of alloy configurations showed similar uncertainty to the approach where all training configurations are considered but at a much reduced computational cost.

1. Introduction

New technological advances require materials with certain properties that may not be found in known material systems. However, even though these properties may not be found in pure materials, it is possible to alloy different material systems to tune the properties of interest [1–3]. Applications in different fields have used this approach to look for desired electronic

[*] Corresponding author at: Department of Aerospace and Mechanical Engineering, University of Notre Dame, 365 Fitzpatrick Hall, Notre Dame, IN 46556, USA.

E-mail addresses: M.A.Aldegunde-Rodriguez@warwick.ac.uk (M. Aldegunde), nzabaras@gmail.com (N. Zabaras), jespertoftkristensen@gmail.com (J. Kristensen).

URLs: http://www.warwick.ac.uk/wcpm/, http://www.zabaras.com (N. Zabaras).

http://dx.doi.org/10.1016/j.jcp.2016.07.016
0021-9991/© 2016 Elsevier Inc. All rights reserved.

and optoelectronic [2,4–7], thermoelectric [8–10] or elastic [11–13] properties, to name a few. Random alloys have been frequently used for this purpose, and in this case the properties are usually well approximated by interpolation between the pure elements, linear or with a bowing parameter [14,15,6]. However, the potential of going beyond this to design specific alloy patterns opens the possibility of obtaining new properties beyond this simple interpolation. For example, a direct gap semiconductor could be found by alloying Si and Ge, which are both indirect gap semiconductors with conduction band minimum in different points of the Brillouin zone [16–19]. This search is not easy as the number of possible alloys grows exponentially with the number of atoms in the system. For a simple binary alloy on a crystal with N sites, the number of possible configurations is 2^N. Most implementations of the *de facto* standard for purely quantum mechanical simulations of solids, *density functional theory* (DFT), scale as $\mathcal{O}(N^3)$, so exhaustive explorations of the configurational space soon become too costly with the growth of the supercell on two fronts: exponentially growing number of configurations and cubic growth in the cost of each of the configurations. Even though some success has been achieved using *ab initio* structure search [20,12,21], the use of surrogate models appears as the most feasible way for the explorations of the configurational space of alloys [22,23,15]. The most important surrogate model in this field is the *cluster expansion* (CE) [24–27], which has been widely used in a number of different applications [22,28,23,29,15,30]. It decomposes a function $f(\cdot)$ of a configuration $\boldsymbol{\sigma}$ in contributions from different clusters of atoms in the same way as a Fourier series decomposes a periodic signal into components of different frequencies, and in the same way a Fourier series is exact if all terms are included, *the CE is exact if all clusters are included*. However, due to practical considerations, only a truncated expansion is possible.

Despite the great attraction from using this surrogate model due to its simplicity and computational efficiency, the truncation of the series means that there is an error in representing any function of the configuration. Furthermore, the coefficients of the model, $\boldsymbol{\gamma}$, need to be estimated from known pairs $\mathcal{D} = \{(\boldsymbol{\sigma}_i, f(\boldsymbol{\sigma}_i))\}$, which adds an extra layer of epistemic uncertainty to the predictions of the model. We will denote this approximation depending on parameters $\boldsymbol{\gamma}$ as $f(\cdot;\boldsymbol{\gamma})$. All this means that any prediction made with the surrogate model will have an error, and it is important to know how much it is to know the confidence we can have in that prediction. Only in this case we can have a truly reliable surrogate that can replace the DFT simulations giving us a *prediction and a level of confidence on it.*

In this work, we adopt a Bayesian perspective to uncertainty quantification. In this context, we interpret probability as a degree of belief [31,32], not as a frequency of the outcome of some experiment. In this context, it is only natural to assign *prior* beliefs which encode our knowledge of a certain process before making any observation. However, through Bayes' theorem, we can update our knowledge with available observations, obtaining a *posterior* probability [31,32] for the process. Therefore, uncertainty quantification in a Bayesian setting starts by assigning a prior probability distribution to the model parameters $\boldsymbol{\gamma}$ which we wish to determine, $p(\boldsymbol{\gamma})$. To improve on this, we will use the observations $\mathcal{D}_N$, which can be obtained either computationally, experimentally or a combination of both. Before we update our beliefs about $\boldsymbol{\gamma}$, we need to encode the compatibility of our model with the data $\mathcal{D}_N$. This is done with the *likelihood* $\mathcal{L}(\mathcal{D}_N \mid \boldsymbol{\gamma})$, which quantifies how likely it is to obtain the data $\mathcal{D}_N$ using the model parameters $\boldsymbol{\gamma}$. Now that we have new information from the data as well as a measure of their compatibility with our model, we can update our belief (probability) about $\boldsymbol{\gamma}$ using Bayes' theorem, $p(\boldsymbol{\gamma} \mid \mathcal{D}_N) \propto p(\boldsymbol{\gamma})\mathcal{L}(\mathcal{D}_N \mid \boldsymbol{\gamma})$, obtaining this way the posterior probability distribution of the parameters, $p(\boldsymbol{\gamma} \mid \mathcal{D}_N)$. Finally, this posterior distribution can be used for the determination of uncertainty of any quantity of interest, QI, which can be obtained from our model with parameters $\boldsymbol{\gamma}$ through a functional $I[f(\boldsymbol{\sigma};\boldsymbol{\gamma})]$. This is done by integrating over all possible model parameters, and in this way we obtain what we call the *predictive distribution*, which contains the information of the uncertainty on the predictions made for I,

$$p(I) = \int p(I \mid \boldsymbol{\gamma}) p(\boldsymbol{\gamma} \mid \mathcal{D}_N)\, d\boldsymbol{\gamma}. \tag{1}$$

If we assume that for a given set of model parameters our functional can only give one value of the QI, then this simplifies to

$$p(I) = \int \delta\left(I[f(\boldsymbol{\sigma};\boldsymbol{\gamma})] - I\right) p(\boldsymbol{\gamma} \mid \mathcal{D}_N)\, d\boldsymbol{\gamma}, \tag{2}$$

where $\delta(\cdot)$ is Dirac's δ-function.

Previous works employing a Bayesian framework for uncertainty quantification in the CE used a Laplace prior for the expansion coefficients and a right-truncated Poisson prior for the number of clusters to induce sparsity in the number of clusters for the expansion [33]. Despite the generality of this approach, the posterior distribution cannot be sampled analytically and a reversible jump Markov Chain Monte Carlo (RJ-MCMC) [34] had to be used to obtain statistics on the predictions [33]. There are other works based on a Bayesian framework, but they did not exploit the resultant uncertainty. For example, in Ref. [35] the authors only kept the maximum of the posterior distribution for the estimation of the coefficients, and in Ref. [36] the uncertainty in the ECI was not propagated to any predicted physical quantity.

In this work, we develop a Bayesian framework that, while keeping a posterior distribution which can be sampled analytically, also enforces sparsity in the model parameters and provides quantification of the uncertainty in the results. To achieve this, we have used the *relevance vector machine* (RVM) [37], a hierarchical Bayesian model that enforces sparsity in the posterior distribution and therefore provides model selection. However, unlike the approach in Ref. [33], our model does not include uncertainty in the chosen basis set, and based on the data just selects one. In Section 2, we describe the CE surrogate model in more detail, and in Section 3 we detail the Bayesian linear regression framework used to obtain its

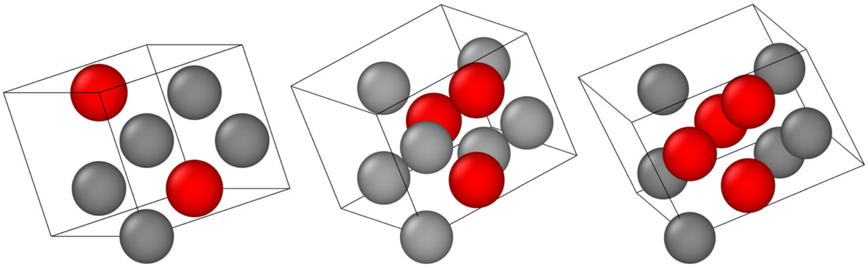


Fig. 1. Examples of (left) 2-, (center) 3- and (right) 4-point clusters in a fcc lattice. Red spheres represent the atoms belonging to the cluster. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

coefficients as random variables. We also discuss how from these parameters we can obtain a predictive distribution for the modeled function and also the use of a RVM to allow for automatic model determination. Section 4 describes the training of the model and details on the DFT simulations used to obtain the necessary data. This is followed, in Section 5, by a discussion on how to incorporate temperature dependence on the CE adding electronic and vibrational free energies and using an approximate model for the phonon calculations which induces additional uncertainty. In the next two sections, we describe numerical results on thermodynamic properties using SiGe and MgLi as test systems. Section 6 describes the procedure to obtain the ground state line (GSL) of an alloy including uncertainty in the energies and also in the structures which belong to it. Section 7 describes the algorithms used to obtain phase diagrams with uncertainty information, both following the phase equilibrium boundary and explicitly simulating the alloy for decreasing temperature using a Sequential Monte Carlo approach including temperature dependence in the CE coefficients. Finally, we end up summarizing the main contributions of this work as well as the numerical results.

2. Cluster expansion

The cluster expansion [24,27] is the most widely used surrogate model in alloy modeling to quickly compute physical properties which are a function of the configuration. In what follows, we present this model using a binary system for simplicity, even though it can be readily extended to systems with more than two components.

Consider a binary alloy on a fixed lattice with components A and B, A_xB_{1-x}, where x is the proportion of component A. Each site of the lattice can be occupied by either A or B, and it is therefore characterized by a *spin* number σ_s with possible values of -1 (A occupancy) and $+1$ (B occupancy). For a crystal with N sites, a vector containing the corresponding value for each of the sites, $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \ldots, \sigma_N\}$ completely defines a configuration of the system. There are 2^N possible configurations, which define the configuration space $\boldsymbol{\Sigma} = \{\boldsymbol{\sigma}\}$. We define a *cluster* as any of the possible groupings of the atoms in the system (e.g. isolated atom, nearest neighbor pairs, etc.). Fig. 1 shows an example of 2-, 3-, and 4-point clusters in an fcc lattice.

The idea of the CE is to construct a set of basis functions for the configurational space with N sites as the direct product of the basis functions for $N = 1$, $\{\varphi_0, \varphi(\sigma)\}$, with $\varphi_0 = 1$ corresponding to the empty cluster and $\varphi(\sigma) = \sigma$. This is a complete and orthonormal basis with the scalar product defined as [27]

$$\langle f, g\rangle = \sum_{\sigma=\pm 1} \frac{f(\sigma)g(\sigma)}{2}. \tag{3}$$

For a N site system, the basis functions will be the direct product of the basis for $N = 1$,

$$\Gamma_\alpha(\boldsymbol{\sigma}_\alpha) = \prod_{p\in\alpha} \varphi(\sigma_p), \tag{4}$$

where the product is over all sites in cluster α and α is one of the 2^N possible clusters in the lattice with N sites, including the empty cluster for which $\Gamma_0(\boldsymbol{\sigma}) = 1$. It is possible to prove that these basis functions form a complete orthonormal basis for the configurational space with N sites [27]. Each basis function, $\Gamma_\alpha(\boldsymbol{\sigma})$, is associated with each of the clusters, and each basis function will have an associated coefficient when expanding any function f of the configuration $\boldsymbol{\sigma}$ given by the scalar product $\langle f(\boldsymbol{\sigma}), \Gamma_\alpha(\boldsymbol{\sigma})\rangle$,

$$f(\boldsymbol{\sigma}) = \sum_\alpha \langle f(\boldsymbol{\sigma}), \Gamma_\alpha(\boldsymbol{\sigma})\rangle\, \Gamma_\alpha(\boldsymbol{\sigma}_\alpha). \tag{5}$$

Symmetry considerations in periodic systems can greatly reduce the number of such independent coefficients. Each cluster α can be characterized by three components: its type η, specified, for example, by shape and size, its position in the lattice π, specified, for example, as the primitive cell containing the center of mass of the cluster, and its “orientation” ω, which is determined by one of the point group symmetries of the lattice. For each cluster $\alpha = (\eta, \pi, \omega)$, all the clusters symmetrically equivalent to it under a space group operation of the underlying lattice form its *orbit*, which we will denote α_η. Therefore, the members of α_η will consist of clusters of the form $\alpha' = (\eta, \pi', \omega')$, i.e., with the same shape but possibly different position and orientation. Since all these clusters have equivalent environments owing to the periodic nature of the system, their corresponding expansion coefficients must be the same [27]. By averaging over the orbits of all inequivalent clusters, we obtain another basis for the cluster expansion,

$$\phi_\eta(\boldsymbol{\sigma}) = \frac{1}{N_\pi}\frac{1}{\omega_\eta}\sum_{\pi=1}^{N_\pi}\sum_{\omega=1}^{\omega_\eta}\Gamma_{\alpha_\eta}(\boldsymbol{\sigma}_{\alpha_\eta}) = \langle \Gamma_{\alpha'}(\boldsymbol{\sigma})\rangle_{\alpha_\eta}, \tag{6}$$

where ω_η is the number of clusters symmetrically equivalent under point group operations for a given η and π, N_π the number of primitive unit cells, and $\langle\cdot\rangle_{\alpha_\eta}$ represents the average over the orbit of a cluster characterized by shape η. Using this basis, we can expand a function of the configuration as:

$$f(\boldsymbol{\sigma} \mid \boldsymbol{\gamma}) = \sum_{i=0}^{M}\gamma_i\phi_i(\boldsymbol{\sigma}), \tag{7}$$

where $\boldsymbol{\gamma} = \{\gamma_i\}$ are the *effective cluster interactions* (ECI) and M is the number of terms in the expansion. In this work, we have included the multiplicity from the space group operations in the ECI γ_i.

It can be shown that the cluster expansion is a multidimensional discrete Fourier transform, so if all clusters are included, then the expansion is exact [27]. However, it is usually necessary to truncate the number of clusters included [26]. This can be done in two different ways: by fixing the maximum number of crystal sites present in any cluster as well as its maximum spatial extent (largest distance between any two crystal sites in the cluster). Since the number of points in the cluster are related to the frequency [27], it is natural to include first the clusters with a small number of points as it would be done in the approximation of a Fourier series.

Remark 1. There are extensions of the CE formalism described in this section, such as the variable basis cluster expansion [27], the mixed-basis cluster expansion [38] or the tensorial cluster expansion [39]. However, as all of them are based on a linear model, the methodology that we will apply in this paper to calculate the uncertainty in the ECI and for predictive modeling of thermodynamic properties can easily be extended to these formulations.

3. The relevance vector machine

The cluster expansion model as formulated in Eq. (7) is a linear model in the basis functions $\phi_i(\boldsymbol{\sigma})$. To determine the coefficients $\boldsymbol{\gamma}$ from the available data, we use a Bayesian linear regression model [31,32], which treats the regression coefficients (ECI) as random variables instead of point estimates.

In the problem of regression we have two spaces, input and output. In our case, the input space is the configurational space $\boldsymbol{\Sigma}$ and the output space depends on the quantity we want to fit, for example, it is the real line $\mathbb{R}$ for energies. We will denote elements of this output space by t and a set of them by $\mathbf{t}$. We will assume that for any input point $\boldsymbol{\sigma}_i$ the observed t_i follows on average our linear model and has an additional error term ε. This quantity represents the model accuracy as it provides a measure of the deviation between our model and the experimental/exact results. Therefore, for a single observation t_i,

$$t_i = \boldsymbol{\gamma}^T\boldsymbol{\phi}[\boldsymbol{\sigma}_i] + \varepsilon_i, \tag{8}$$

where $\boldsymbol{\phi}[\boldsymbol{\sigma}_i]$ is a $M \times 1$ matrix with entry j representing the basis function ϕ_j evaluated at $\boldsymbol{\sigma}_i$. ε_i is an error term and we will assume it to be Gaussian with the same precision for all data points, $\beta = 1/v = 1/\sigma^2$, where v is the variance and σ the standard deviation. Therefore, the probability of obtaining a particular value for the observation of a material system with configuration $\boldsymbol{\sigma}$ will follow a Gaussian distribution with mean $\boldsymbol{\gamma}^T\boldsymbol{\phi}[\boldsymbol{\sigma}]$:

$$t \sim \mathcal{N}(t \mid \boldsymbol{\gamma}^T\boldsymbol{\phi}[\boldsymbol{\sigma}], \beta^{-1}). \tag{9}$$

This model defines the likelihood function $\mathcal{L}$, which depends on the model parameters $\boldsymbol{\gamma}$ and β and on the experimentally observed data $\mathcal{D}_N$, which consists of N pairs of observations $t_{t,i}$ at given configurations $\boldsymbol{\sigma}_{t,i}$. The likelihood gives a measure of how likely it is to obtain the observed data $\mathcal{D}_N = \{(\boldsymbol{\sigma}_{t,i}, t_{t,i})\}_{i=0}^{N-1}$ given the assumed model. In our case, it will be

$$\mathcal{L}(\mathcal{D}_N \mid \boldsymbol{\gamma}, \beta) = \prod_{(\boldsymbol{\sigma}_{t,i}, t_{t,i})\in\mathcal{D}_N}\mathcal{N}(t_i \mid \boldsymbol{\gamma}^T\boldsymbol{\phi}[\boldsymbol{\sigma}_{t,i}], \beta^{-1}) = \mathcal{N}(\mathbf{t}_t \mid \boldsymbol{\Phi}\boldsymbol{\gamma}, \beta^{-1}\mathbf{I}), \tag{10}$$

where $\mathbf{t}_t = (t_{t,0} \ldots t_{t,N-1})^T$ and $\boldsymbol{\Phi}$ is the *design matrix*, which is defined as a matrix whose entries are $\boldsymbol{\Phi}_{ij} = \phi_j[\boldsymbol{\sigma}_{t,i}]$, i.e., a matrix where each row contains all the basis functions evaluated for a given training system, or, equivalently, where each column contains a given basis function evaluated for all the training systems,

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0[\boldsymbol{\sigma}_{t,0}] & \cdots & \phi_{M-1}[\boldsymbol{\sigma}_{t,0}] \\ \phi_0[\boldsymbol{\sigma}_{t,1}] & \cdots & \phi_{M-1}[\boldsymbol{\sigma}_{t,1}] \\ \vdots & \ddots & \vdots \\ \phi_0[\boldsymbol{\sigma}_{t,N-1}] & \cdots & \phi_{M-1}[\boldsymbol{\sigma}_{t,N-1}] \end{pmatrix} = \begin{pmatrix} \boldsymbol{\phi}[\boldsymbol{\sigma}_{t,0}]^T \\ \boldsymbol{\phi}[\boldsymbol{\sigma}_{t,1}]^T \\ \vdots \\ \boldsymbol{\phi}[\boldsymbol{\sigma}_{t,N-1}]^T \end{pmatrix}, \tag{11}$$

where $\boldsymbol{\phi}[\boldsymbol{\sigma}]$ is the column vector with the basis functions evaluated at a given configuration $\boldsymbol{\sigma}$, $\phi_i[\boldsymbol{\sigma}]$, as its entries, $\boldsymbol{\phi}[\boldsymbol{\sigma}] = (\phi_0[\boldsymbol{\sigma}]\, \phi_1[\boldsymbol{\sigma}] \cdots \phi_{M-1}[\boldsymbol{\sigma}])^T$.

Since we do not know *a priori* which of the basis functions are necessary or relevant to describe our objective function, we use a relevance vector machine, which through the use of a sparsity inducing prior provides automatic model selection by discarding the basis functions which are not required to describe the data. In this case, the prior distribution is described by the following hierarchical model [37],

$$p(\boldsymbol{\gamma} \mid \boldsymbol{\alpha}) = \prod_i \mathcal{N}(\boldsymbol{\gamma} \mid \alpha_i^{-1}) = \mathcal{N}(\boldsymbol{\gamma} \mid 0, \mathrm{diag}(\alpha_i^{-1})), \tag{12}$$

$$p(\boldsymbol{\alpha} \mid c_0, d_0) = \prod_i \mathcal{G}(\alpha_i \mid c_0, d_0), \tag{13}$$

$$p(\beta \mid c_0, d_0) = \mathcal{G}(\beta \mid c_0, d_0), \tag{14}$$

where $\mathcal{G}(x \mid a, b)$ is a gamma distribution on x with shape parameter a and rate parameter b. We set $a_0 = b_0 = c_0 = d_0 = 0$ to obtain uniform hyperpriors (over a logarithmic scale), which has the added benefit of resulting in the scale invariance of the predictions [37].

Given the likelihood and the priors, we can obtain the posterior probability distribution of the parameters given the data using Bayes' theorem [31],

$$p(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \beta \mid \mathcal{D}_N) = \frac{\mathcal{L}(\mathcal{D}_N \mid \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \beta)}{\int \mathcal{L}(\mathcal{D}_N \mid \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \beta)\, d\boldsymbol{\gamma}\, d\beta\, d\boldsymbol{\alpha}}. \tag{15}$$

The normalizing integral cannot be carried out analytically, so we seek an approximation. To do this, we decompose the posterior over the parameters as $p(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \beta \mid \mathcal{D}_N) = p(\boldsymbol{\gamma} \mid \mathcal{D}_N, \boldsymbol{\alpha}, \beta) p(\boldsymbol{\alpha}, \beta \mid \mathcal{D}_N)$. This way we can obtain the posterior over the weights conditioned on the hyperparameters $\boldsymbol{\alpha}$ and β as [37] (see Appendix A for details)

$$p(\boldsymbol{\gamma} \mid \mathcal{D}_N, \boldsymbol{\alpha}, \beta) = \frac{\mathcal{L}(\mathcal{D}_N \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) p(\boldsymbol{\gamma} \mid \boldsymbol{\alpha})}{\int \mathcal{L}(\mathcal{D}_N \mid \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(\boldsymbol{\gamma} \mid \boldsymbol{\alpha})\, d\boldsymbol{\gamma}} = \mathcal{N}(\boldsymbol{\gamma} \mid \boldsymbol{\mu}, \mathbf{S}_N), \tag{16}$$

where

$$\mathbf{S}_N^{-1} = \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi} + \mathbf{A}, \tag{17}$$

$$\boldsymbol{\mu} = \beta \mathbf{S}_N \boldsymbol{\Phi}^T \mathbf{t}_t, \tag{18}$$

and we have defined $\mathbf{A} = \mathrm{diag}(\alpha_0, \ldots, \alpha_M)$. To obtain an analytical expression for the posterior distribution, we approximate $p(\boldsymbol{\alpha}, \beta \mid \mathcal{D}_N)$ by a delta function at its mode, $p(\boldsymbol{\alpha}, \beta \mid \mathcal{D}_N) \approx \delta(\boldsymbol{\alpha}_{MP}, \beta_{MP})$. Using this approximation, learning the RVM becomes the search for the mode of the hyperparameter posterior distribution [37], which is the maximization of $p(\boldsymbol{\alpha}, \beta \mid \mathcal{D}_N) = p(\mathcal{D}_N \mid \boldsymbol{\alpha}, \beta) p(\boldsymbol{\alpha}) p(\beta)$ with respect to $\boldsymbol{\alpha}$ and β. With the adopted approximation of uniform hyperpriors this is equivalent to the maximization of $p(\mathcal{D}_N \mid \boldsymbol{\alpha}, \beta)$, which is the marginal likelihood given by:

$$p(\mathcal{D}_N \mid \boldsymbol{\alpha}, \beta) = \int p(\mathcal{D}_N \mid \boldsymbol{\gamma}, \beta) p(\boldsymbol{\gamma} \mid \boldsymbol{\alpha})\, d\boldsymbol{\gamma} = \mathcal{N}(\mathbf{t}_t \mid 0, \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T). \tag{19}$$

Details of the derivation and also on its maximization with respect to $\boldsymbol{\alpha}$ and β can be found in Appendix B. $\boldsymbol{\alpha}$ and β are updated sequentially using Eqs. (B.12) and (B.20) [40,31] until convergence is achieved. In the numerical implementation of the pruning of the model basis functions, we consider a maximum value of $\alpha_{max} = 10^{16}$ as an approximation to the limit $\alpha \to \infty$. This process is detailed in Algorithm 1.

Remark 2. The use of a maximum posterior approximation for the parameters $\boldsymbol{\alpha}$ and β will lead to an underestimation of the uncertainty in the results. This restriction used in the original formulation of the RVM [37] is typical of Bayesian hierarchical models and can be easily lifted at the expense of losing the analytical expression for the posterior distribution of the parameters. The impact of such an approach to the predictive uncertainty requires further study.

Remark 3. We can understand better the reason for the sparsification if we look at the covariance $\mathbf{C}$ of the marginal likelihood in Eq. (19),

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T = \beta^{-1}\mathbf{I} + \sum_{i=0}^{M-1} \frac{1}{\alpha_i} \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^T, \tag{20}$$

Algorithm 1 Evidence approximation: Hyperparameter optimization.

```
Select convergence criterion for the inner and outer iterations: $\theta_{inner}$, $\theta_{outer}$
Select maximum number of iterations for outer and inner loops, $t_{max}$ and $j_{max}$, respectively
Select numerical threshold $\alpha_{max}$ to detect $\alpha \to \infty$
Initialize outer iteration $t = 0$
Initialize $\boldsymbol{\alpha}^0 = (\alpha_{max} \dots \alpha_{max})^T$
Initialize $\beta^0 = 0$
for $t = 0, 1, \dots, t_{max}$ do
    Obtain $\mathbf{s}$, $\mathbf{q}$ according to Eqs. (B.15) and (B.16)
    Set $\boldsymbol{\alpha}^t = \boldsymbol{\alpha}^{t-1}$, $\beta^t = \beta^{t-1}$
    for $i = 0, 1, \dots, M-1$ do
        if $q_i^2 > s_i$ then
            $\alpha_i = s_i^2/(q_i^2 - s_i)$, Eq. (B.12)
        else
            $\alpha_i = \alpha_{max}$
        end if
    end for
    Set $\boldsymbol{\alpha}^t = (\alpha_0 \dots \alpha_M)^T$
    Set $\Delta\alpha = \| \boldsymbol{\alpha}^t - \boldsymbol{\alpha}^{t-1} \|_2$
    Set $\beta_0^t = \beta^{t-1}$
    for $j = 0, 1, \dots, j_{max}$ do
        $\frac{1}{\beta_{j+1}^t} = \frac{(\mathbf{t}_t - \boldsymbol{\Phi}\boldsymbol{\mu}_j^{t-1})^T(\mathbf{t}_t - \boldsymbol{\Phi}\boldsymbol{\mu}_j^{t-1})}{N - \mathrm{trace}(\mathbf{I} - \mathbf{A}\mathbf{S}_{N,j}^{t-1})}$, Eq. (B.20).
        Update $\mathbf{S}_{N,j}^t$, $\boldsymbol{\mu}_j^t$ using Eqs. (17) and (18)
        Set $\Delta\beta = |\beta_{j+1}^t - \beta_j^t|$
        if $\Delta\beta < \theta_{inner}$ then
            break
        end if
    end for
    Set $\beta^t = \beta_{j+1}^t$
    Set $\Delta\beta = |\beta^t - \beta^{t-1}|$
    if $\Delta\alpha, \Delta\beta < \theta_{outer}$ then
        break
    end if
end for
```

where $\boldsymbol{\varphi}_i$ is the i-th column of the design matrix $\boldsymbol{\Phi}$. Equation (19) for the marginal likelihood gives the probability of obtaining the given observations $t_{t,i}$ for each of the input training points $\boldsymbol{\sigma}_{t,i}$ given the hyperparameters $\boldsymbol{\alpha}$ and β. Each of the terms in the summation in Eq. (20) represents the covariance matrix associated to each of the basis functions. The objective of the *evidence approximation* is to maximize the marginal likelihood for the data $\mathcal{D}_N$. Since the marginal likelihood is a unimodal distribution centered at the origin, its maximization at the experimental observations $\mathbf{t}_t$ will happen when the covariance is aligned with the data points. Therefore, the RVM will prune all the covariance matrices $\varphi_i\varphi_i^T$ which are not aligned with the experimental training set $\mathbf{t}_t$ by taking their α_i to infinity. More details on this behavior can be found in [37,31].

3.1. Predictive distribution

We are usually interested in obtaining a prediction with our model after carrying out the training process. Once we have a distribution for the model parameters, we can calculate a probability distribution of the output of our model, t^*, given an input system with configuration $\boldsymbol{\sigma}^*$. This is called the *predictive distribution* and it is computed by averaging the likelihood of observing t^* given $\boldsymbol{\sigma}^*$ over all possible values of the model parameters with weight defined by the posterior distribution $p(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \beta \,|\, \mathcal{D}_N)$,

$$\begin{aligned} p(t^* \,|\, \mathcal{D}_N) &\approx \int p(t^* \,|\, \boldsymbol{\gamma}, \beta) p(\boldsymbol{\gamma} \,|\, \mathcal{D}_N, \boldsymbol{\alpha}, \beta) \delta(\boldsymbol{\alpha}_{MP}, \beta_{MP}) \, d\boldsymbol{\gamma} d\boldsymbol{\alpha} d\beta \\ &= \int p(t^* \,|\, \boldsymbol{\gamma}, \beta_{MP}) p(\boldsymbol{\gamma} \,|\, \mathcal{D}_N, \boldsymbol{\alpha}_{MP}, \beta_{MP}) d\boldsymbol{\gamma} \\ &= \int \mathcal{N}(t^* \,|\, \boldsymbol{\mu}^T \boldsymbol{\phi}[\boldsymbol{\sigma}^*], \beta_{MP}^{-1}) \mathcal{N}(\boldsymbol{\gamma} \,|\, \boldsymbol{\mu}, \mathbf{S}_N) d\boldsymbol{\gamma} = \mathcal{N}(t^* \,|\, \mu^*, s^*), \end{aligned} \tag{21}$$

where

$$s^* = \beta_{MP}^{-1} + \boldsymbol{\phi}[\boldsymbol{\sigma}^*]^T \mathbf{S}_N \boldsymbol{\phi}[\boldsymbol{\sigma}^*], \tag{22}$$

$$\mu^* = \boldsymbol{\mu}^T \boldsymbol{\phi}[\boldsymbol{\sigma}^*], \tag{23}$$

and we have used Eq. (A.3). Therefore, for the assumed model, each prediction for a new system is given by a normal probability distribution and is dependent on the particular system.

Remark 4. The framework described in this section assumes a single noise parameter β which represents the model accuracy only, as we will assume that the training energies obtained from DFT are exact. This is in general not true, and they will have an error associated to numerical noise and also to the lack of knowledge of the exact exchange–correlation functional. This source of error can also be added to the model as an additional term in the precision of the likelihood in Eq. (10).

4. Model training

To study the effect of the uncertainty in the CE, we will explore the thermodynamic properties of two different alloys: SiGe and MgLi. Silicon–germanium (SiGe) is a solid–solution semiconductor with a diamond cubic structure, which is the same as the one of its components Si and Ge. SiGe alloys are commonly used in nanoelectronic [41,42], optoelectronic [43,44] or thermoelectric devices [45,46] in part because of the opportunities they present for band-gap and strain engineering, which leads to effects on fundamental properties absent in both Si and Ge. On the other hand, bcc Magnesium–Lithium alloys belong to the family of ultra-light materials [47,12,23]. Magnesium is the sixth most abundant element in the earth's mantle and its low density offers possibilities for its use in situations where weight savings are important, such as automotive or aerospace applications. However in its most stable hcp phase, Mg has some undesirable mechanical properties. Alloying it with Li stabilizes it in a bcc phase and makes it more attractive to be used in manufacturing [23].

In the study of these materials, we will be using the *Alloy Theoretic Automated Toolkit* (ATAT) [48,49]. Also, for the study of the stability of different phases of an alloy and therefore the ground state line, we are interested in the ability of the CE to reproduce the formation energy, $\Delta_f E$, of the different phases. For example, for MgLi:

$$\Delta_f E(\boldsymbol{\sigma}) = E(\boldsymbol{\sigma}) - \left[x_{Mg} E(\boldsymbol{\sigma}_{Mg}) + (1 - x_{Mg}) E(\boldsymbol{\sigma}_{Li})\right], \tag{24}$$

where x_{Mg} is the fraction of Mg for the configuration σ and $\sigma_{Li/Mg}$ are the configurations of pure Li and Mg, respectively. Therefore, this is the quantity we will use to fit the ECI to obtain a surrogate model,

$$\Delta_f E(\boldsymbol{\sigma} \mid \boldsymbol{\gamma}) = \sum_{i=0}^{M} \gamma_i \phi_i(\boldsymbol{\sigma}). \tag{25}$$

Remark 5. Since the ECI provide us with the coefficients of a linear model for a given function of the configuration, a different set is needed if we want to study other properties of the materials. For example, we may be interested on the elastic properties of a material and fit a set of ECI to its bulk modulus or in the study of electronic properties of SiGe, for which it would be interesting to obtain the ECI for different energy band-gaps or effective masses in the Γ, Δ and L valleys in the Brillouin zone.

The training energies are obtained from DFT simulations using the *Vienna Ab-initio Simulation Package* (VASP) [50,51] as in Ref. [52]. The Perdew–Burke–Ernzerhof (PBE) exchange–correlation functional [53] was used throughout. A convergence criterion of 1 meV/atom was used to choose the plane-wave energy cut-off, which was achieved around 340 eV. Spin polarization was not included. The Brillouin zone integration was done on a Γ-centered Monkhorst–Pack [54] grid with just over 7500 **k** points per reciprocal atom and scaled according to the size of each supercell being computed.

The configurations used for the fitting, $\boldsymbol{\Sigma}_t = \{\boldsymbol{\sigma}_i\}$, were generated using the *MIT Ab-initio Phase Stability* (maps) tool in ATAT. The problem of finding the optimal next structure to simulate, known as active learning [55,56], is done by the maps tool using a trade-off between minimizing the expected variance of a least-squares fit and minimizing the cost of simulating the new structure [49]. To do this, the code scans through candidate structures [57] in order of increasing computational cost, which is estimated as $\mathcal{O}(N^3)$ for DFT simulations. For each structure, the *gain* G is calculated as $G = \Delta V / C$, where ΔV is the expected variance gain and C the estimated computational cost. The search is aborted when $\Delta V_{max}/C < G_{max}$, where G_{max} is the best gain found so far and ΔV_{max} is the estimated maximum possible variance reduction [49].

Remark 6. To use a fully Bayesian framework, the active learning would be driven by a Bayesian design of experiments [58,59] to determine which structures to simulate based on the information available at each step, including the uncertainty in the current model.

One last point that needs to be specified for the training of the expansion is the truncation criterion. The number of terms included in the expansion is determined by two different parameters: maximum number of points in a given cluster and maximum size (spatial extent) of the cluster. The necessary values for convergence depend on the material being simulated. For SiGe, including only 2-point clusters already gives a good fit [33], but for MgLi, clusters with up to 5 points may be necessary [23,33]. This can be seen as well using the RVM, which removes clusters that are not relevant for the fit. For the case of SiGe, this can be seen in Fig. 2, which shows the fitted ECI for at least 2-point clusters using as a basis all 2-point clusters up to 9 nm in diameter and all 3-point clusters up to 5 nm. Of the 6 3-point clusters included, only one of them is kept and it has a small associated ECI value. This means that the 3-point clusters do not add much information to

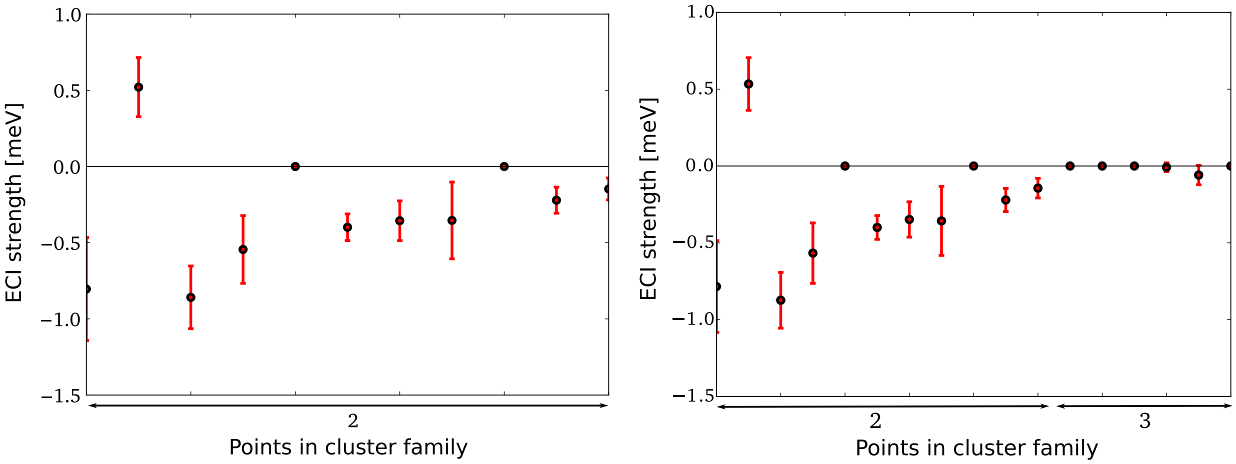


Fig. 2. ECI of the cluster expansion for the SiGe formation energy obtained using a RVM starting with clusters up to (left) 2- and (right) 3-points. Black dots represent the posterior mean for each of the coefficients and the red error bars the 95% confidence interval from Eqs. (16)–(18), respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the description of the formation energies for SiGe. Even though the posterior probability of the ECI is integrated out in the predictions, Eq. (21), it is interesting to look at their values as they can provide information about the physics of the system. Specifically, for a given type of cluster, the sign of its ECI will determine if it is favored or not, and therefore it is important to discern the physics of the model [33]. For example, for the 2-point clusters, if the corresponding ECI is negative then its associated cluster family prefers its two atoms to be of the same type ($\sigma_1 \cdot \sigma_2 = +1$). We notice that for all the ECI, most of the mass of their probability distributions has the same sign, which means that even though their impact can change for different realizations of the ECI, the physics remains the same.

Fig. 3 shows the same information for MgLi, including up to (top) 3-, (middle) 4- and (bottom) 5-point clusters. In this case, most of the strength is also in the small 2-point clusters, but the smallest 3-point cluster also has a large strength. However, most of the other 3-point clusters are suppressed by the RVM. As we increase the maximum number of points in the clusters we can see that, even though their strength remains weaker than small 2- and 3-point clusters, they are not removed by our model. Looking at the signs again, most of the clusters have a well defined ECI sign, but a few of the weaker 2-point clusters have their probability mass divided between positive and negative strength, which means that the model does not give a conclusive answer about the physics of those cluster families.

Remark 7. With the use of a truncated cluster expansion a question may arise: *Are we capturing the data trends or the data source/actual physics?*. Even though more research would be needed to answer it conclusively, we can make a few observations. To start with, we are limited by the quality of the data in learning about the physics of the problem. For example, it could be that the data is poorly collected so that it does not provide a faithful reflection of the real physics. However, we assume that the data quality is good and that the real physical behavior is contained therein. Therefore, we can think of the parametrization (CE fitting) as an attempt to extract the physical behavior and learn it as well as possible with as few terms as possible. We expect that this low-fidelity representation of the high-fidelity data captures the behavior needed for the application. In some cases, the physical interactions are simple enough that a few parameters are able to capture these effects very well. This is the case, for example, of SiGe, where the strong nearest neighbor interactions are dominant. However, in other systems, such as MgLi, we need a lot of interactions and could be getting just the general trends right, but not capture most of the actual physics. It is conceivable that in systems with complex physics, while capturing the general trends with the parameters, we cannot put any physical meaning to these parameters.

In the end, if the parameters are not associated with real physics, but still capture the data trends, we should be happy with the predictive ability of the model. However, we have to be careful with concluding anything about physics being told by the parameters, but this is true for any statistical meta-model we build. In the case of this work, alloy energy landscapes tend to be explainable with few parameters simply because the energy binding together atoms tends to be local in nature (there are exceptions when considering, e.g., ionic systems) and hence, by the form of the cluster expansion (expanding nearest neighbors first, etc.) they will be picked up by the expansion.

5. Temperature dependence and its impact on the uncertainty

All the discussions presented so far assumed the use of DFT energies for the fitting, so the ECI are obtained at 0 K. Although it is sometimes assumed that these are valid for any temperature [33], as the temperature increases the degrees of freedom of the system are not only the configurational ones accounted for in the 0 K DFT simulations. In general, at a finite temperature T, the equilibrium of a system is defined by the minimum of its free energy F, not its internal energy U. For a constant volume V, the Helmholtz Free energy is

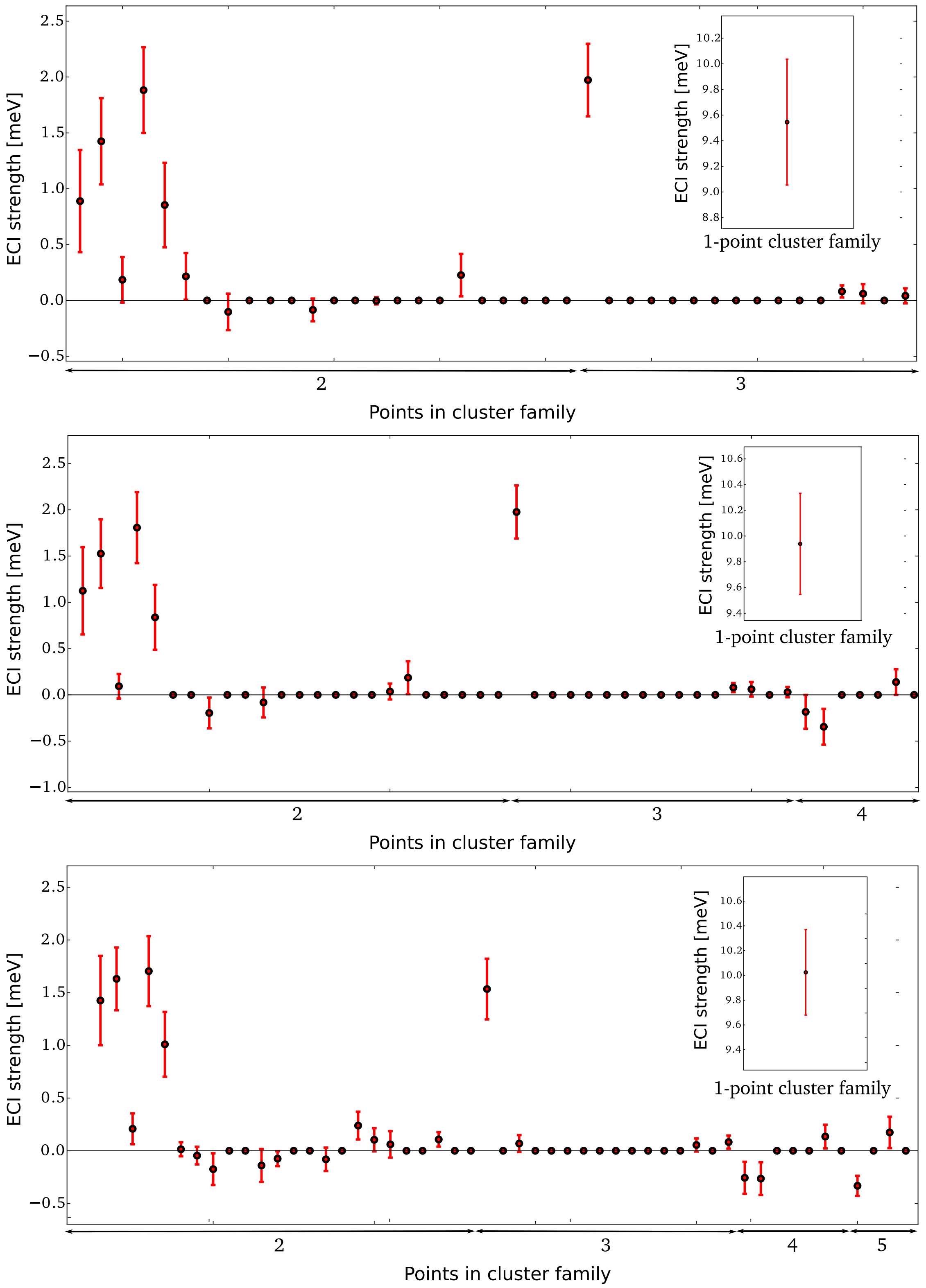


Fig. 3. ECI of the cluster for MgLi including an increasing number of points in the expansion (from top to bottom). Black dots represent the posterior mean for each of the coefficients and the red error bars the 95% confidence interval from Eqs. (16)–(18), respectively. The inset shows the ECI for the 1-point cluster. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$F = U - TS, \tag{26}$$

where S is the entropy of the system. Using the adiabatic approximation [60,61], we include two contributions to the free energy in the system for finite temperatures, the electronic (including the static lattice) and the vibrational free energies,

$$F(T,V) = F^{el}(T,V) + F^{vib}(T,V), \tag{27}$$

where $F^{el}(T,V)$ is the electronic free energy and $F^{vib}(T,V)$ is the vibrational free energy. Both of these contributions can be obtained from DFT simulations: F^{el} using a generalization to finite temperatures [62,63], and F^{vib} using the direct method within the quasi-harmonic approximation to calculate the phonon spectrum [64,65].

Remark 8. Even though the description of the cluster expansion formalism was done with only configurational degrees of freedom, it can be shown that it works as well for finite temperature systems just by replacing the energy from 0 K DFT calculations by the free energy $F(T,V)$ [66].

5.1. *Electronic free energy*

At a finite temperature T, the occupied states are not just those below the Fermi level E_F, but there is a finite probability of occupation of states of higher energy E. This probability is given by the Fermi–Dirac distribution, $f_{FD}(E;T)$ [67]:

$$f_{FD}(E;T) = \frac{1}{1 + \exp\left((E - E_F)/k_B T\right)}, \tag{28}$$

where k_B is the Boltzmann constant. Therefore, the spectral electron density $n(E;T)$ will be given by the product of this probability times the density of states of the system, $g(E)$:

$$n(E;T) = g(E) f_{FD}(E;T) = \frac{g(E)}{1 + \exp\left((E - E_F)/k_B T\right)}, \tag{29}$$

where we have assumed that the density of states does not depend on the temperature [68] and therefore can be obtained from 0 K DFT simulations.

The electronic free energy including finite temperature effects can be written as [63]:

$$F^{el} = E^{DFT} + \int_{-\infty}^{\infty} E f_{FD}(E;T) g(E)\, dE - TS^{el}, \tag{30}$$

where $E^{DFT} = U^{ext} + U^H + E^{xc}$ includes the interaction with external potential U^{ext}, the Hartree energy U^H, and the exchange–correlation energy E^{xc} [69]. The second term in the summation is a generalization to finite temperature of the sum of Kohn–Sham energies $\sum_i^{occ} \varepsilon_i \approx \int_{-\infty}^{E_F} E g(E)\, dE$ which at 0 K only runs over states below the Fermi energy. Finally, S_{el} is the electron entropy and is given by [63,68]

$$S^{el} = -\int_{-\infty}^{\infty} g(E) \left\{ f_{FD}(E;T) \log[f_{FD}(E;T)] + [1 - f_{FD}(E;T)] \log[1 - f_{FD}(E;T)] \right\} dE. \tag{31}$$

All the additional terms can be obtained using the density of states calculated from a 0 K DFT simulation of the system and therefore require almost no additional computational cost. Also, note that at 0 K, F^{el} reduces to the DFT calculated energy.

5.2. *Vibrational free energy*

Another source of free energy at finite temperatures are the vibrations of the lattice, which are usually represented by the associated quasi-particles, phonons [67]. The vibrational free energy for the phonon system can be written as [70,60]

$$F_{vib} = \frac{1}{2} \sum \hbar\omega_i + k_B T \sum \log\left(1 - e^{-\hbar\omega_i/k_B T}\right) \approx \int_0^{\infty} g_p(\omega) \left\{ \frac{1}{2}\hbar\omega + k_B T \log\left(1 - e^{-\hbar\omega/k_B T}\right) \right\} d\omega, \tag{32}$$

where $\hbar$ is the reduced Planck's constant, ω_i is the frequency of the phonon and the summation runs over all phonon modes. $g_p(\omega)$ is the phonon density of states that can be used to approximate the summation.

To calculate the free energy of any system, we need its phonon spectrum, and we use the same methodology implemented in ATAT, which is based on DFT simulations and uses the direct method [64,65]. Using this method, a supercell repeating the primitive cell of the material is created. One atom a of the original primitive cell (position 0) is displaced by a vector $\mathbf{u}(0, a)$ and the forces $\mathbf{F}(\mathbf{n}, b)$ at the rest of the atoms b of every copy of the primitive cell, with origin at $\mathbf{n}$, are recorded. These forces are related to the displacements through the force constant tensor $\mathbf{B}$,

$$\mathbf{F}(\mathbf{n}, b) = \sum_{\mathbf{m}} \mathbf{B}(\mathbf{n}, b, \mathbf{m}, a)\mathbf{u}(0, a), \tag{33}$$

where $\mathbf{m}$ denotes the origin of the periodic replicas of the cell 0 containing atom a. By running DFT simulations with independent displacements $\mathbf{u}(0, a)$ in the different atoms of the primitive cell and calculating the resulting forces in other atoms, we obtain an overdetermined system for the elements of $\mathbf{B}$ which can be solved using the least squares method [65,66]. In practice, not all atoms in a supercell need to be displaced and the symmetries of the crystal can be exploited to reduce the number of DFT simulations. The force constant tensor is then used to build the *dynamical matrix* $\mathbf{D}(\mathbf{k})$ of the crystal [65] for any value of crystal momentum $\mathbf{k}$. For a unit cell with n atoms, this matrix is of size $3n \times 3n$ (3 coordinates per atom) and its 3×3 entry for the atom pair ab is

$$\mathbf{D}(\mathbf{k}, ab) = \frac{1}{\sqrt{M_a M_b}} \sum_{\mathbf{m}} \mathbf{B}(0, b, \mathbf{m}, a) e^{-2\pi i \mathbf{k}\mathbf{m}}, \tag{34}$$

where M_a and M_b are the masses of atoms a and b, respectively. The eigenvalues of the dynamical matrix define the phonon frequencies of the system [65] for any given value of $\mathbf{k}$,

$$\mathbf{D}(\mathbf{k})\mathbf{w}(\mathbf{k}) = [\omega(\mathbf{k})]^2 \mathbf{w}(\mathbf{k}), \tag{35}$$

where $\omega(\mathbf{k})$ are the frequencies for a given momentum $\mathbf{k}$ and $\mathbf{w}(\mathbf{k})$ are their corresponding eigenvectors. Details of the derivation of these equations for the phonon calculations can be found in Appendix C.

Remark 9. In this work, we use forces from 0 K DFT simulations to obtain the phonon spectrum. However, in principle, these forces will also depend on the temperature and should be calculated using the full free energy F^{el}, Eq. (30), and not just E^{DFT}. This means that the phonon spectrum would also depend on the temperature. However, the effect of the electronic temperature is expected to be small [60], so we only use the 0 K calculation, thus removing the explicit temperature dependence in the phonon frequencies. This corresponds to the quasi-harmonic approximation [60].

5.2.1. The bond stiffness versus bond length approximation

The methodology described in the previous section to calculate the force constant tensor is well established, but it can be very costly for alloy simulations since the determination of the components of this tensor requires several calculations with a large supercell for each configuration. Even though these are done with frozen atomic positions, they still result in a considerable additional computational cost, especially if we use configurations with a large number of atoms in the training process. Therefore, instead of calculating the phonon modes using the direct method for every training configuration of the alloy, a subset of them, $\mathbf{\Sigma}_{fc} \subset \mathbf{\Sigma}_t$, is used to build a regression model for the dependence of the stiffness of the springs connecting nearest neighbor sites and their bond length L [66,71]. This regression can then be used to obtain the force constants for the rest of the configurations. We will call the new data set used for the regression of the stiffness $\mathcal{D}_{fc} = \{(\boldsymbol{\sigma}_i, \mathbf{E}(\Delta\boldsymbol{\sigma}_i))\}_{\boldsymbol{\sigma}_i \in \mathbf{\Sigma}_{fc}}$, where $\mathbf{E}(\Delta\boldsymbol{\sigma}_i)$ represents the set of all energies necessary to calculate the force constant tensor for the configuration $\boldsymbol{\sigma}_i$. This procedure is based on approximating the force constant tensor as

$$\mathbf{B}(\mathbf{m}, a, \mathbf{n}, b) = \mathbf{k}_{\sigma_a \sigma_b}(L) \tag{36}$$

where $\mathbf{k}_{\sigma_a\sigma_b}(L)$ is the (bond length-dependent) 3×3 stiffness tensor of the spring connecting the two sites $(\mathbf{m}, a)$ and $(\mathbf{n}, b)$ occupied by atoms of type σ_a and σ_b. These bond stiffnesses are assumed to take the same values for all configurations with the same parent lattice and only depend on the atomic identities σ_a and σ_b [66]. Furthermore, we only consider stretching and bending components as they are expected to be the most important [66]. This means that for a binary alloy we have six independent functions $k_{\sigma_a\sigma_b}(L)$, two for each of the three possible atomic pairings.

This step, while considerably reducing the computational effort, adds an extra level of uncertainty to the calculation. As with the CE coefficients, we capture this uncertainty using a Bayesian linear regression model with a RVM prior. Using an independent linear model for each the independent stiffness component $k_{\sigma_a\sigma_b}$ of the force constant tensor $\mathbf{B}$ with coefficients $\boldsymbol{\xi}_{\sigma_a\sigma_b} = (\xi_1 \dots \xi_P)^T$ and basis $\boldsymbol{\chi}_{\sigma_a\sigma_b}(L) = \left(\chi_1(L) \dots \chi_P(L)^T\right)$, we have

$$k_{\sigma_a\sigma_b}(L) = \boldsymbol{\xi}^T_{\sigma_a\sigma_b} \boldsymbol{\chi}_{\sigma_a\sigma_b}(L). \tag{37}$$

Therefore, we obtain a probability distribution for the coefficients of the stiffness given the subset of structures for the fit $\mathbf{\Sigma}_{fc}$,

$$p(\boldsymbol{\xi}_{\sigma_a\sigma_b} \mid \mathcal{D}_{fc}) = \mathcal{N}(\boldsymbol{\xi}_{\sigma_a\sigma_b} \mid \boldsymbol{\mu}_{\sigma_a\sigma_b}, \mathbf{S}_{N,\sigma_a\sigma_b}), \tag{38}$$

where $\boldsymbol{\mu}^i$ and $\mathbf{S}_N^i$ are given by Eqs. (18) and (17). The predictive distribution for the stiffness used to sample its value is given by

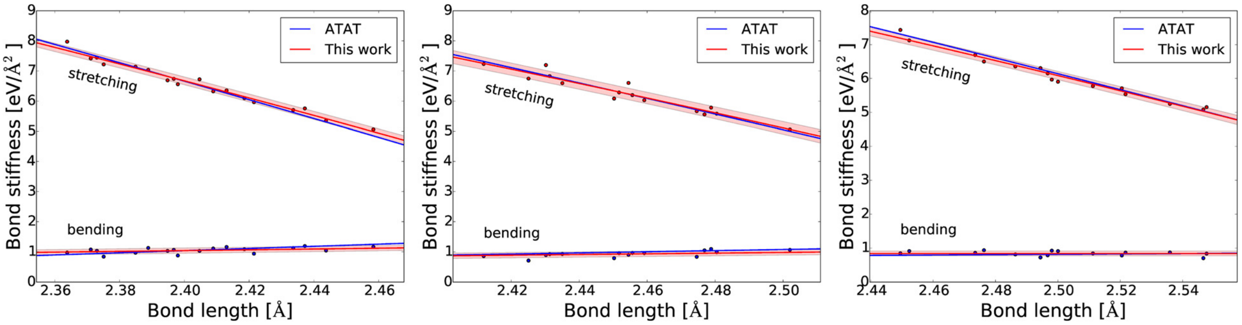


Fig. 4. Fit of the bond stiffness k_{ab} as a function of the bond length L for SiGe using a second-order polynomial. The fit for Si–Si (left), Ge–Si (middle) and Ge–Ge (right) bonds are shown. Each figure contains the stretching (red dots) and bending (blue dots) components of the tensor calculated for five configurations, a fit using Bayesian linear regression (red lines plus shades indicating one standard deviation uncertainty) and the least-squares fits from ATAT (blue lines). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$p(k_{\sigma_a\sigma_b} \,|\, \mathcal{D}_{fc}) = \mathcal{N}(k_{\sigma_a\sigma_b} \,|\, \mu_{\sigma_a\sigma_b}, s_{\sigma_a\sigma_b}), \tag{39}$$

where the mean μ^i and variance s^i are given by Eqs. (23) and (22), respectively. Fig. 4 shows the fit of the two independent force constant tensor components, stretching and bending, for SiGe using Bayesian linear regression and the least-squares method as implemented in ATAT. Since there are two types of atoms, there are three bonds that need to be fitted, Si–Si, Si–Ge and Ge–Ge bonds. Eight configurations were used to generate the data for fitting, pure Si, pure Ge, four $Si_{0.5}Ge_{0.5}$ configurations with space groups $\mathtt{F}\overline{4}\mathtt{3m}$ (2 atoms/unit cell), $\mathtt{R}\overline{3}\mathtt{m}$ (2 configurations with 4 atoms/unit cell) and $\mathtt{C2/c}$ (8 atoms/unit cell), one $Si_{0.25}Ge_{0.75}$ configuration ($\mathtt{R3m}$, 4 atoms/unit cell) and one $Si_{0.75}Ge_{0.25}$ configuration ($\mathtt{C2/m}$, 8 atoms/unit cell). We can see that the standard deviation of the predictive model using this approximation with the Bayesian fit stays below 5% for all stretching terms and therefore we do not expect the stiffness versus length approximation to increase the uncertainty too much.

Using this approximation, the dynamical matrix can now be easily obtained for any alloy configuration using Eqs. (34) and (36) and used to calculate the vibrational free energy F^{vib}, Eq. (32).

Remark 10. This error includes limited fitting data and model inadequacy. As for the model, we have only considered a polynomial fit as this is also the implemented method in ATAT. The appropriate polynomial order depends on the material, but we set a maximum of third order and let the RVM select the order that maximizes the evidence. The limited data error can be reduced by including more configurations to the fit of the bond stiffness to bond length. For example, if we want more data for the Ge–Ge bond model, we would include Ge-rich configurations, as these contain more Ge–Ge bonds.

Remark 11. Since there is a probability distribution associated to the nearest neighbor bond stiffness and therefore the force constants, $p(B^i \,|\, \mathcal{D}_{fc})$, the vibrational free energy also has an associated probability distribution $p(F^{vib} \,|\, \mathcal{D}_{fc})$. Even though we do not have an analytical expression, we can sample from it in a three step process: (i) independently sample the bond stiffness for the different atom types from Eq. (39), (ii) construct the force constant tensor using Eq. (36) (iii) calculate the phonon spectrum for the given force constants using Eqs. (34) and (35) and (iv) calculate F^{vib} using Eq. (32) with this phonon spectrum.

Remark 12. The calculation of the vibrational free energy using the full model as described in the previous section is the recommended method. However, as its cost can be very high in some circumstances, such as the use of many training configurations with large unit cells, the bond stiffness versus bond length approximation provides an alternative at a much lower cost. The user of this approximation must be aware that the saving in cost comes at the price of an increased uncertainty in the results, and this can be significant.

5.3. Temperature dependent uncertainty in the ECI

Using the methods described above, we can compute the free energy of the system at any temperature, which is the quantity whose minimization will lead to the stable configuration. These values of the free energy can be used to fit the ECI at different temperatures and these can then be used in thermodynamic calculations [71]. In order to do this, we first select a temperature range and temperature step to calculate the ECI. In between these temperatures, we use an interpolation. In our case, we chose a temperature range from 0 to 2000 K rand a step of 100 K. For each of these temperatures, we calculate the electronic free energy as in Eq. (30) and the vibrational free energy as in Eq. (32). Using the adiabatic approximation we add them to obtain the total free energy for each temperature, Eq. (27). Fitting the free energy at each temperature T, we obtain a predictive distribution for the free energy given the temperature, $p(F \,|\, \mathcal{D}_N, T)$, Eq. (21). This process is summarized in Algorithm 2:

Algorithm 2 Uncertainty in free energy calculation.

```
Select $\Delta T$ and $N_T$ that define the temperature range, $\mathbf{T} = \{i\Delta T\}_{i=0}^{N_T}$
for all $\boldsymbol{\sigma}_i \in \boldsymbol{\Sigma}_t$ do
    Generate the electronic free energy $F^{elec}$ using Eq. (30) for all $T \in \mathbf{T}$
    Generate the vibrational free energy $F_i^{vib}$ using Eq. (32) for all $T_i \in \mathbf{T}$
    Calculate the total free energy using $F = F^{elec} + F_i^{vib}$, Eq. (27)
end for
for all $T \in \mathbf{T}$ do
    Fit the coefficients $\boldsymbol{\gamma}$ of the cluster expansion to the free energy $F$, obtain $p(\boldsymbol{\gamma} \mid \mathcal{D}_N, \boldsymbol{\alpha}_{MP}, \beta_{MP}, T) = \mathcal{N}(\boldsymbol{\gamma} \mid \boldsymbol{\mu}(T), \mathbf{S}_N(T))$, Eqs. (16)–(18)
    Calculate parameters $s_i^*(T)$, $\mu_i^*(T)$ of the predictive distribution $p(F \mid \mathcal{D}_N, T)$, Eqs. (22) and (23).
end for
```

Algorithm 3 Uncertainty in free energy calculation using the stiffness versus length approximation.

```
Select the number of samples $N_s$ of $F^{vib}$ to approximate $p(F \mid \mathcal{D}_N, \mathcal{D}_{fc}, T)$, Eq. (41)
Select $\Delta T$ and $N_T$ that define the temperature range, $\mathbf{T} = \{i\Delta T\}_{i=0}^{N_T}$
for all $\boldsymbol{\sigma}_i \in \boldsymbol{\Sigma}_t$ do
    Generate the electronic free energy $F^{elec}$ using Eq. (30) for all $T \in \mathbf{T}$
    for $i = 1, \ldots, N_s$ do
        Sample the stiffness from $p(k_{\sigma_a \sigma_b} \mid \mathcal{D}_{fc})$, Eq. (39)
        Build the force constant tensor from Eq. (36)
        Calculate the phonon spectrum, Eqs. (34) and (35)
        Generate the vibrational free energy $F_i^{vib}$ using Eq. (32) for all $T_i \in \mathbf{T}$
    end for
    for $i = 1, \ldots, N_s$ do
        Calculate the total free energy using $F = F^{elec} + F_i^{vib}$, Eq. (27)
        for all $T \in \mathbf{T}$ do
            Fit the coefficients $\boldsymbol{\gamma}$ of the cluster expansion to the free energy $F$, obtain $p(\boldsymbol{\gamma} \mid \mathcal{D}_N, \boldsymbol{\alpha}_{MP}, \beta_{MP}, T, F_i^{vib}) = \mathcal{N}(\boldsymbol{\gamma} \mid \boldsymbol{\mu}_i(T), \mathbf{S}_{N,i}(T))$, Eqs. (16)–(18)
        end for
    end for
end for
for $i = 1, \ldots, N_s$ do
    for all $T \in \mathbf{T}$ do
        Calculate parameters $s_i^*(T)$, $\mu_i^*(T)$ of the predictive distribution $p(F \mid \mathcal{D}_N, \mathcal{D}_{fc}, T, F_i^{vib})$, Eqs. (22) and (23)
    end for
end for
Estimate the predictive distribution for each temperature using Eq. (41)
```

5.3.1. *Uncertainty using the bond stiffness versus the bond length approximation*

In the case of the stiffness versus length approximation, for the vibrational free energy, since we used a Bayesian regression for the force constants, we have a probability distribution for each temperature, $p(F^{vib} \mid \mathcal{D}_{fc}, T)$ which can be sampled as described in Remark 11. For each sample of F^{vib}, we will have a different free energy, Eq. (27), and therefore each of the samples will require a separate cluster expansion for F. Each of these cluster expansions provides us with a posterior distribution for the ECI $p(\boldsymbol{\gamma} \mid \mathcal{D}_N, \boldsymbol{\alpha}_{MP}, \beta_{MP}, T, F^{vib})$, Eq. (16), and a predictive distribution for the free energy given the vibrational free energy, $p(F \mid \mathcal{D}_N, T, F^{vib})$, Eq. (21). If we integrate out the dependence on F^{vib}, we obtain the final predictive distribution for the free energy at the given temperature T,

$$p(F \mid \mathcal{D}_N, \mathcal{D}_{fc}, T) = \int p(F \mid \mathcal{D}_N, T, F^{vib}) p(F^{vib} \mid \mathcal{D}_{fc}, T)\, dF^{vib}. \tag{40}$$

In this case the predictive distribution for the free energy depends on two different data sets. The first one, $\boldsymbol{\Sigma}_t$ is the training set of configurations from which we fit the ECI, whereas the second set, $\boldsymbol{\Sigma}_{fc} \subset \boldsymbol{\Sigma}_t$, appears because of the approximation for the vibrational free energy, which only uses a subset of the training configurations. Even though we cannot obtain the distribution $p(F^{vib} \mid \boldsymbol{\Sigma}_{fc}, T)$ analytically, we can approximate it using N_s samples as described before, so that the predictive distribution for the free energy becomes

$$p(F \mid \mathcal{D}_N, \mathcal{D}_{fc}, T) \approx \frac{1}{N_s} \sum_i p(F \mid \mathcal{D}_N, T, F_i^{vib}), \tag{41}$$

which is a mixture of Gaussians. This distribution, which is not Gaussian, has mean and variance $\mu^*(T) = \frac{1}{N_s} \sum \mu_i^*(T)$ and $s^*(T) = \frac{1}{N_s} \sum (s_i^*(T) + \mu_i^*(T)^2) - \mu^*(T)^2$ [72]. Using this approximation, Algorithm 2 is modified in the form of Algorithm 3 shown next.

Fig. 5 shows the free energy per atom for a test configuration predicted by the cluster expansion together with its uncertainty as the 95% confidence interval. The values calculated from the DFT simulations using Eqs. (30) and (32), including the full phonon model to obtain F^{vib}, are also shown as black dots. The uncertainty of the predictive distribution increases with the temperature, and the values calculated using DFT are within the 95% confidence interval for all the temperature range.

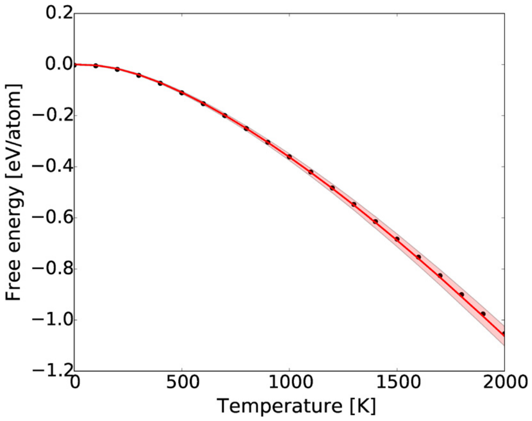


Fig. 5. Predicted dependence of the free energy with the temperature for one test configuration with composition $Si_{0.25}Ge_{0.75}$. The 95% confidence interval is shown as a shaded red region around the mean prediction. The values calculated from the DFT simulations using Eqs. (30) and (32) are shown as black dots. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6. Ground state search

One of the most important steps in the study of a new alloy is the determination of its ground state line (GSL), which contains the concentrations and formation energies of the structures that are thermodynamically stable at a given temperature. When calculated at 0 K using DFT, these structures can also be used as a starting point to calculate the (T, x) phase diagram.

Once the ECI are fitted for the formation energy as explained in Section 4, using the model in Eq. (25), we calculate the formation energy for a large pool of alloy configurations $\boldsymbol{\Sigma}_s = \{\boldsymbol{\sigma}_j\}$. For instance, in the following examples this pool will consist of all symmetrically inequivalent configurations with up to ten atoms, generated again using ATAT's `maps` tool. Not all minimum energy configurations for a given concentration belong to the ground state line. Given a configuration $\boldsymbol{\sigma}$, consider two configurations $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ with concentrations x_{σ_1} and x_{σ_2} such that $x_{\sigma_1} < x_\sigma < x_{\sigma_2}$. If the average energy of $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ is lower than the energy of $\boldsymbol{\sigma}$, then a mixed phase of $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ is more favorable energetically than a pure phase of $\boldsymbol{\sigma}$. Therefore, the GSL is formed by the structures with non-positive formation energies in the convex hull of the set $\left\{(x_{\sigma_j}, \Delta_f E(\boldsymbol{\sigma}_j))\right\}_{\boldsymbol{\sigma}_j \in \boldsymbol{\Sigma}_s}$ [23], where x_{σ_j} is the concentration of configuration $\boldsymbol{\sigma}_j$ and $\Delta_f E(\boldsymbol{\sigma}_j)$ its formation energy.

Since the ECI that we determined from the Bayesian fit are random variables, the predicted formation energies for a given structure will have an uncertainty which will be reflected not only in the energy of the ground state line but, more importantly, also in which structures belong to the ground state line. The uncertainty is calculated as follows. Given the probability distribution of the ECI, we draw a sample from it. For this sample, we calculate the formation energies of all structures in the generated pool $\boldsymbol{\Sigma}_s$. From these energies we calculate the convex hull and therefore obtain the GSL. We repeat this process for a given number of random ECI samples N_s. If we define a quantity of interest as the ground state configuration for a given concentration x, $I = \boldsymbol{\sigma}_x \in GSL$, by using Eq. (2) we obtain

$$\begin{aligned} p(I) = p(\boldsymbol{\sigma}_x \in GSL) &= \int \delta\left(I(\boldsymbol{\gamma}) - I\right) p(\boldsymbol{\gamma} \,|\, \mathcal{D}_N)\, d\boldsymbol{\gamma} \\ &\approx \frac{1}{N_s} \sum_{i=1}^{N_s} \delta\left(I(\boldsymbol{\gamma}_i) - I\right) = N_{\sigma_x}/N_s, \end{aligned} \tag{42}$$

where we have used a Monte Carlo approximation with N_s samples of $\boldsymbol{\gamma} \sim p(\boldsymbol{\gamma} \,|\, \mathcal{D}_N)$ for the integral. $I(\boldsymbol{\gamma})$ is the configuration predicted by the CE with coefficients $\boldsymbol{\gamma}$ to be the ground state at concentration x and N_{σ_x} is the number of times the configuration $\boldsymbol{\sigma}_x$ is predicted by the CE to be the ground state for concentration x. Even though in principle $\boldsymbol{\sigma}_x$ is arbitrary, in the implementation it is limited to the structures in the pool $\boldsymbol{\Sigma}_s$, so only a finite set of concentrations x will be represented. This process is summarized in Algorithm 4.

Fig. 6 shows the free energies of formation of SiGe alloys, both the training values calculated using DFT (black circles) and the mean predictions calculated with a trained cluster expansion with 2- and 3-point clusters with a size up to 9 and 5 Å, respectively, in the initial basis functions (blue crosses). Fig. 6(a) shows that this is an example of an alloy where at 0 K all configurations have positive formation energies, which means that the only thermodynamically stable configurations are the pure components, Si and Ge. On the other hand, Fig. 6(b) shows that if we look at the free energy of formation at a finite temperature T using the T-dependent ECI calculated as described in Section 5, it becomes favorable for some of the alloy configurations to form instead of remaining in a phase with pure Si and pure Ge coexisting. This will become clearer as well when we look at the phase diagrams in the next section.

Algorithm 4 Uncertainty in the GSL.

```
Given the posterior distribution of the ECI, $p(\boldsymbol{\gamma} \mid \mathcal{D}_N, \boldsymbol{\alpha}_{MP}, \beta_{MP})$, Eqs. (16)–(18)
Select a pool of configurations to be considered, $\boldsymbol{\Sigma}_s = \{\boldsymbol{\sigma}_i\}$
Select number of samples of the GSL, $N_s$
Let $N_{\boldsymbol{\sigma}}$ be the number of appearances of configuration $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}_s$ in samples of the GSL
Initialize $N_{\boldsymbol{\sigma}} = 0 \; \forall \boldsymbol{\sigma} \in \boldsymbol{\Sigma}_s$
for $i = 1, \ldots, N_s$ do
    Sample $\boldsymbol{\gamma}_i$ from $p(\boldsymbol{\gamma} \mid \mathcal{D}_N, \boldsymbol{\alpha}_{MP}, \beta_{MP})$, Eqs. (16)–(18)
    for all $\boldsymbol{\sigma}_j \in \boldsymbol{\Sigma}_s$ do
        Calculate $x_{\boldsymbol{\sigma}_j}$, concentration of one of the elements of the binary alloy
        Calculate the formation energy $\Delta_f E(\boldsymbol{\sigma}_j \mid \boldsymbol{\gamma}_i) = \boldsymbol{\gamma}_i^T \boldsymbol{\phi}(\boldsymbol{\sigma}_j)$, Eq. (25)
    end for
    Obtain the set of configurations $\mathcal{C}_i$ which form the convex hull of $\left\{(x_{\boldsymbol{\sigma}_j}, \Delta_f E(\boldsymbol{\sigma}_j \mid \boldsymbol{\gamma}_i))\right\}_{\boldsymbol{\sigma}_j \in \boldsymbol{\Sigma}_s}$
    for all $\boldsymbol{\sigma}_j \in \mathcal{C}_i$ do
        $N_{\boldsymbol{\sigma}_j} \to N_{\boldsymbol{\sigma}_j} + 1$
    end for
end for
for every concentration $x$ represented in $\boldsymbol{\Sigma}_s$ do
    for every $\boldsymbol{\sigma} \mid x_{\boldsymbol{\sigma}} = x$ do
        $p(\boldsymbol{\sigma} \in \mathrm{GSL}) \approx N_{\boldsymbol{\sigma}} / N_s$, Eq. (42)
    end for
end for
```

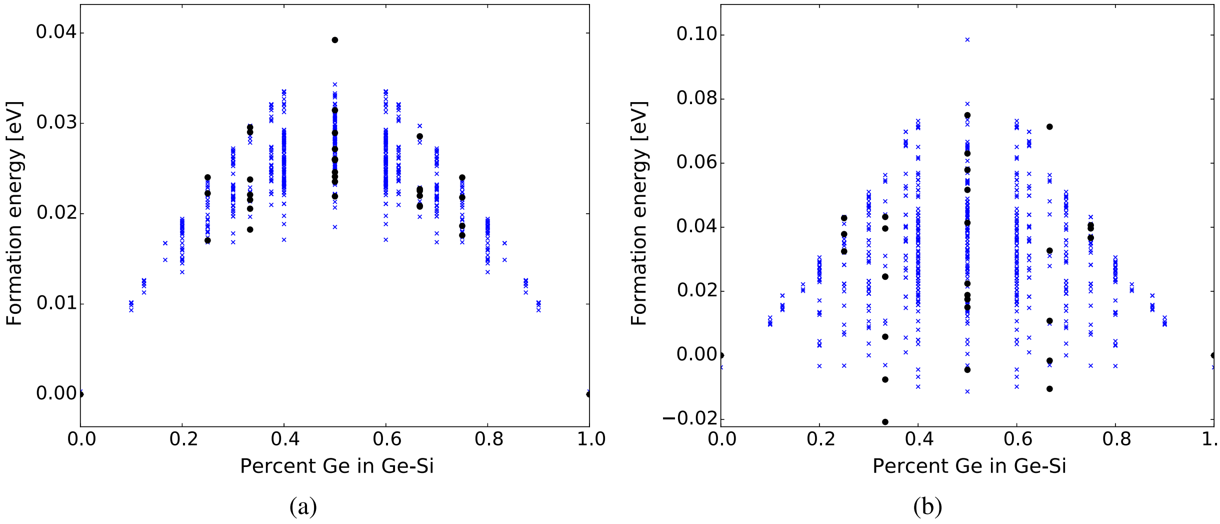


Fig. 6. Free energy of formation of SiGe at (a) 0 K and (b) 2000 K. Training data calculated with DFT are shown as black circles and the mean prediction from CE for all symmetrically different configurations containing up to ten atoms in the unit cell as blue crosses. The CE was trained using an initial basis containing 2- and 3-point clusters. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 7 shows the formation energies of MgLi alloys, again including the 81 DFT training data as black circles, and the median of the calculated values with a trained cluster expansion for all symmetrically different configurations containing up to ten atoms in the unit cell as blue crosses. In this case, in the training process, we included up to 5-point clusters with the following sizes: 2-point: 2 nm, 3-point: 1.4 nm, 4- and 5-point: 0.9 nm. The median ground state line with a 95% confidence interval predicted using the CE surrogate model to generate 5000 GSL samples is also shown. The insets in the figure show, for a subset of the concentrations, all the structures with non-zero probability of belonging to the GSL as predicted using Algorithm 4 with a plot of the most likely atomic configuration. For example, at a Mg concentration of $x_{Mg} = 0.8$, our probabilistic CE predicts that there is a 48% probability that no structure belongs to the GSL. The next most likely outcome, with a probability of 37%, is that the structure labeled as `00056` (space group `I4/m`) belongs to the GSL. Since only structures with up to 10 atoms in the unit cell were considered, our results cannot say anything about the probability of belonging to the GSL of other configurations with larger unit cells.

Suppose that for a given concentration all structures with non-zero probability of belonging to the GSL as predicted by the CE were in the training set. In this case, we know the exact energies of the configurations as the training data is considered ground truth. Therefore, we know that in this case the uncertainty in which structure belongs to the GSL for the given concentration comes only from the error in the ECI fit. This type of error can be seen, for example, at the point $x = 1/3$, where two of the structures predicted to be in the GSL with non-zero probability are in the training set, `00008` (`I4/mmm`) and `00067` (`Cmcm`). Configuration `00067` has the lowest training energy, so we will call it $\boldsymbol{\sigma}_{min,x=1/3}$ and its

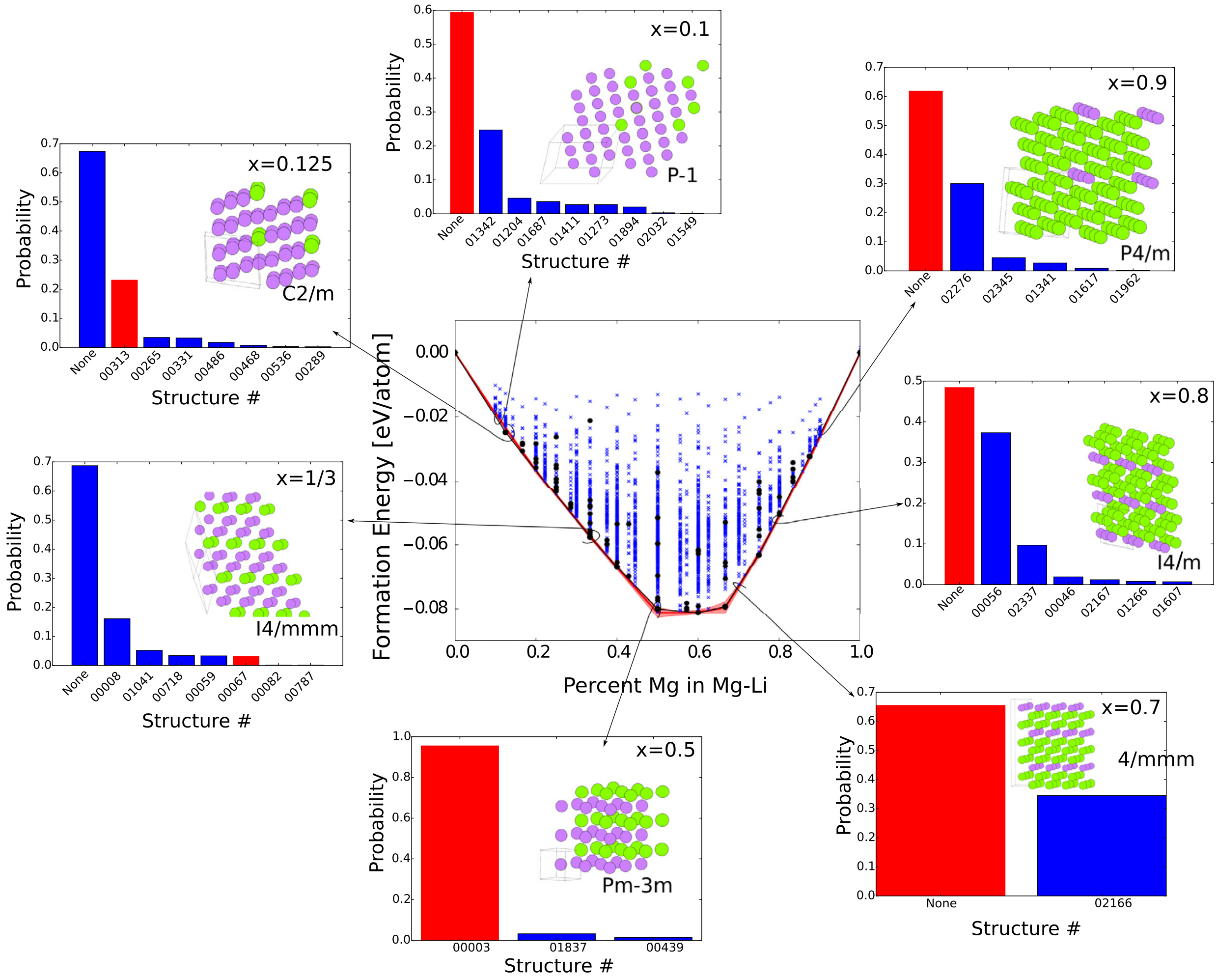


Fig. 7. Formation energies of MgLi. Training data calculated using DFT are shown as black circles and mean predictions from CE as blue crosses. CE energies are shown for all symmetrically different configurations containing up to ten atoms in the unit cell. The ground state line calculated from the training data is shown with a black line and the one from CE predictions with a red line at the median and a 95% confidence interval around it. Results were obtained using up to 5 point clusters (2-pt: 20 Å, 3-pt: 14 Å, 4-pt: 9 Å, 5-pt: 9 Å) for the initial basis functions. The probability distribution for the predicted structure in the GSL is shown for a set of concentrations. The red bar is the one that belongs to the training GSL at each of the concentrations. The labels represent the order of enumeration from ATAT and None means that no structure of that concentration in the GSL is predicted. The insets contain the atomic arrangement of the most likely structure according to the CE together with its space group using the international short symbol. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

formation energy $\Delta_f E_{min,x=1/3} = \Delta_f E(\boldsymbol{\sigma}_{min,x=1/3})$. However, the cluster expansion ranks it only as the fifth most probable one, behind the structure 00008, which appears as the most likely configuration to be in the GSL at $x = 1/3$, ignoring the case that no configuration with $x = 1/3$ belongs to the GSL. On the other hand, in cases such as $x = 0.5$, the CE has most of the weight of its prediction at $\boldsymbol{\sigma}_{min,x=0.5}$, with only a low probability assigned to other structures not included in the training set. Finally, in cases such as $x = 0.7$ or $x = 0.9$, the CE has its maximum weight on $\boldsymbol{\sigma}_{min,x=0.7}$ and $\boldsymbol{\sigma}_{min,x=0.9}$, but it also gives a high probability to some structures not in the training set. These configurations can then be further studied by DFT simulations to verify the CE predictions.

6.1. Impact of training set and maximum cluster size

To study the convergence of the results, we consider different training set sizes and maximum number of points in the clusters used in the expansion. Fig. 8 shows the bcc MgLi ground state line and the 95% confidence intervals for different CE training configurations. We consider the cases using up to 3-, 4- and 5-point clusters for the initial basis functions in the regression model and also using 20, 40 and 81 training points, chosen in the order that they are incorporated into the training set by the maps tool. As expected, the simulations using all the training data but different number of clusters show an increase in the uncertainty in the GSL energies as we reduce the maximum number of points in any cluster, even though this increase in the uncertainty is not very large. This is expected to a certain extent if we look at the ECI strengths of 4-

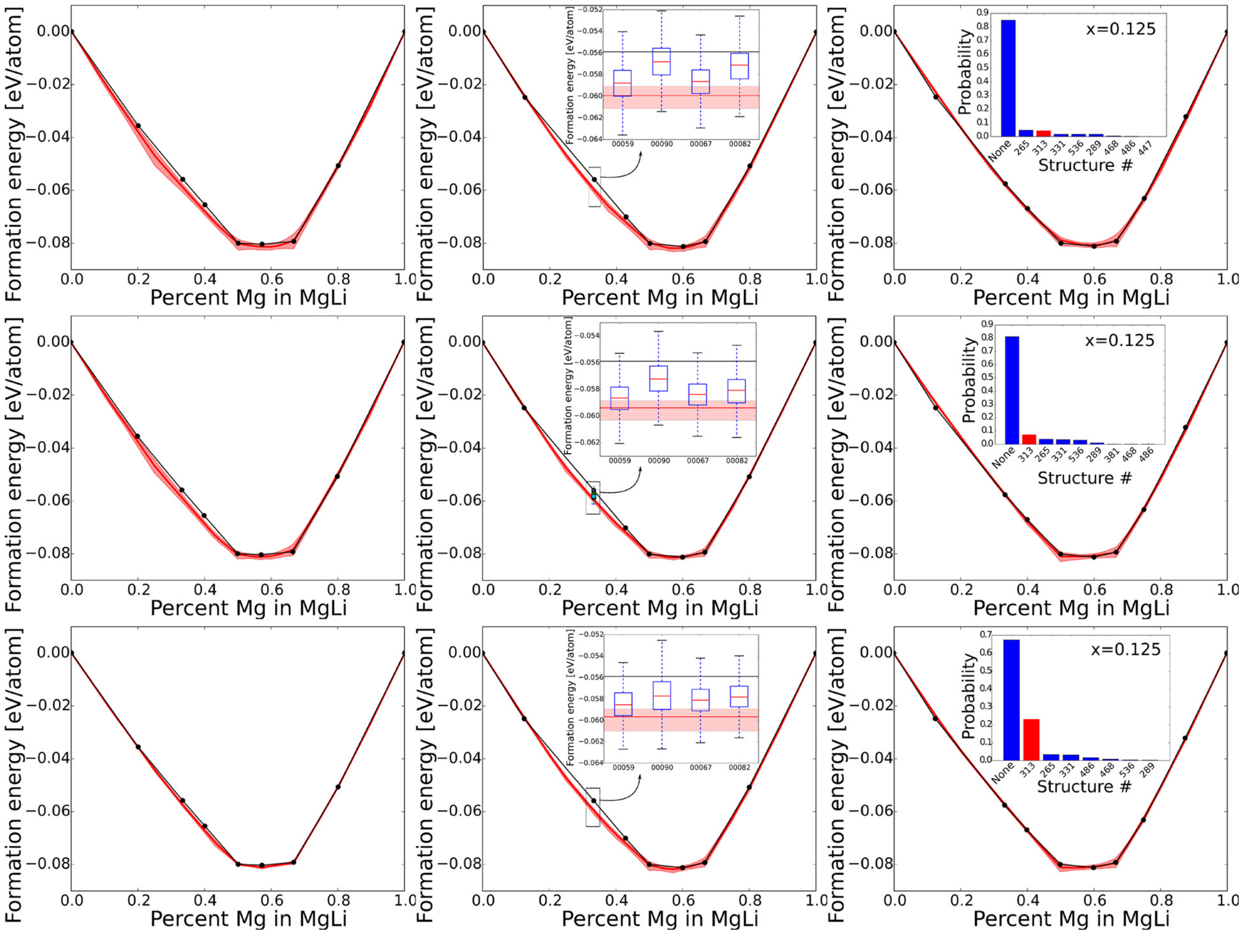


Fig. 8. bcc MgLi ground state line with 95% confidence intervals for different CE training configurations. From top to bottom: using up to 3-, 4- and 5-point clusters in the regression model. From left to right: using 20, 40 and 81 training points. The insets for 81 training points show, for $x = 0.125$, the probability of the different configurations of belonging to the GSL. The insets for 40 training points show a boxplot of the CE predicted energies for the four lowest energy configurations at 33% Mg content together with the energy of the lowest training energy (black line) and the GSL (red line with the 95% confidence interval as shaded region). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and 5-point clusters in Fig. 3, which are considerably smaller that the 2-point and smallest 3-point clusters. If we look at the uncertainty in the structures belonging to the GSL, we find that the impact is quite large in some cases. For example, as shown in the insets on the right column, the point at $x = 0.125$ has a very low probability to be in the GSL if we only include up to 4-point clusters and even lower if only include up to 3-point clusters.

As we decrease the number of training configurations to 40, the uncertainty in the energy of the GSL as predicted by the CE increases, and so does the deviation between it and the energy of the GSL constructed only from the training data, the "training GSL". There are two different errors playing a role here and unfortunately they are not clearly shown by the error bars. This is especially clear at $x = 1/3$ for the GSL using 40 training configurations. In this case, the CE predicted mean energies of several structures fall below $\Delta_f E_{min,x=1/3}$, even though this value is within their individual 95% confidence intervals (∼5–10 meV) in all cases. However, the GSL only contains the lowest of the energies of these configurations, so the 95% confidence interval of the predicted GSL energy at $x = 1/3$ appears very small, ∼2 meV. This effect is shown in the insets in the second column of Fig. 8, which show a boxplot of the CE predicted energies for the four lowest energy configurations at 33% Mg content, `00059` (`Cmmm`), `00067` (`Cmcm`), `00082` (`C2/m`) and `00090` (`P`$\overline{3}$`m1`). The lowest training energy $\Delta_f E_{min,x=1/3}$ is shown as a black line and the CE GSL with a red line with a 95% confidence interval as shaded region. The span of the CE energy predictions includes $\Delta_f E_{min,x=1/3}$, but the fact that only the lowest of them can belong to the GSL for any single sample of the ECI leads to a smaller confidence interval in the GSL that does not include $\Delta_f E_{min,x=1/3}$. This effect is further enhanced by the cases where no configuration with $x = 1/3$ belongs to the GSL, since this means that the energy of the GSL at that concentration is lower than the lowest of the energies of any configuration with $x = 1/3$.

Finally, as we decrease the number of training configurations to only 20, we find that even with the use of a RVM for pruning the initial 54 model basis functions, a large overconfidence in the predictions of the model appears [73], and the prediction errors are as low as 1 meV. The situation improves if we reduce the original number of basis functions for the regression. Therefore, column 1 of Fig. 8 shows the GSL prediction with 95% confidence intervals reducing the maximum radius of three point clusters to 10 Å, which is justified as we know that only the smallest 3-point cluster play a role as shown in Fig. 3. In this case, the predictions of the model show a clear increase in the uncertainty, even though, as in the case of 40 training points, the interplay in the uncertainty of different simulations keeps the uncertainty in the energy of the GSL down.

Remark 13. Since in every case most configurations belonging to the training GSL were included in the reduced training sets, there is not any major departure in energy from this training GSL in any case, which also shows the effectiveness of the algorithm used by ATAT to choose the training configurations. However, the probabilities of individual configurations of belonging to the GSL can differ quite significantly with different initial maximum number of clusters or number of training data.

7. Phase diagrams

The GSL only contained information of the thermodynamically stable configurations at 0 K. However, as the temperature is increased, thermal energy starts playing a role in determining which configurations are the most stable ones. (T, x) phase diagrams include this information giving the most stable structure for a given temperature and concentration. These diagrams are usually computed using a Monte Carlo method within the semi-grand-canonical or transmutation ensemble, where the energy E and alloy concentration x are allowed to change on a lattice with a fixed number of sites N, and the temperature T and chemical potential μ are externally fixed for each of the points of the diagram. The natural thermodynamic potential (per atom) for the semi-grand canonical ensemble is

$$\phi(T,\mu) = -\frac{1}{\beta N}\log\sum_i \exp\left[-\beta N(E_i - \mu x_i)\right], \tag{43}$$

where β is the inverse temperature defined in terms of the temperature T as $1/k_B T$ and the summation is over all possible states of the system characterized by a different occupation of each site of the lattice. Equivalently, it can also be defined through the following differential

$$d(\beta\phi) = (E - \mu x)\,d\beta - \beta x\,d\mu, \tag{44}$$

where E and x are now the system's average energy and concentration, respectively.

To construct a phase diagram, we use the phb tool from ATAT, which follows the phase boundary between phases found from the GSL constructed from 0 K formation energies. This is done using the fact that the potentials for each phase are the same at the phase boundary. For example, equating Eq. (44) for phases α and γ,

$$(E^\alpha - \mu x^\alpha)\,d\beta - \beta x^\alpha\,d\mu = (E^\gamma - \mu x^\gamma)\,d\beta - \beta x^\gamma\,d\mu, \tag{45}$$

which can be recast as an equation for the derivative $d\mu/d\beta$,

$$\frac{d\mu}{d\beta} = \frac{E^\gamma - E^\alpha}{\beta(x^\gamma - x^\alpha)} - \frac{\mu}{\beta}. \tag{46}$$

The necessary averages E^γ, E^α, x^γ and x^α are computed from a MC simulation of the system at the given T and μ. The energies for a given configuration are obtained from the trained cluster expansion, whereas the concentrations are directly obtained from the number of atoms of each species in the simulation domain. Using the value of the slope of the boundary, $d\mu/d\beta$ and an initial state (μ^0, β^0), the boundary at step $t+1$ is obtained from the previous step as

$$\mu^{t+1} = \mu^t + \left.\frac{d\mu}{d\beta}\right|^t (\beta^{t+1} - \beta^t) = \mu^t + \left(\frac{E^{\gamma,t} - E^{\alpha,t}}{x^{\gamma,t} - x^{\alpha,t}} - \mu^t\right)\left(\frac{\beta^{t+1}}{\beta^t} - 1\right), \tag{47}$$

with the process starting from the $T = 0$ ground states. A more detailed description of the calculation of phase diagrams can be found in Ref. [48].

Remark 14. In this section, we do not consider temperature dependence of the ECI, even though it is possible to include it as described in Section 5. Even though this will affect the values for the phase boundary, it does not affect the methodology presented.

Since we have a probability distribution for the ECI, Eq. (16), the phase boundary is not unique and depends on the particular value of the ECI used to calculate the system energy. To obtain the uncertainty in the phase boundaries, we use a similar MC approach as that used for the GSL. For each realization of the ECI from Eq. (16), we obtain a different realization of the phase diagram for a given alloy (see Algorithm 5).

Algorithm 5 Uncertainty in the phase boundary calculation.

```
Given the posterior distribution of the ECI, $p(\boldsymbol{\gamma} \mid \mathcal{D}_N, \boldsymbol{\alpha}_{MP}, \beta_{MP})$ Eqs. (16)–(18)
Select initial and minimum temperature steps, $\Delta T_0$ and $\Delta T_{min}$
Select maximum concentration step, $\Delta x_{max}$
Select number of samples, $N_s$
for $i = 1, \ldots, N_s$ do
    Sample $\gamma_i$ from $p(\boldsymbol{\gamma} \mid \mathcal{D}_N, \boldsymbol{\alpha}_{MP}, \beta_{MP})$, Eqs. (16)–(18)
    Calculate the GSL as the convex hull of $\left\{(x_{\boldsymbol{\sigma}_j}, \Delta_f E(\boldsymbol{\sigma}_j \mid \boldsymbol{\gamma}_i))\right\}_{\boldsymbol{\sigma}_j \in \boldsymbol{\Sigma}_s}$, Algorithm 4
    Define $N_{GSL,i}$ as the number of configurations in the GSL with ECI $\gamma_i$
    for $j = 1, \ldots, N_{GSL,i} - 1$ do
        Select phases $j$ and $j + 1$ to follow their boundary
        Set $\Delta T = \Delta T_0$
        while $\Delta T \geq \Delta T_{min}$ and not end of boundary do
            Run phb to follow boundary with step $\Delta T$
            if $\Delta x > \Delta x_{max}$ then
                $\Delta T \to \Delta T/2$
            end if
        end while
        if Third phase $k$ appeared then
            Run phb to follow boundaries $j - k$ and $(j + 1) - k$ from last $(T, \mu)$ point
        else if Phase $j + 1$ disappeared then
            Run phb to follow boundary $j - (j + 2)$ from last $(T, \mu)$ point
        else
            continue
        end if
    end for
end for
```

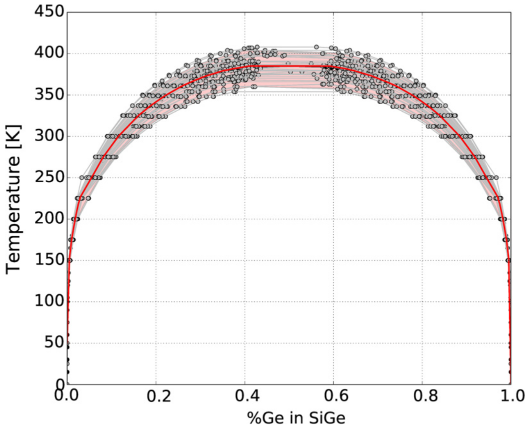


Fig. 9. Phase diagram of SiGe calculated using a minimum temperature step $\Delta T_{min} = 1$ K in Algorithm 5. The continuous red line is the median of the simulations and the shaded red region the 95% confidence interval. The phase boundary obtained using the ATAT calculated ECI overlaps the median of our model and therefore it is not visible. The gray circles are the (x, T) points of each of the realizations of the phase boundary obtained using Algorithm 5, and the lines interpolating between these points are shown for guidance. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Remark 15. Since different ECIs result in different 0 K ground states, the phase diagrams resulting from using different ECIs may have different phase boundaries. Therefore, the MC approach will not just result in a confidence interval for each phase boundary, but also in a probability for each phase boundary to be present.

phb takes as an input the temperature step, ΔT; however, in regions where the change in concentration Δx is large, this may result in missing the point where two-phase boundaries intersect. To avoid this, we call phb within an adaptive temperature loop which looks for changes in concentration larger that a given threshold as described in Algorithm 5.

As the program looks for phase transitions at fixed temperature steps, the natural way to obtain the uncertainty in the boundary is to look at the concentration spread for a fixed T. However, since we used an adaptive temperature step, we have different temperature points for each realization of the phase boundary, and therefore we use interpolation between the available data points for each phase boundary to construct the confidence interval at every temperature/concentration.

Fig. 9 shows an ensemble of phase diagrams for SiGe calculated using a minimum temperature step $\Delta T_{min} = 1$ K. 44 phase diagrams and linear interpolation between the calculated points in each of them were used to construct the confidence intervals. Since all realizations of the ECI predict only two 0 K ground states, the only result of the uncertainty is a confidence interval for a single phase boundary. In this particular case, the states below the lines correspond to a mixed phase of pure Si and pure Ge, since as we saw in the ground state line, those are the only thermodynamically stable states at 0 K. States above this line are disordered states, with random mixing of the Si and Ge atoms in the lattice.

7.1. Phase transition from disordered to two-phase coexistence

Another way to construct the phase diagram is to calculate the energy of the system at every (T, μ) point instead of following the phase boundary. Specifically, it may be of interest to study the behavior of an alloy at a given concentration as the temperature changes.

In this section, we calculate as an example the disorder to two-phase coexistence phase transition for the diamond SiGe alloy at 50% composition. Computationally, the phase transition is found by starting the alloy at a high temperature which for SiGe is taken to be 2000 K. The temperature has to be large enough for the entropic energy in the free energy to dominate, so that the system is in a disordered phase. Then, the temperature is gradually lowered while monitoring the specific heat at constant pressure [74],

$$C_p(T) = \beta^2 k_B \left\langle (U - \langle U \rangle)^2 \right\rangle, \tag{48}$$

where $\langle \cdot \rangle$ denotes an average, and U is the internal energy of the alloy. Lowering the temperature decreases the entropic energy allowing the configurational energy to become comparatively stronger and finally demanding a certain ordering of the atoms. In a two-phase coexistence the atoms do not want to mix on the lattice but stay separated into pure forms. A peak (divergence, in the limit of an infinite lattice) in the specific heat signals the phase transition [75].

Let us now discuss the numerical method used to compute the phase transition. We used an adaptive sequential Monte Carlo (ASMC) method [76,33] coupled via an in-house code to the MCMC library of ATAT [48,49]. The ASMC code works as follows. A set of N_p so-called particles, 64 in our examples, is started at a high temperature of 2000 K. Each particle is an initial state of the system: a $41 \times 41 \times 41$ Monte Carlo cell with periodic boundary conditions. Using double-spin-flip dynamics [48] they are independently evolved via the ASMC method to 50 K. Also, each particle has a normalized weight W_i such that the expected value of any function K of the configuration of the system can be approximated as [76]

$$\langle K \rangle \approx \sum_{i=1}^{N_p} W_i K_i, \tag{49}$$

where K_i is the value of the function obtained by particle i. The weights are also used to define the *effective sample size (ESS)* as $ESS = 1/\sum_i (W_i)^2$, which provides a measure of the population variance. $ESS = 1$ happens when a single particle i has weight $W_i = 1$ whereas the rest have zero weights. The other extreme, $ESS = N_p$, happens when all the particles have the same weight $W = 1/N_p$. The step size is adaptive and guided by the ESS, but we maximally allow a 10 K jump to ensure we do not miss a transition. Since the weight update can be obtained as a function of the temperature step [76], we can use fixed reduction factor for the ESS, ξ, to update to the next temperature by requiring $ESS(T + \Delta T) = \xi ESS(T)$. If at any step the ESS falls below a minimum threshold, ESS_{min}, then the particles are resampled using a multinomial distribution from the old samples according to their weights so that it is more likely to obtain replicas of the particles with larger weight [76]. At each step (including the initialization), 100 flips per lattice site were performed. The methodology to obtain the uncertainty in the phase transition temperature is summarized in Algorithm 6. For more details please refer to [76].

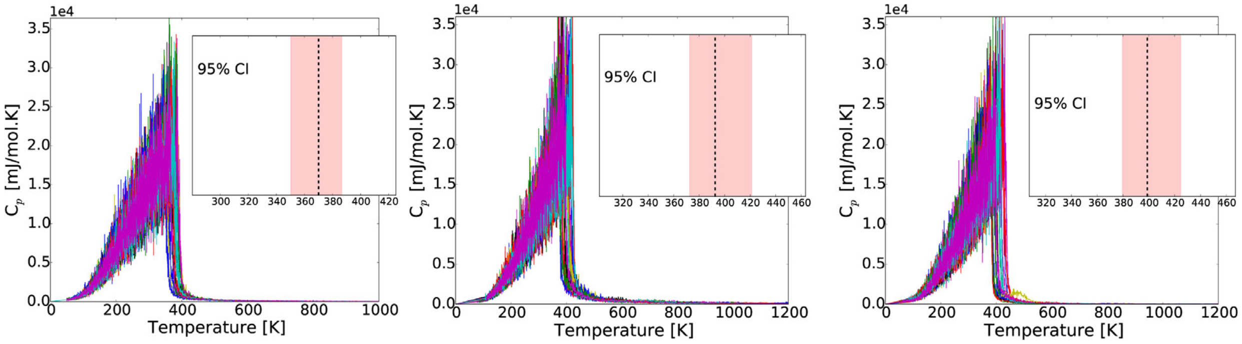
Remark 16. There are uncertainties associated with the number of particles and the seed used for the internal pseudo-random-number generator. If we allow this, we are not only capturing the uncertainties from the truncated cluster expansion and limited data but also from the particular implementation of the ASMC method. To avoid the ASMC uncertainties, we use the same number of particles with each ECI sample and the same seed.

Fig. 10(a) shows the uncertainty in the disorder to order transition by looking at the specific heat at constant pressure for 19 realizations of the temperature independent 0 K ECI. The inset shows the median transition temperature with a 95% confidence interval. As mentioned previously, due to the finite size of the system, the phase transition is signaled by a finite peak, so we take the temperature at the maximum of the specific heat as the phase transition temperature. Fig. 10(b) shows the same information but including the temperature dependence in the ECI and their uncertainty as described in Section 5. The median transition temperature increases from 370 K to ~393 K, which is very close to the value of 390 K obtained with ATAT using temperature dependent ECI. Regarding the uncertainty, the 95% confidence interval increases from 38 K (10%) to 49 K (12%) with the introduction of temperature dependence to the ECI. Even though there is an increase in the uncertainty, it remains at a similar level. These values of the uncertainty are moderately higher than those obtained in [33] using a different model to estimate the uncertainty. In that work, using temperature independent 0 K ECI the 95% confidence interval was of 20 K (6%). However, this difference could be explained by the low number of samples used in that work, where only five values of the transition temperature were used.

As described in Section 5.2.1, it is possible to reduce the computational cost substantially by using an approximation for the force constant tensor of all training configurations based on a fit of the dependence between bond stiffness and bond

Algorithm 6 Uncertainty in the disorder to two-phase coexistence transition temperature.

```
Given the posterior distribution of the ECI, possibly temperature dependent $p(\boldsymbol{\gamma} \mid \mathcal{D}_N, \boldsymbol{\alpha}_{MP}, \beta_{MP}, T)$, Eqs. (16)–(18)
Select the number of samples of the ECI, $N_s$
Select the initial and final temperatures, $T_i$ and $T_f$, and maximum temperature step $\Delta T_{max}$
Select the number of particles $N_p$
Select the size of simulation system
Select $ESS$ reduction factor $\xi$ and minimum allowed value, $ESS_{min}$
for $j = 1, \dots, N_s$ do
    Sample $\gamma_i$ from $p(\boldsymbol{\gamma} \mid \mathcal{D}_N, \boldsymbol{\alpha}_{MP}, \beta_{MP})$, Eqs. (16)–(18)
    {Adaptive Sequential Monte Carlo:}
    while $T > T_f$ do
        Given $W_i(T)$, compute $\Delta T$ such that $ESS(T + \Delta T) = \xi ESS(T)$
        Update particle weights $W_i$ for all particles
        if $ESS(T + \Delta T) < ESS_{min}$ then
            Resample particles
        end if
        $T \to T + \Delta T$
        for $i = 1, \dots, N_p$ do
            Equilibrate system using double-spin-flip dynamics
            Compute average energy $U_i$
        end for
        Approximate energy as $U \approx \sum_{i=1}^{N_p} W_i U_i$, Eq. (49)
        Compute $C_p(T) = \beta^2 k_B \left\langle (U - \langle U \rangle)^2 \right\rangle$, Eq. (48)
    end while
    Find transition temperature $T^j_{trans} = \text{argmax}_T \, C_p(T)$
end for
Gather statistics for $\{T^j_{trans}\}_{j=1}^{N_s}$
```



(a) (b) (c)

Fig. 10. Uncertainties in the disorder to two-phase coexistence of Si–Ge at 50% composition identified by a peak in the specific heat using (a) temperature independent 0 K ECI, (b) temperature dependent ECI using the full model and (c) temperature dependent ECI using the bond stiffness versus bond length approximation. 19 different runs are shown each in a different color. The inset summarizes the main figure via the posterior median as a black dotted vertical line surrounded by the 95% confidence interval as the shaded area.

length using only a subset of the training configurations. This approximation introduced an extra source of uncertainty from the fitting of the bond stiffness which is not intrinsic to the CE of the free energies we have used. In order to quantify the influence of this simplified modeling of the vibrational free energy, we also use it to carry out the calculation of the disorder to order transition temperature. In this case the vibrational free energy, F^{vib}, is a random variable and the calculation is done as described in Algorithm 3. Fig. 10(c) shows the uncertainty in the disorder to order transition by looking again at the specific heat at constant pressure. Compared to the calculations with T-dependent ECI using the full model, the median transition temperature increased from 393 K to 399 K, whereas the 95% confidence interval stayed at 49 K. There is a negligible increase in the uncertainty from using the approximate stiffness versus length model to calculate the force constant tensor in this system. There is, however, an increase of 6 K (1.5%) in the value of the predicted median transition temperature. This value is well within the uncertainty predicted by both simulations, and could be due to the limited number of samples used in the study.

8. Conclusions

We have presented a new approach based on machine learning using a Bayesian framework to obtain the uncertainty in thermodynamic properties of alloys based on the cluster expansion. In this framework, the cluster expansion coefficients are not point estimates, but random variables, so that the resulting expansion is also a random variable. The model accounts for uncertainty originating from limited data for the training and also from the inaccuracy of the truncated cluster expansion.

Since the cluster expansion is exact if an infinite number of basis functions is used, both of these uncertainties would vanish in the limit of infinite training points and basis functions.

The use of a relevance vector machine allowed us to obtain sparsity in the cluster expansion coefficients while providing an analytical solution for the predictive distribution of the cluster expanded quantity. This contrasts with previous work which required numerical sampling of the posterior distribution over the model parameters using an RJ-MCMC algorithm.

We have tested the framework on two binary alloys, SiGe and MgLi, using the formation energy as the training objective so that we could find the ground state line, i.e., the stable configurations and their energies. We have studied the accuracy of the model with the number of training points and basis functions and seen that the error from the limited number of points in each cluster decreases rapidly. The uncertainty quantification of the GSL energies showed that the cluster expansion is reliable to *ab initio* level accuracy even with limited data as shown by the uncertainties in the order of eV/atom. Regarding the selection of training data, we used the methodology implemented in ATAT, but this could be changed to a Bayesian method that fully takes into account the uncertainty in our model predictions [77]. The study of the error in the cluster expansion was extended to temperature dependent effective cluster interactions, and we showed that the error in the predicted free energy increases with the temperature. The same was seen in the uncertainty in order–disorder phase transition temperature for SiGe at 50% concentration, where the 95% confidence interval increased from 38 K (10%) to 49 K (12%) when including the temperature dependence of the effective cluster interactions. An attractive approach to save computational time is the bond stiffness versus bond length scheme, which approximates the force constant tensor using interpolation from a few selected configurations, thus saving an important amount of time but at the expense of increased uncertainty. To evaluate the impact on the uncertainty of the predictions, we carried out the simulations for the phase transition using this approximation, and found that the median transition temperature increased from 393 to 399 K, but the 95% confidence interval remained at 49 K. This means that in our test case, the approximation traded close to a 10-fold reduction in computational cost for a very small change in the median of the predictions of 1.5% and a negligible change in the uncertainty of the prediction.

The methodology presented in this work opens the gate to a new approach to the determination of alloy properties in the presence of uncertainty. As the formalism presented is completely general, it can be applied to the cluster expansion of any alloy property beyond what we have shown in this paper, such as elastic properties, energy band gaps or effective masses. It can also be the basis for applications to inverse design problems [1] in the presence of uncertainties, where the outcome would be a set of structures with the corresponding probabilities of satisfying the required thermodynamic constraint.

Acknowledgements

This work was supported by EPSRC (Grant No. EP/L027682/1). Additional support from the Computer Science and Mathematics Division of ORNL under the DARPA EQUiPS program is acknowledged. N.Z. thanks the Royal Society for support through a Wolfson Research Merit Award and the Technische Universität München, Institute for Advanced Study for support through a Hans Fisher Senior Fellowship, funded by the German Excellence Initiative and the European Union Seventh Framework Programme under Grant Agreement No. 291763. The work at Cornell was supported by 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-0809062).

Appendix A. Derivation of the posterior distribution over the model parameters

The posterior parameter distribution over the parameters of the model, $\boldsymbol{\gamma}$ is given by Eq. (16),

$$p(\boldsymbol{\gamma}\,|\,\mathcal{D}_N,\boldsymbol{\alpha},\beta)=\frac{\mathcal{L}(\mathcal{D}_N\,|\,\boldsymbol{\gamma},\boldsymbol{\beta})p(\boldsymbol{\gamma}\,|\,\boldsymbol{\alpha})}{\int\mathcal{L}(\mathcal{D}_N\,|\,\boldsymbol{\gamma},\boldsymbol{\alpha},\boldsymbol{\beta})p(\boldsymbol{\gamma}\,|\,\boldsymbol{\alpha})\,d\boldsymbol{\gamma}}. \tag{A.1}$$

Using Eqs. (10) and (12) it can be written as

$$p(\boldsymbol{\gamma}\,|\,\mathcal{D}_N,\boldsymbol{\alpha},\beta)=\frac{\mathcal{N}(\mathbf{t}_t\,|\,\boldsymbol{\Phi}\boldsymbol{\gamma},\beta^{-1}\mathbf{I})\mathcal{N}(\boldsymbol{\gamma}\,|\,0,\mathbf{A}^{-1})}{\int\mathcal{N}(\mathbf{t}_t\,|\,\boldsymbol{\Phi}\boldsymbol{\gamma},\beta^{-1}\mathbf{I})\mathcal{N}(\boldsymbol{\gamma}\,|\,0,\mathbf{A}^{-1})\,d\boldsymbol{\gamma}}=\frac{\mathcal{N}(\mathbf{t}_t\,|\,\boldsymbol{\Phi}\boldsymbol{\gamma},\beta^{-1}\mathbf{I})\mathcal{N}(\boldsymbol{\gamma}\,|\,0,\mathbf{A}^{-1})}{\mathcal{N}(\mathbf{t}_t\,|\,0,\beta^{-1}\mathbf{I}+\boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T)}, \tag{A.2}$$

where we have used the result [31]

$$\int\mathcal{N}(\mathbf{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1})\mathcal{N}(\mathbf{y}\,|\,\mathbf{A}\mathbf{x}+\mathbf{b},\mathbf{L}^{-1})\,d\mathbf{x}=\mathcal{N}(\mathbf{y}\,|\,\mathbf{A}\boldsymbol{\mu}+\mathbf{b},\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T). \tag{A.3}$$

to do the integration in the denominator. Using $\mathcal{N}(\mathbf{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1})=(2\pi)^{-N/2}\,|\boldsymbol{\Lambda}|^{1/2}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right]$,

$$p(\boldsymbol{\gamma}\,|\,\mathcal{D}_N,\boldsymbol{\alpha},\beta)=(2\pi)^{-M/2}\left(|\mathbf{A}|\,|\beta\mathbf{I}|\,\left|\beta^{-1}\mathbf{I}+\boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T\right|\right)^{1/2} \tag{A.4}$$

$$\exp\left\{-\frac{1}{2}\left[(\mathbf{t}_t-\boldsymbol{\Phi}\boldsymbol{\gamma})^T\beta\mathbf{I}(\mathbf{t}_t-\boldsymbol{\Phi}\boldsymbol{\gamma})+\boldsymbol{\gamma}^T\mathbf{A}\boldsymbol{\gamma}-\mathbf{t}_t^T(\beta^{-1}\mathbf{I}+\boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T)^{-1}\mathbf{t}_t\right]\right\}. \tag{A.5}$$

Rearranging the exponent, using the determinant identity [78]

$$\left|\mathbf{A} + \boldsymbol{\Phi}^T \beta \mathbf{I} \boldsymbol{\Phi}\right| = |\mathbf{A}|\, |\beta \mathbf{I}| \left|\beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T\right|, \tag{A.6}$$

and defining $\mathbf{S}_N^{-1} = \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \mathbf{A}$, we arrive at

$$p(\boldsymbol{\gamma} \,|\, \mathcal{D}_N, \boldsymbol{\alpha}, \beta) = (2\pi)^{-M/2} \left|\mathbf{S}_N^{-1}\right|^{1/2} \exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{\gamma} - \beta\mathbf{S}_N\boldsymbol{\Phi}^T\mathbf{t}_t\right)^T \mathbf{S}_N^{-1}\left(\boldsymbol{\gamma} - \beta\mathbf{S}_N\boldsymbol{\Phi}^T\mathbf{t}_t\right)\right]\right\}, \tag{A.7}$$

which is a normal distribution with covariance and mean given by

$$\mathbf{S}_N = \left(\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \mathbf{A}\right)^{-1}, \tag{A.8}$$

$$\boldsymbol{\mu} = \beta\mathbf{S}_N\boldsymbol{\Phi}^T\mathbf{t}_t. \tag{A.9}$$

Appendix B. Derivation of the evidence function

The marginal likelihood is the integration over the model parameters of the likelihood times the prior over the model parameters. Using Eqs. (10) and (12),

$$p(\mathcal{D}_N \,|\, \boldsymbol{\alpha}, \beta) = \int \mathcal{L}(\mathcal{D}_N \,|\, \boldsymbol{\gamma}, \beta)\, p(\boldsymbol{\gamma} \,|\, \boldsymbol{\alpha})\, d\boldsymbol{\gamma} = \int \mathcal{N}(\mathbf{t}_t \,|\, \boldsymbol{\Phi}\boldsymbol{\gamma}, \beta^{-1}\mathbf{I})\mathcal{N}(\boldsymbol{\gamma} \,|\, 0, \mathbf{A}^{-1})\, d\boldsymbol{\gamma} = \mathcal{N}(\mathbf{t}_t \,|\, 0, \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T), \tag{B.1}$$

where $\mathbf{A} = \mathrm{diag}(\alpha_i)$ and we have used the identity in Eq. (A.3).

As seen before, the covariance of the marginal likelihood can be decomposed in a summation where each term depends on only one of the hyperparameters,

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T = \beta^{-1}\mathbf{I} + \sum_{i=0}^{M-1} \alpha_i^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T, \tag{B.2}$$

where $\boldsymbol{\varphi}_i$ is the i-th column of the design matrix $\boldsymbol{\Phi}$. For each basis function i, the matrix $\mathbf{C}$ can be decomposed as

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \sum_{j \neq i} \alpha_j^{-1}\boldsymbol{\varphi}_j\boldsymbol{\varphi}_j^T + \alpha_i^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T = \mathbf{C}_{\setminus i} + \alpha_i^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T, \tag{B.3}$$

which, in turn, will lead to a decomposition of the marginal likelihood where we can isolate the dependence on α_i. To do this, we use the *matrix determinant lemma* [31],

$$|\mathbf{C}| = \left|\mathbf{C}_{\setminus i}\right| \left|1 + \alpha_i^{-1}\boldsymbol{\varphi}_i^T\mathbf{C}_{\setminus i}^{-1}\boldsymbol{\varphi}_i\right|, \tag{B.4}$$

and the *Sherman–Morrison formula* [31],

$$\mathbf{C}^{-1} = \mathbf{C}_{\setminus i}^{-1} - \frac{\mathbf{C}_{\setminus i}^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T\mathbf{C}_{\setminus i}^{-1}}{\alpha_i + \boldsymbol{\varphi}_i^T\mathbf{C}_{\setminus i}^{-1}\boldsymbol{\varphi}_i}. \tag{B.5}$$

With Eqs. (B.4) and (B.5), we can factorize the marginal likelihood as

$$p(\mathcal{D}_N \,|\, \boldsymbol{\alpha}, \beta) = \mathcal{N}(\mathbf{t}_t \,|\, 0, \mathbf{C}_{\setminus i}) \left|1 + \alpha_i^{-1}\boldsymbol{\varphi}_i^T\mathbf{C}_{\setminus i}^{-1}\boldsymbol{\varphi}_i\right|^{-1/2} \exp\left(\frac{1}{2}\mathbf{t}_t^T \frac{\mathbf{C}_{\setminus i}^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T\mathbf{C}_{\setminus i}^{-1}}{\alpha_i + \boldsymbol{\varphi}_i^T\mathbf{C}_{\setminus i}^{-1}\boldsymbol{\varphi}_i}\mathbf{t}_t\right). \tag{B.6}$$

For convenience, we define the *sparsity* s_i and *quality* q_i for every vector $\boldsymbol{\varphi}_i$ [40,31],

$$s_i = \boldsymbol{\varphi}_i^T\mathbf{C}_{\setminus i}^{-1}\boldsymbol{\varphi}_i, \tag{B.7}$$

$$q_i = \boldsymbol{\varphi}_i^T\mathbf{C}_{\setminus i}^{-1}\mathbf{t}_t. \tag{B.8}$$

Using these definitions, the marginal likelihood can be written as

$$p(\mathcal{D}_N \,|\, \boldsymbol{\alpha}, \beta) = \mathcal{N}(\mathbf{t}_t \,|\, 0, \mathbf{C}_{\setminus i}) \left|1 + \alpha_i^{-1}s_i\right|^{-1/2} \exp\left(\frac{1}{2}\frac{q_i^2}{\alpha_i + s_i}\right). \tag{B.9}$$

Taking the logarithm of this function, we obtain the evidence function, whose maximization is equivalent to that of the marginal likelihood

$$\mathcal{E}(\boldsymbol{\alpha}, \beta) = \log \mathcal{N}(\mathbf{t}_t \,|\, 0, \mathbf{C}_{\backslash i}) + \frac{1}{2}\left[\log \alpha_i - \log(\alpha_i + s_i) + \frac{q_i^2}{\alpha_i + s_i}\right]. \tag{B.10}$$

To find the maximum of the evidence function with respect to α_i, we only need to worry about the second term, as we removed all dependence on α_i from $\mathbf{C}_{\backslash i}$:

$$\frac{\partial \mathcal{E}}{\partial \alpha_i} = \frac{1}{2}\frac{\alpha_i^{-1} s_i^2 - (q_i^2 - s_i)}{(\alpha_i + s_i)^2}. \tag{B.11}$$

By equating this derivative to zero, we find a stationary point which can be shown to be a maximum [40]. Recalling that $\alpha_i \geq 0$, we find the maximum as

$$\alpha_i^* = \begin{cases} \infty & \text{if } q_i^2 - s_i \leq 0, \\ \frac{s_i^2}{q_i^2 - s_i} & \text{if } q_i^2 - s_i > 0. \end{cases} \tag{B.12}$$

Even though we have a closed form for the maximum with respect to α_i, it depends on the value of the rest of the hyperparameters, so the maximization of the marginal likelihood still requires an iterative process as detailed in Algorithm 1.

Even though we have found an expression for α_i^*, its computation in this form is not efficient as it requires calculating $\mathbf{C}_{\backslash i}^{-1}\boldsymbol{\varphi}_i$ and $\mathbf{C}_{\backslash i}^{-1}\mathbf{t}$ for every basis function i at every iteration of the optimization A way to speed up this calculation is to introduce two new quantities, S_i and Q_i,

$$S_i = \boldsymbol{\varphi}_i^T \mathbf{C}^{-1} \boldsymbol{\varphi}_i, \tag{B.13}$$

$$Q_i = \boldsymbol{\varphi}_i^T \mathbf{C}^{-1} \mathbf{t}_t, \tag{B.14}$$

which are related to s_i and q_i as

$$s_i = \alpha_i S_i / (\alpha_i - S_i), \tag{B.15}$$

$$q_i = \alpha_i Q_i / (\alpha_i - S_i). \tag{B.16}$$

Using S_i and Q_i, we can obtain s_i and q_i with a single factorization of the matrix $\mathbf{C}$ per iteration. A further optimization is possible rewriting $\mathbf{C}^{-1}$ using the *Woodbury matrix identity* [31],

$$\mathbf{C}^{-1} = \left(\beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T\right)^{-1} = \beta\mathbf{I} - \beta^2 \boldsymbol{\Phi}\Sigma\boldsymbol{\Phi}^T, \tag{B.17}$$

where Σ is the posterior variance as defined in Eq. (17), which will have the size of the number of active basis (those with $\alpha_i \neq \infty$) at the given iteration.

To find the maximum with respect to β, we follow the same procedure and find the stationary point of $\partial\mathcal{E}/\partial\beta$. In this case, it can be shown that [37]

$$\frac{\partial \mathcal{E}}{\partial \beta} = \frac{1}{2}\left[\frac{N}{\beta} - \text{trace}(\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{\Phi}) - (\mathbf{t}_t - \boldsymbol{\Phi}\boldsymbol{\mu})^T(\mathbf{t}_t - \boldsymbol{\Phi}\boldsymbol{\mu})\right] = \frac{1}{2}\left[\frac{N}{\beta} - \beta^{-1}\text{trace}(\mathbf{I} - \mathbf{A}\boldsymbol{\Sigma}) - (\mathbf{t}_t - \boldsymbol{\Phi}\boldsymbol{\mu})^T(\mathbf{t}_t - \boldsymbol{\Phi}\boldsymbol{\mu})\right], \tag{B.18}$$

where we have used the definition of the posterior covariance, Eq. (17). By equating Eq. (B.18) to zero, we find

$$\frac{1}{\beta} = \frac{(\mathbf{t}_t - \boldsymbol{\Phi}\boldsymbol{\mu})^T(\mathbf{t}_t - \boldsymbol{\Phi}\boldsymbol{\mu})}{N - \text{trace}(\mathbf{I} - \mathbf{A}\boldsymbol{\Sigma})}. \tag{B.19}$$

Since both the posterior mean, Eq. (18), and covariance, Eq. (17), depend on β, this equation defines an iterative update. Therefore, at iteration i, we would update β as

$$\frac{1}{\beta^{i+i}} = \frac{(\mathbf{t}_t - \boldsymbol{\Phi}\boldsymbol{\mu}^i)^T(\mathbf{t}_t - \boldsymbol{\Phi}\boldsymbol{\mu}^i)}{N - \text{trace}(\mathbf{I} - \mathbf{A}\boldsymbol{\Sigma}^i)}. \tag{B.20}$$

Appendix C. Phonons in a crystal

In the harmonic approximation, the total energy of a periodic crystal under small lattice distortions from the equilibrium is [70,79]

$$H=\frac{1}{2}\sum_{i} M_i\left(\frac{d\mathbf{u}_i}{dt}\right)^2+\frac{1}{2}\sum_{i,j,\alpha,\beta} u_{i\alpha}B_{\alpha\beta}(i,j)u_{j\beta}, \tag{C.1}$$

where $\mathbf{u}_i=(u_{ix}\,u_{iy}\,u_{iz})$ is the atomic displacement of atom i respect to its equilibrium position, α and β run over the Cartesian components of the vector, i and j run over the atoms of the lattice and we have defined the force constant tensor $\mathbf{B}(i,j)$ with components given by

$$B_{\alpha\beta}(i,j)=\frac{\partial^2 V}{\partial u_{i\alpha}\partial u_{j\beta}}, \tag{C.2}$$

where V is the potential energy of the system here approximated using a quadratic (harmonic) expansion. In a periodic system, we can specify the indexing of each point in the system i with the pair $(\mathbf{m},a)$, where $\mathbf{m}$ specifies the unit cell and a the atom of the unit cell. Using this notation, the force constant tensor becomes $\mathbf{B}(\mathbf{n},b,\mathbf{m},a)$. This tensor can be used to obtain the force on the atom $(\mathbf{n},b)$ when a displacement $\mathbf{u}_{\mathbf{m},a}$ is applied to atom $(\mathbf{m},a)$ [79],

$$\mathbf{F}(\mathbf{n},b)=\mathbf{B}(\mathbf{n},b,\mathbf{m},a)\mathbf{u}_{\mathbf{m},a}. \tag{C.3}$$

If the same displacement $\mathbf{u}_{0,a}$ is applied to atom a in all the copies of the supercell, we can write:

$$\mathbf{F}(\mathbf{n},b)=\sum_{\mathbf{m}}\mathbf{B}(\mathbf{n},b,\mathbf{m},a)\mathbf{u}_{0,a}. \tag{C.4}$$

At this point we can introduce a change of variables $\mathbf{u}_{\mathbf{m},a}\to\mathbf{v}_{\mathbf{m},a}=\sqrt{M_a}\mathbf{u}_{\mathbf{m},a}$ before proceeding any further to factor the possibly different masses of a and b into the term with the force constants only. The energy of the system then becomes

$$H=\frac{1}{2}\sum_{\mathbf{m},a}\left(\frac{d\mathbf{v}_{\mathbf{m},a}}{dt}\right)^2+\frac{1}{2}\sum_{\mathbf{m},a,\mathbf{n},b,\alpha,\beta} v_{\mathbf{m},a,\alpha}\frac{B_{\alpha\beta}(\mathbf{m},a,\mathbf{n},b)}{\sqrt{M_aM_b}}v_{\mathbf{n},b,\beta}=K+\frac{1}{2}\sum_{\mathbf{m},a,\mathbf{n},b,\alpha,\beta} v_{\mathbf{m},a,\alpha}\frac{B_{\alpha\beta}(\mathbf{m},a,\mathbf{n},b)}{\sqrt{M_aM_b}}v_{\mathbf{n},b,\beta}, \tag{C.5}$$

where we have defined the kinetic part of the Hamiltonian as K. Using Bloch's theorem, we now do a further change of variables to *normal coordinates* [79],

$$\mathbf{e}_{\mathbf{k},a}=\frac{1}{\sqrt{NV}}\sum_{\mathbf{m}}\mathbf{v}_{\mathbf{m},a}e^{2\pi i\mathbf{k}\mathbf{m}}, \tag{C.6}$$

where $\mathbf{k}$ is a point in the first Brillouin zone and V is the volume of the system. It can be inverted using Fourier's inversion theorem as

$$\mathbf{v}_{\mathbf{m},a}=\frac{1}{\sqrt{NV}}\sum_{\mathbf{k}}\mathbf{e}_{\mathbf{k},a}e^{-2\pi i\mathbf{k}\mathbf{m}}. \tag{C.7}$$

Substituting this back into Eq. (C.5),

$$\begin{aligned}H&=K+\frac{1}{2NV}\sum_{\mathbf{k},\mathbf{k}'}\sum_{\mathbf{m},a,\mathbf{n},b}\mathbf{e}^T_{\mathbf{k},a}\frac{\mathbf{B}(\mathbf{m},a,\mathbf{n},b)}{\sqrt{M_aM_b}}\mathbf{e}_{\mathbf{k}',b}e^{-2\pi i\mathbf{k}\mathbf{m}}e^{-2\pi i\mathbf{k}'\mathbf{n}}\\&=K+\frac{1}{2NV}\sum_{\mathbf{k},\mathbf{k}',a,b}\mathbf{e}^T_{\mathbf{k},a}\left(\sum_{\mathbf{m},\mathbf{n}}\frac{\mathbf{B}(\mathbf{m},a,\mathbf{n},b)}{\sqrt{M_aM_b}}e^{2\pi i\mathbf{k}'(\mathbf{m}-\mathbf{n})}e^{-2\pi i(\mathbf{k}+\mathbf{k}')\mathbf{m}}\right)\mathbf{e}_{\mathbf{k}',b}.\end{aligned} \tag{C.8}$$

Because of the translational invariance of the system, the force constants cannot depend on the absolute value of $\mathbf{m}$ and $\mathbf{n}$, but only on their difference $\mathbf{h}=\mathbf{m}-\mathbf{n}$ [79]. Using this we have,

$$\begin{aligned}H&=K+\frac{1}{2NV}\sum_{\mathbf{k},\mathbf{k}',a,b}\mathbf{e}^T_{\mathbf{k},a}\left(\sum_{\mathbf{h}}\frac{\mathbf{B}(a,b,\mathbf{h})}{\sqrt{M_aM_b}}e^{2\pi i\mathbf{k}'(\mathbf{h})}\left(\sum_{\mathbf{m}}e^{-2\pi i(\mathbf{k}+\mathbf{k}')\mathbf{m}}\right)\right)\mathbf{e}_{\mathbf{k}',b}\\&=K+\frac{1}{2}\sum_{\mathbf{k},a,b}\mathbf{e}^T_{\mathbf{k},a}\left(\sum_{\mathbf{h}}\frac{\mathbf{B}(a,b,\mathbf{h})}{\sqrt{M_aM_b}}e^{-2\pi i\mathbf{k}\mathbf{h}}\right)\mathbf{e}_{-\mathbf{k},b}.\end{aligned} \tag{C.9}$$

The term in brackets in the last line can be used as the definition of the *dynamical matrix* $\mathbf{D}(\mathbf{k}, ab)$,

$$\mathbf{D}(\mathbf{k}, ab) = \sum_{\mathbf{h}} \frac{\mathbf{B}(a, b, \mathbf{h})}{\sqrt{M_a M_b}} e^{-2\pi i \mathbf{kh}} = \sum_{\mathbf{m}} \frac{\mathbf{B}(\mathbf{m}, a, 0, b)}{\sqrt{M_a M_b}} e^{-2\pi i \mathbf{km}}. \tag{C.10}$$

Using this definition, we can write

$$H = K + \sum_{\mathbf{k}} \frac{1}{2} \sum_{a,b} \mathbf{e}^T_{\mathbf{k},a} \mathbf{D}(\mathbf{k}, ab) \mathbf{e}_{-\mathbf{k},b} = \sum_{\mathbf{k}} \left[K(\mathbf{k}) + \frac{1}{2} \sum_{a,b} \mathbf{e}^T_{\mathbf{k},a} \mathbf{D}(\mathbf{k}, ab) \mathbf{e}_{-\mathbf{k},b} \right] = \sum_{\mathbf{k}} H(\mathbf{k}). \tag{C.11}$$

For each value of $\mathbf{k}$ we have a system of $3n$ particles, for n the number of particles in the unit cell:

$$H(\mathbf{k}) = K(\mathbf{k}) + \frac{1}{2} \sum_{a,b} \mathbf{e}^T_{\mathbf{k},a} \mathbf{D}(\mathbf{k}, ab) \mathbf{e}_{-\mathbf{k},b}. \tag{C.12}$$

These particles are coupled through the dynamical matrix, so we want to diagonalize it to find the normal modes for each $\mathbf{k}$. This is the Hamiltonian of a classical oscillator, so the eigenvalues of $\mathbf{D}(\mathbf{k})$ will give us the frequencies of the phonons for each value of $\mathbf{k}$, $\omega(\mathbf{k})$ [70,79],

$$\mathbf{D}(\mathbf{k})\mathbf{w}(\mathbf{k}) = [\omega(\mathbf{k})]^2 \mathbf{w}(\mathbf{k}). \tag{C.13}$$

References

[1] A. Franceschetti, A. Zunger, The inverse band-structure problem of finding an atomic configuration with given electronic properties, Nature 402 (6757) (1999) 60–63, http://dx.doi.org/10.1038/46995.

[2] P. Piquini, P.A. Graf, A. Zunger, Band-gap design of quaternary (In, Ga) (As, Sb) semiconductors via the inverse-band-structure approach, Phys. Rev. Lett. 100 (18) (2008) 186403, http://dx.doi.org/10.1103/PhysRevLett.100.186403.

[3] Y.-Y. Zhang, W. Gao, S. Chen, H. Xiang, X.-G. Gong, Inverse design of materials by multi-objective differential evolution, Comput. Mater. Sci. 98 (2015) 51–55, http://dx.doi.org/10.1016/j.commatsci.2014.10.054.

[4] Y. Chen, J. Xi, D.O. Dumcenco, Z. Liu, K. Suenaga, D. Wang, Z. Shuai, Y.-S. Huang, L. Xie, Tunable band gap photoluminescence from atomically thin transition-metal dichalcogenide alloys, ACS Nano 7 (5) (2013) 4610–4616, http://dx.doi.org/10.1021/nn401420h.

[5] A. Kutana, E.S. Penev, B.I. Yakobson, Engineering electronic properties of layered transition-metal dichalcogenide compounds through alloying, Nanoscale 6 (11) (2014) 5820–5825, http://dx.doi.org/10.1039/c4nr00177j.

[6] H. Li, X. Duan, X. Wu, X. Zhuang, H. Zhou, Q. Zhang, X. Zhu, W. Hu, P. Ren, P. Guo, L. Ma, X. Fan, X. Wang, J. Xu, A. Pan, X. Duan, Growth of alloy $MoS_{2x}Se_{2(1-x)}$ nanosheets with fully tunable chemical compositions and optical properties, J. Am. Chem. Soc. 136 (10) (2014) 3756–3759, http://dx.doi.org/10.1021/ja500069b.

[7] J. Xi, T. Zhao, D. Wang, Z. Shuai, Tunable electronic properties of two-dimensional transition metal dichalcogenide alloys: a first-principles prediction, J. Phys. Chem. Lett. 5 (2) (2014) 285–291, http://dx.doi.org/10.1021/jz402375s.

[8] B.N. Pantha, R. Dahal, J. Li, J.Y. Lin, H.X. Jiang, G. Pomrenke, Thermoelectric properties of $In_xGa_{1-x}N$ alloys, Appl. Phys. Lett. 92 (4) (2008) 042112, http://dx.doi.org/10.1063/1.2839309.

[9] H. Goldsmid, Bismuth telluride and its alloys as materials for thermoelectric generation, Materials 7 (4) (2014) 2577–2592, http://dx.doi.org/10.3390/ma7042577.

[10] S. Bhattacharya, G.K.H. Madsen, High-throughput exploration of alloying as design strategy for thermoelectrics, Phys. Rev. B 92 (8) (2015) 085205, http://dx.doi.org/10.1103/PhysRevB.92.085205.

[11] E. Kobayashi, S. Matsumoto, H. Doi, T. Yoneyama, H. Hamanaka, Mechanical properties of the binary titanium–zirconium alloys and their potential for biomedical materials, J. Biomed. Mater. Res. 29 (8) (1995) 943–950, http://dx.doi.org/10.1002/jbm.820290805.

[12] W.A. Counts, M. Friák, D. Raabe, J. Neugebauer, Using ab initio calculations in designing bcc MgLi–X alloys for ultra-lightweight applications, Adv. Eng. Mater. 12 (12) (2010) 1198–1205, http://dx.doi.org/10.1002/adem.201000225.

[13] X. Song, L. You, B. Zhang, A. Song, Design of low elastic modulus Ti–Nb–Zr alloys for implant materials, in: Advanced Materials and Processing 2010, World Scientific, 2011, pp. 334–338.

[14] S.-H. Wei, A. Zunger, Predicted band-gap pressure coefficients of all diamond and zinc-blende semiconductors: chemical trends, Phys. Rev. B 60 (8) (1999) 5404–5411, http://dx.doi.org/10.1103/PhysRevB.60.5404.

[15] J. Kang, S. Tongay, J. Li, J. Wu, Monolayer semiconducting transition metal dichalcogenide alloys: stability and band bowing, J. Appl. Phys. 113 (14) (2013) 143703, http://dx.doi.org/10.1063/1.4799126.

[16] S. Froyen, D.M. Wood, A. Zunger, New optical transitions in strained Si–Ge superlattices, Phys. Rev. B 36 (8) (1987) 4547–4550, http://dx.doi.org/10.1103/PhysRevB.36.4547.

[17] T. Pearsall, J. Vandenberg, R. Hull, J. Bonar, Structure and optical properties of strained Ge–Si superlattices grown on (001) Ge, Phys. Rev. Lett. 63 (19) (1989) 2104–2107, http://dx.doi.org/10.1103/PhysRevLett.63.2104.

[18] C. Tserbak, H.M. Polatoglou, G. Theodorou, $(Si)_3/(Ge)_4$ superlattices: direct-gap semiconductors?, Europhys. Lett. 18 (5) (1992) 451–456, http://dx.doi.org/10.1209/0295-5075/18/5/013.

[19] M. D'Avezac, J.-W. Luo, T. Chanier, A. Zunger, Genetic-algorithm discovery of a direct-gap and optically allowed superstructure from indirect-gap Si and Ge semiconductors, Phys. Rev. Lett. 108 (2) (2012) 027401, http://dx.doi.org/10.1103/PhysRevLett.108.027401.

[20] J. Feng, R.G. Hennig, N.W. Ashcroft, R. Hoffmann, Emergent reduction of electronic state dimensionality in dense ordered Li–Be alloys, Nature 451 (7177) (2008) 445–448, http://dx.doi.org/10.1038/nature06442.

[21] G. Schusteritsch, S.P. Hepplestone, C.J. Pickard, First-principles structure determination of interface materials: the Ni_xInAs nickelides, Phys. Rev. B 92 (5) (2015) 054105, http://dx.doi.org/10.1103/PhysRevB.92.054105.

[22] V. Blum, A. Zunger, Mixed-basis cluster expansion for thermodynamics of bcc alloys, Phys. Rev. B 70 (15) (2004) 155108, http://dx.doi.org/10.1103/PhysRevB.70.155108.

[23] R.H. Taylor, S. Curtarolo, G.L.W. Hart, Ordered magnesium–lithium alloys: first-principles predictions, Phys. Rev. B 81 (2) (2010) 024112, http://dx.doi.org/10.1103/PhysRevB.81.024112.

[24] J.M. Sanchez, Cluster expansions and the configurational energy of alloys, Phys. Rev. B 48 (18) (1993) 14013–14015, http://dx.doi.org/10.1103/PhysRevB.48.14013.

[25] M.H.F. Sluiter, Y. Kawazoe, Cluster expansion method for adsorption: application to hydrogen chemisorption on graphene, Phys. Rev. B 68 (8) (2003) 085410, http://dx.doi.org/10.1103/PhysRevB.68.085410.

[26] N.A. Zarkevich, D.D. Johnson, Reliable first-principles alloy thermodynamics via truncated cluster expansions, Phys. Rev. Lett. 92 (25 Pt 1) (2004) 255702, http://dx.doi.org/10.1103/PhysRevLett.92.255702.

[27] J.M. Sanchez, Cluster expansion and the configurational theory of alloys, Phys. Rev. B 81 (22) (2010) 224202, http://dx.doi.org/10.1103/PhysRevB.81.224202.

[28] A. Van der Ven, G. Ceder, Vacancies in ordered and disordered binary alloys treated with the cluster expansion, Phys. Rev. B 71 (5) (2005) 054102, http://dx.doi.org/10.1103/PhysRevB.71.054102.

[29] M.Y. Lavrentiev, D. Nguyen-Manh, S.L. Dudarev, Magnetic cluster expansion model for bcc–fcc transitions in Fe and Fe–Cr alloys, Phys. Rev. B 81 (18) (2010) 184202, http://dx.doi.org/10.1103/PhysRevB.81.184202.

[30] J. Kristensen, N.J. Zabaras, Predicting low-thermal-conductivity Si–Ge nanowires with a modified cluster expansion method, Phys. Rev. B 91 (5) (2015) 054105, http://dx.doi.org/10.1103/PhysRevB.91.054105.

[31] Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006, http://www.springer.com/us/book/9780387310732.

[32] D. Barber, Bayesian Reasoning and Machine Learning, Cambridge University Press, 2012.

[33] J. Kristensen, N.J. Zabaras, Bayesian uncertainty quantification in the evaluation of alloy properties with the cluster expansion method, Comput. Phys. Commun. 185 (11) (2014) 2885–2892, http://dx.doi.org/10.1016/j.cpc.2014.07.013.

[34] P.J. Green, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, Biometrika 82 (4) (1995) 711–732, http://dx.doi.org/10.1093/biomet/82.4.711.

[35] T. Mueller, G. Ceder, Bayesian approach to cluster expansions, Phys. Rev. B 80 (2) (2009) 024103, http://dx.doi.org/10.1103/PhysRevB.80.024103.

[36] L.J. Nelson, V. Ozoliņš, C.S. Reese, F. Zhou, G.L.W. Hart, Cluster expansion made easy with Bayesian compressive sensing, Phys. Rev. B 88 (2013) 155105, http://dx.doi.org/10.1103/PhysRevB.88.155105, http://link.aps.org/doi/10.1103/PhysRevB.88.155105.

[37] M.E. Tipping, Sparse Bayesian learning and the relevance vector machine, J. Mach. Learn. Res. 1 (2001) 211–244.

[38] D.B. Laks, L.G. Ferreira, S. Froyen, A. Zunger, Efficient cluster expansion for substitutional systems, Phys. Rev. B 46 (19) (1992) 12587–12605, http://dx.doi.org/10.1103/PhysRevB.46.12587.

[39] A. van de Walle, A complete representation of structure–property relationships in crystals, Nat. Mater. 7 (2008) 455–458, http://dx.doi.org/10.1038/nmat2200.

[40] A.C. Faul, M.E. Tipping, Analysis of sparse Bayesian learning, in: T.G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems, MIT Press, 2002, pp. 383–389.

[41] M.V. Fischetti, S.E. Laux, Band structure, deformation potentials, and carrier mobility in strained Si, Ge, and SiGe alloys, J. Appl. Phys. 80 (4) (1996) 2234, http://dx.doi.org/10.1063/1.363052.

[42] S. Gupta, V. Moroz, L. Smith, Q. Lu, K.C. Saraswat, 7-nm FinFET CMOS design enabled by stress engineering using Si, Ge, and Sn, IEEE Trans. Electron Devices 61 (5) (2014) 1222–1230, http://dx.doi.org/10.1109/TED.2014.2311129.

[43] D.J. Paul, Si/SiGe heterostructures: from material and physics to devices and circuits, Semicond. Sci. Technol. 19 (10) (2004) R75–R108, http://dx.doi.org/10.1088/0268-1242/19/10/R02.

[44] P. Chaisakul, D. Marris-Morini, J. Frigerio, D. Chrastina, M.-S. Rouifed, S. Cecchi, P. Crozat, G. Isella, L. Vivien, Integrated germanium optical interconnects on silicon substrates, Nat. Photonics 8 (6) (2014) 482–488, http://dx.doi.org/10.1038/nphoton.2014.73.

[45] G. Joshi, H. Lee, Y. Lan, X. Wang, G. Zhu, D. Wang, R.W. Gould, D.C. Cuff, M.Y. Tang, M.S. Dresselhaus, G. Chen, Z. Ren, Enhanced thermoelectric figure-of-merit in nanostructured p-type silicon germanium bulk alloys, Nano Lett. 8 (12) (2008) 4670–4674, http://dx.doi.org/10.1021/nl8026795.

[46] A. Samarelli, L. Ferre Llin, S. Cecchi, J. Frigerio, D. Chrastina, G. Isella, E. Müller Gubler, T. Etzelstorfer, J. Stangl, Y. Zhang, J. Weaver, P. Dobson, D. Paul, Prospects for SiGe thermoelectric generators, Solid-State Electron. 98 (2014) 70–74, http://dx.doi.org/10.1016/j.sse.2014.04.003.

[47] S. Kudela, Magnesium–lithium matrix composites – an overview, Int. J. Mater. Prod. Technol. 18 (1–3) (2003) 91–115, http://dx.doi.org/10.1504/IJMPT.2003.003587.

[48] A. van de Walle, M. Asta, G. Ceder, The alloy theoretic automated toolkit: a user guide, Calphad-Comput. Coupling Ph. Diagrams Thermochem. 26 (4) (2002) 539–553, http://dx.doi.org/10.1016/S0364-5916(02)80006-2, arXiv:cond-mat/0212159.

[49] A. van de Walle, G. Ceder, Automating first-principles phase diagram calculations, J. Phase Equilib. 23 (4) (2002) 348–359, http://dx.doi.org/10.1361/105497102770331596.

[50] G. Kresse, J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B 54 (16) (1996) 11169–11186, http://dx.doi.org/10.1103/PhysRevB.54.11169.

[51] G. Kresse, J. Furthmüller, Efficiency of *ab-initio* total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci. 6 (1) (1996) 15–50, http://dx.doi.org/10.1016/0927-0256(96)00008-0.

[52] J. Kristensen, I. Bilionis, N.J. Zabaras, Relative entropy as model selection tool in cluster expansions, Phys. Rev. B 87 (17) (2013) 174112, http://dx.doi.org/10.1103/PhysRevB.87.174112.

[53] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (18) (1996) 3865–3868, http://dx.doi.org/10.1103/PhysRevLett.77.3865.

[54] H.J. Monkhorst, J.D. Pack, Special points for Brillouin-zone integrations, Phys. Rev. B 13 (12) (1976) 5188–5192, http://dx.doi.org/10.1103/PhysRevB.13.5188.

[55] J. Sacks, W.J. Welch, T.J. Mitchell, H.P. Wynn, Design and analysis of computer experiments, Stat. Sci. 4 (4) (1989) 409–423.

[56] D.A. Cohn, Z. Ghahramani, M.I. Jordan, Active learning with statistical models, J. Artif. Intell. Res. 4 (1996) 129–145, http://dx.doi.org/10.1613/jair.295.

[57] L. Ferreira, S.-H. Wei, A. Zunger, Stability, electronic structure, and phase diagrams of novel inter-semiconductor compounds, Int. J. High Perform. Comput. Appl. 5 (1) (1991) 34–56, http://dx.doi.org/10.1177/109434209100500103.

[58] C. Currin, T. Mitchell, M. Morris, D. Ylvisaker, Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments, J. Am. Stat. Assoc. 86 (416) (1991) 953–963.

[59] J. Snoek, H. Larochelle, R.P. Adams, Practical Bayesian optimization of machine learning algorithms, in: Advances in Neural Information Processing Systems, 2012, pp. 2951–2959.

[60] B. Grabowski, T. Hickel, J. Neugebauer, *Ab initio* study of the thermodynamic properties of nonmagnetic elementary fcc metals: exchange–correlation-related error bars and chemical trends, Phys. Rev. B 76 (2) (2007) 024309, http://dx.doi.org/10.1103/PhysRevB.76.024309.

[61] F. Körmann, A. Dick, B. Grabowski, B. Hallstedt, T. Hickel, J. Neugebauer, Free energy of bcc iron: integrated ab initio derivation of vibrational, electronic, and magnetic contributions, Phys. Rev. B 78 (3) (2008) 033102, http://dx.doi.org/10.1103/PhysRevB.78.033102.

[62] N.D. Mermin, Thermal properties of the inhomogeneous electron gas, Phys. Rev. 137 (5A) (1965) A1441–A1443, http://dx.doi.org/10.1103/PhysRev.137.A1441.

[63] D.M.C. Nicholson, G.M. Stocks, Y. Wang, W.A. Shelton, Z. Szotek, W.M. Temmerman, Stationary nature of the density-functional free energy: application to accelerated multiple-scattering calculations, Phys. Rev. B 50 (19) (1994) 14686–14689, http://dx.doi.org/10.1103/PhysRevB.50.14686.

[64] G. Kresse, J. Furthmüller, J. Hafner, Ab initio force constant approach to phonon dispersion relations of diamond and graphite, Europhys. Lett. 32 (9) (1995) 729–734, http://dx.doi.org/10.1209/0295-5075/32/9/005.

[65] K. Parlinski, Z.Q. Li, Y. Kawazoe, First-principles determination of the soft mode in cubic ZrO_2, Phys. Rev. Lett. 78 (21) (1997) 4063–4066, http://dx.doi.org/10.1103/PhysRevLett.78.4063.

[66] A. van de Walle, G. Ceder, The effect of lattice vibrations on substitutional alloy thermodynamics, Rev. Mod. Phys. 74 (1) (2002) 11–45, http://dx.doi.org/10.1103/RevModPhys.74.11.

[67] N.W. Ashcroft, N.D. Mermin, Solid State Physics, Brooks/Cole, 1976.

[68] C. Wolverton, A. Zunger, First-principles theory of short-range order, electronic excitations, and spin polarization in Ni–V and Pd–V alloys, Phys. Rev. B 52 (12) (1995) 8813–8828, http://dx.doi.org/10.1103/PhysRevB.52.8813.

[69] M. Levy, J.P. Perdew, Density functionals for exchange and correlation energies: exact conditions and comparison of approximations, Int. J. Quant. Chem. 49 (4) (1994) 539–548, http://dx.doi.org/10.1002/qua.560490416.

[70] M. Born, K. Huang, Dynamical Theory of Crystal Lattices, Clarendon Press, 1954.

[71] A. van de Walle, Multicomponent multisublattice alloys, nonconfigurational entropy and other additions to the Alloy Theoretic Automated Toolkit, Calphad-Comput. Coupling Ph. Diagrams Thermochem. 33 (2) (2009) 266–278, http://dx.doi.org/10.1016/j.calphad.2008.12.005, arXiv:0906.1608.

[72] B.G. Lindsay, Mixture Models: Theory, Geometry and Applications, Institute of Mathematical Statistics, 1995.

[73] J. Quiñonero-Candela, C.E. Rasmussen, A unifying view of sparse approximate Gaussian process regression, J. Mach. Learn. Res. 6 (2005) 1935–1959.

[74] D.P. Landau, K. Binder, A Guide to Monte Carlo Simulations in Statistical Physics, 3rd edition, Cambridge University Press, 2009.

[75] T. Tsuji, Heat capacity of solids, in: S.L. Chaplot, R. Mittal, N. Choudhury (Eds.), Thermodynamic Properties of Solids: Experiments and Modeling, Wiley VCH, 2010, pp. 159–196, Ch. 5.

[76] I. Bilionis, P. Koutsourelakis, Free energy computations by minimization of Kullback–Leibler divergence: an efficient adaptive biasing potential method for sparse representations, J. Comput. Phys. 231 (9) (2012) 3849–3870, http://dx.doi.org/10.1016/j.jcp.2012.01.033.

[77] J. Kristensen, I. Bilionis, N. Zabaras, Adaptive simulation selection for the discovery of the ground state line of binary alloys with a limited computational budget, in: R. Melnik, R. Makarov, J. Belair (Eds.), Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science, Springer-Verlag, 2016.

[78] K.M. Abadir, J.R. Magnus, Matrix Algebra, Cambridge University Press, Cambridge, 2005.

[79] J. Ziman, Electrons and Phonons: The Theory of Transport Phenomena in Solids, reprint edition, Oxford Classic Texts in the Physical Sciences, Oxford University Press, 2001.