ICET – A Python library for constructing and sampling alloy cluster expansions

Mattias Ångqvist,[1] William A. Muñoz,[1] J. Magnus Rahm,[1] Erik Fransson,[1]
Céline Durniak,[2] Piotr Rozyczko,[2] Thomas Holm Rod,[2] and Paul Erhart[1, *]

[1]*Chalmers University of Technology, Department of Physics, Gothenburg, Sweden*
[2]*Data Management and Software Centre, European Spallation Source, Copenhagen, Denmark*

Alloy cluster expansions (CEs) provide an accurate and computationally efficient mapping of the potential energy surface of multi-component systems that enables comprehensive sampling of the many-dimensional configuration space. Here, we introduce ICET, a flexible, extensible, and computationally efficient software package for the construction and sampling of CEs. ICET is largely written in Python for easy integration in comprehensive workflows, including first-principles calculations for the generation of reference data and machine learning libraries for training and validation. The package enables training using a variety of linear regression algorithms with and without regularization, Bayesian regression, feature selection, and cross-validation. It also provides complementary functionality for structure enumeration and mapping as well as data management and analysis. Potential applications are illustrated by two examples, including the computation of the phase diagram of a prototypical metallic alloy and the analysis of chemical ordering in an inorganic semiconductor.

I. INTRODUCTION

Ordering phenomena are ubiquitous in materials science, physics, and chemistry. They are particularly relevant in multi-component systems, where they are associated for example with phase transitions, segregation as well as chemical order.[1] The underlying energetics can usually be assessed by first-principles calculations, based on e.g., density functional theory (DFT), with good accuracy. The computational cost of such calculations, however, precludes a statistically adequate sampling of the relevant configuration space. In this context, alloy cluster expansions (CEs) in combination with Monte Carlo (MC) simulations provide a powerful means to balance computational efficiency and accuracy.[2,3]

The CE approach has been widely and very successfully adopted to analyze for example phase diagrams in metallic[4,5] and semiconducting alloys,[6–8] including surfaces[9–15] as well as nanoparticles.[16–23] CEs have also been applied to study the temperature and composition dependence of ordering in various materials.[24–28] The CE approach is not limited to the mapping of total and mixing energies but can also be applied to model for example activation barriers,[29] vibrational properties,[30,31] chemical expansion,[26] or transport properties.[25]

Here, we introduce the integrated cluster expansion toolkit (ICET) to enable efficient construction and sampling of CEs. ICET is designed to be modular, extensible, and flexible, while maintaining high computational efficiency. This enables integration of ICET in extended workflows, reflecting the ongoing shift toward machine learning and large data initiatives in computational material science.[32,33] ICET is primarily developed in Python whereas computationally more demanding parts are written in C++, providing performance while maintaining portability and ease-of-use. This approach enables easy integration for example with countless first-principles codes and analysis tools accessible via the atomic simulation environment (ASE)[34] as well as state-of-the-art regression techniques via SCIKIT-LEARN.[35]

ICET provides a feature set that is comparable to or extends beyond earlier monolithic codes, such as the ATAT,[36–38] the UNCLE[39] or CASM codes.[40] Since ICET is written in Python, it is, however, straightforward to add new functionality. This enables one for example to implement advanced algorithms for finding ground states.[41,42] ICET is available under an open-source license and hosted on GITLAB to encourage community participation. Current functionality includes for example:

- support for multiple species and multiple coexisting sub-lattices, e.g., $Ba_{8-x}Sr_xGa_yGe_{46-y}$ or $Au_{1-x}Pd_x$:H_y

- advanced linear regression techniques with regularization (including compressive sensing[43]), cross-validation, and ensemble optimization via SCIKIT-LEARN[35]

- MC simulations in various ensembles using observers and multiple CEs in parallel via the MCHAMMER module

- supplemental functionality including e.g., structure enumeration,[44,45] structure mapping, convex hull extraction,[46] and ground state finding[42]

The remainder of this paper is organized as follows. The next section describes the methodologies implemented in ICET, including an overview of the CE formalism, algorithms available for CE construction, and the MCHAMMER module for sampling CEs via MC simulations. The components and workflow of ICET are summarized in Sect. V. Section VI demonstrates the potential of ICET via examples. The first example addresses the construction and sampling of CEs for the Ag–Pd system as well as the subsequent generation of a phase diagram from these data. The second example summarizes the application of ICET for the simulation of chemical ordering in a semiconducting system, specifically an inorganic clathrate.

II. CLUSTER EXPANSION FORMALISM

A. Clusters and orbits

The objective of a CE is to describe the variation of a property of interest, most commonly the energy, with the chemical configuration, i.e. the distribution of different species over a lattice. To this end, the structure is decomposed into a set of
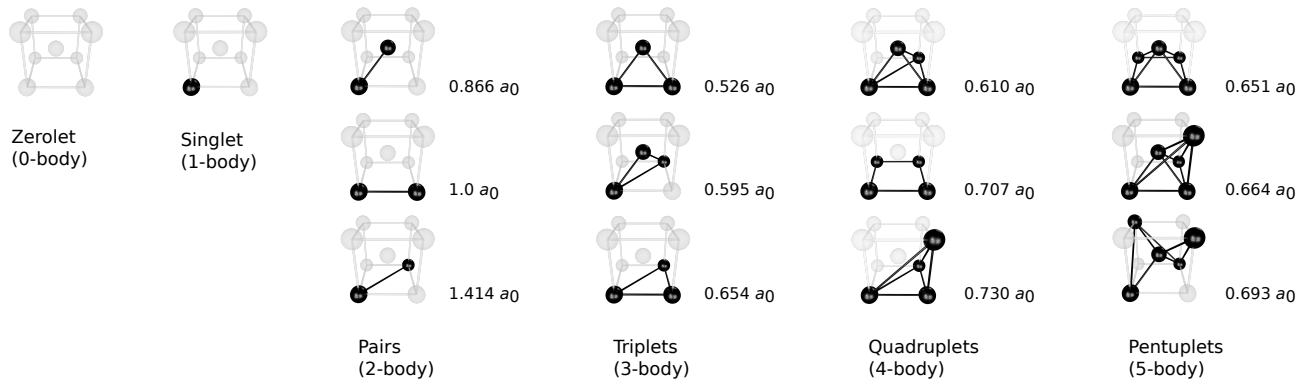
FIG. 1. Illustration of clusters on a BCC lattice. The clusters are ordered by radius from top to bottom, with the size given in units of the lattice constant a_0. The cluster radius is defined as the average distance of the sites to the center of the cluster.

clusters, where a cluster of order k is defined as a list of k sites. Clusters are commonly referred to by order as singlets, pairs, triplets and so on.

The clusters are subject to the symmetry of the underlying lattice, as described by the associated space group (for periodic systems) or point group (for non-extended systems). Clusters that can be mapped onto each other as a result of the application of an intrinsic symmetry operation of the lattice, are said to belong to the same orbit (Fig. 2). Each orbit in turn is represented by a symmetry inequivalent cluster, from which every other cluster of the orbit can be generated by application of the symmetry operations and a permutation of the sites.

While the number of clusters is in principle infinite, the short-ranged nature of physical interactions implies that clusters with notable contributions to the CE are commonly shorter-ranged and few-bodied. It is therefore customary to only include clusters up a certain size and order, as shown for the case of a BCC lattice in Fig. 1.

B. Point functions

It can be formally shown that a CE is able to represent any function of the configuration $Q(\boldsymbol{\sigma})$ if one can construct a complete orthogonal basis.[2] Here, $\boldsymbol{\sigma}$ denotes the occupation vec-

tor, the N elements of which indicate the species that resides on the corresponding site. To obtain a practical procedure, for each lattice point p one can define M orthogonal point functions $\Theta_n(\sigma_p)$

$$\Theta_n(\sigma_p) = \begin{cases} 1 & \text{if } n = 0 \\ -\cos\left(\pi(n+1)\sigma_p/M\right) & \text{if } n \text{ is odd} \\ -\sin\left(\pi n \sigma_p/M\right) & \text{if } n \text{ is even,} \end{cases}$$

where M is the allowed number of species and n goes from 0 to $M-1$. These point functions form an orthogonal set over all possible occupation numbers,[37]

$$\langle \Theta_n, \Theta_{n'} \rangle = \sum_{\sigma_p=0}^{M-1} \Theta_n(\sigma_p)\Theta_{n'}(\sigma_p) = \begin{cases} 0 & \text{if } n \neq n' \\ \neq 0 & \text{if } n = n'. \end{cases}$$

In the case of multiple sublattices M can assume different values for different sublattices. For example, in the case of the zincblende alloy $Al_{1-x-y}Ga_xIn_yAs_{1-z}Sb_z$, $M = 3$ for the cation and $M = 2$ for the anion lattice.

C. CE expression

A set of functions $\Pi_\alpha(\boldsymbol{\sigma})$ in the M^N-dimensional configuration space can now be constructed as products of point functions,

$$\Pi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) = \Theta_{n_1}(\sigma_1)\Theta_{n_2}(\sigma_2)\ldots\Theta_{n_l}(\sigma_l).$$

Here, $\boldsymbol{\alpha} = [n_1, n_2, \ldots n_l]$, where n_i are point function indices and l is the number of sites in the structure. Each $\boldsymbol{\alpha}$ corresponds to a cluster in the sense that $n_i = 0$ if site i is not part of the cluster (such that the corresponding point function is 1) and nonzero otherwise. In the binary case, there is one $\boldsymbol{\alpha}$ for each cluster, each having elements that are either 0 or 1. In systems with more than two components, there are multiple $\boldsymbol{\alpha}$ corresponding to the same cluster, due to the the possible combinations of nonzero point function indices n_i.

It can be shown that the functions $\Pi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma})$ form a complete orthogonal set, so that any function of the configuration can
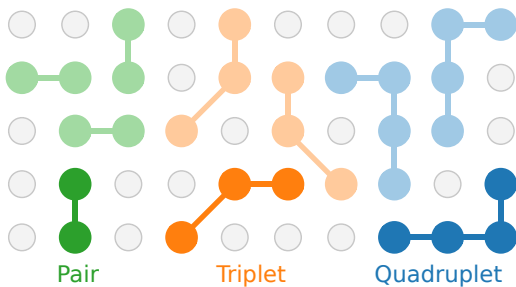

FIG. 2. Examples for pair (green), triplet (orange) and quadruplet (blue) clusters on a square lattice. The representative (symmetry inequivalent) clusters are shown in dark colors, whereas examples for other clusters in the orbits are shaded.

be expressed as

$$Q(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\alpha}} J_{\boldsymbol{\alpha}} \Pi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}).$$

All basis functions $\Pi_{\boldsymbol{\alpha}}$ have one configuration invariant component that is equal to 1 when $\boldsymbol{\alpha} = \mathbf{0}$. One can therefore omit the latter from the basis functions and instead use one configuration invariant term Q_0. Furthermore, taking into account the symmetry of the clusters the summation can be carried out over orbits (or representative clusters) rather than all clusters, which yields the full CE expression

$$Q(\boldsymbol{\sigma}) = Q_0 + \sum_{\boldsymbol{\alpha}} \langle \Pi_{\boldsymbol{\alpha}'}(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\alpha}} m_{\boldsymbol{\alpha}} J_{\boldsymbol{\alpha}}. \tag{1}$$

The $\langle \ldots \rangle_{\boldsymbol{\alpha}}$ bracket indicates the average over the basis functions for all $\boldsymbol{\alpha}'$ that are symmetry equivalent to $\boldsymbol{\alpha}$. The effective cluster interactions (ECIs) $J_{\boldsymbol{\alpha}}$ are the free parameters of the CE and the target of the training procedure described in the next section. Finally, $m_{\boldsymbol{\alpha}}$ denotes the multiplicity of the representative cluster $\boldsymbol{\alpha}$.

III. CLUSTER EXPANSION CONSTRUCTION

A. Matrix form

To obtain a CE for a specific material one must determine the ECIs. To this end, one requires reference data in the form of a set of configurations $\{\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, \ldots, \boldsymbol{\sigma}_n\}$ as well as an associated vector of target data $\boldsymbol{Q}^T = [Q_1, Q_2, \ldots Q_n]$, which is usually obtained from first-principles calculations. Equation (1) can be cast in the form

$$\boldsymbol{Q} = \boldsymbol{\Pi} \boldsymbol{J}, \tag{2}$$

where the rows of $\boldsymbol{\Pi}$ are given by

$$\boldsymbol{\Pi}_i = \left[1, \langle \Pi_{\boldsymbol{\alpha}_1'}(\boldsymbol{\sigma}_i) \rangle_{\boldsymbol{\alpha}_1} m_{\boldsymbol{\alpha}_1}, \ldots, \langle \Pi_{\boldsymbol{\alpha}_n'}(\boldsymbol{\sigma}_i) \rangle_{\boldsymbol{\alpha}_n} m_{\boldsymbol{\alpha}_n} \right], \tag{3}$$

and $\boldsymbol{J}$ denotes the ECIs with $J_0 = Q_0$. Note that it can sometimes be useful to exclude $m_{\boldsymbol{\alpha}}$ from $\boldsymbol{\Pi}$ and let the target values $\boldsymbol{Q}$ refer to the primitive unit cell. This will ensure all elements in $\boldsymbol{\Pi}$ are in the interval $[-1, 1]$ and avoids a bias due to the number of elements in $\boldsymbol{\sigma}$.

B. Linear regression techniques

Following the decomposition of a lattice into clusters and the construction of the sensing matrix, it remains to determine the ECIs by solving the linear system given by Eq. (2). This is equivalent to finding the parameter vector $\boldsymbol{J}$ that minimizes $\|\boldsymbol{\Pi} \boldsymbol{J} - \boldsymbol{Q}\|_2$. In principle a solution can be determined by conventional least-squares, which works well in the overdetermined limit. Due to the computational cost associated with

DFT calculations the linear system is, however, often underdetermined and/or rows of the sensing matrix $\boldsymbol{\Pi}$ are correlated. At the same time, the nearsightedness of physical interactions suggests that the solution vector $\boldsymbol{J}$ be sparse.[43] We are thus faced with the task of feature selection, a common process in machine learning, which yields models that are less prone to overfitting and more transferable. It can also reduce the computational expense during sampling (see Sect. IV).

To achieve sparse solutions and combat overfitting one can include regularization terms in the objective function in the form of the ℓ_1 or ℓ_2-norm of the solution vector. Consider, for example, elastic net regularization, for which

$$\boldsymbol{J}_{\text{opt}} = \underset{\boldsymbol{J}}{\text{argmin}} \left\{ \|\boldsymbol{\Pi} \boldsymbol{J} - \boldsymbol{Q}\|_2^2 + \alpha \|\boldsymbol{J}\|_1 + \beta \|\boldsymbol{J}\|_2^2 \right\}.$$

For $\alpha = 0$ this expression reduces to Ridge regression while for $\beta = 0$ one obtains the expression commonly used for solving the least absolute shrinkage and selection operator (LASSO) problem. Other techniques include the split-Bregman algorithm, which has been previously used for constructing CEs,[43] as well as recursive feature elimination (RFE), which iteratively removes the weakest parameters from a model and can be applied to different objective functions. Furthermore, there are Bayesian linear regression techniques such as automatic relevance detection regression (ARDR), which provide probabilistic models of the regression problem at hand, and Bayesian compressive sampling.[47] The parameters that are intrinsic to a regression algorithm, such as α in the case of LASSO or the allowed number of features in the case of RFE, are known as hyper-parameters.

The performance of different models can be assessed by cross-validation (CV). To this end, the available reference data is split into training and validation sets. The former is used to fit a model, whereas the latter is used to measure the predictive power of the model, usually via the root-mean-square error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{N_s} \sum_i \left(Q_i^{\text{model}} - Q_i^{\text{target}} \right)^2},$$

where the summation extends over the N_s structures comprising the validation set. To reduce the statistical error, the RMSE is furthermore averaged over several different splits of the reference data.

ICET supports various regression techniques, including the ones named above, via the SCIKIT-LEARN machine learning library[35] and allows one to compute CV scores in a number of different ways. Since sensing matrix and target data are readily available via the Python interface, one can also interface directly with SCIKIT-LEARN or other machine learning libraries. ICET also provides functionality for generating ensembles of models from a single sensing matrix. Thereby it is possible to check the sensitivity and stability of more advanced prediction, as illustrated in the examples below.

IV. CLUSTER EXPANSION SAMPLING

A CE can be employed in a number of different ways, including finding ground states,[42] but it is most commonly sampled using MC simulations. For this purpose, ICET includes the MCHAMMER module, which supports various thermodynamic ensembles and provides supplemental functionality for data management and analysis.

MC sampling is carried out using the Metropolis algorithm, in which a trial is accepted with probability

$$\mathcal{P} = \min \left\{ 1, \exp\left(-\beta \Delta \psi\right) \right\}.$$

Here, $\beta = 1/k_B T$ and $\Delta \psi$ is the change in the thermodynamic potential associated with the ensemble being sampled (excluding the entropy term).

In the case of the canonical ($\{N_i\}VT$) ensemble the thermodynamic potential equals the internal energy, $\psi = E$, and a trial step involves swapping the species of two sites (conserving the concentrations).

Especially when exploring phase diagrams it is often useful to sample along the concentration axes. This can be achieved for example by using the semi-grand canonical (SGC) ($N\{\Delta\mu_i\}VT$) ensemble. In this case, the trial move involves only one site, for which the occupation is swapped and the underlying thermodynamic potential is $\psi = E - N \sum_{i=2} c_i \Delta\mu_i$, where N is the total number of sites, c_i is the concentration of species i, and $\Delta\mu_i = \mu_i - \mu_1$ the chemical potential difference of species i relative to the first species.

In the SGC ensemble the mapping from chemical potential difference to concentration is multi-valued in two-phase regions. It therefore cannot be used to sample across miscibility gaps. This shortcoming can be overcome by employing the variance-constrained semi-grand canonical (VCSGC) ensemble,[48] which uses the same trial move as the SGC ensemble but for which $\psi = E + N k_B T \bar{\kappa}(c + \bar{\phi}/2)^2$. The intensive parameters $\bar{\phi}$ and $\bar{\kappa}$ constrain respectively the average and the fluctuation of the concentration.

The choice of ensemble is motivated by the characteristics of the system and the objective of the simulations. The canonical ensemble conserves concentrations, which makes it ideal for studying systems at specific compositions, extracting structural order parameters or obtaining ground states by simulated annealing. The SGC and VCSGC ensembles, on the other hand, allow the composition to be continuously varied and by extension the integration of the free energy. In the SGC ensemble, the concentration derivative of the canonical free energy is given (for simplicity for a binary system) by

$$\partial \Delta F / \partial c = -N \Delta\mu. \qquad (4)$$

This relation is, however, useful only in single-phase regions of the phase diagram, where $\Delta\mu$ maps to one and only one concentration. In SGC simulations, multi-phase regions manifest themselves by discontinuous jumps between phase boundaries. Such discontinuities, if carefully studied, can be exploited for tracking phase boundaries but hinder integration of the free energy.[38]

In the VCSGC ensemble, the canonical free energy derivative is (for a sufficiently large system) given by[48]

$$\partial \Delta F / \partial c = -2 N k_B T \bar{\kappa} \left(\langle c \rangle + \bar{\phi}/2\right), \qquad (5)$$

where $\langle c \rangle$ is the average observed concentration. Unlike the SGC ensemble, the mapping between $\bar{\phi}$ and concentration is always one-to-one for a sufficiently large value of $\bar{\kappa}$, whence the free energy can be recovered by integration across multi-phase regions. The phase diagram can then be constructed with standard techniques for free energy minimization.

MC simulations in both the SGC and VCSGC ensembles are illustrated in Sect. VI A whereas the canonical ensemble is employed in Sect. VI B.

V. WORKFLOW

ICET integrates the CE formalism (Sect. II) with linear regression (Sect. III) and sampling techniques (Sect. IV) into one workflow (Fig. 3). Assuming reference data, say from DFT calculations, is available for a set of structures that have been generated for example by enumeration[44] the construction and sampling of a CE proceeds as follows.
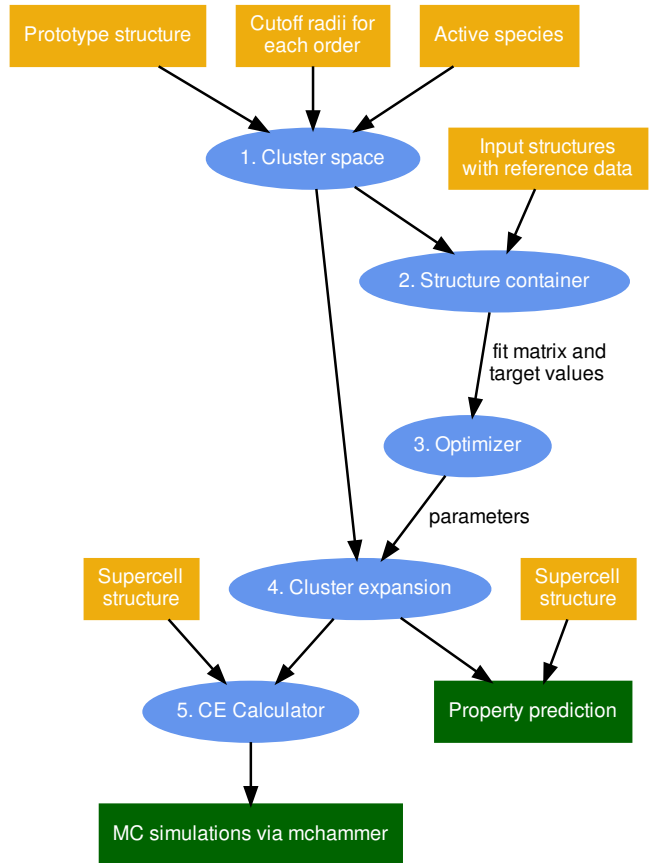

FIG. 3. Illustration of the ICET workflow. Entities represented by Python objects are shown in blue, input parameters and data in orange, and additional functionalities in green.
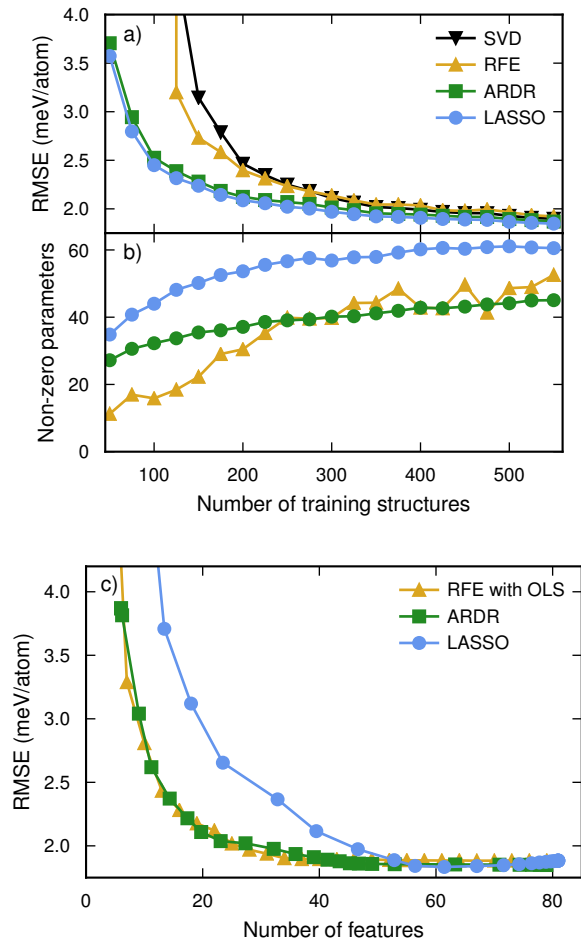
FIG. 4. Construction of CEs for Ag–Pd. (a) RMSE obtained by CV and (b) number of features (non-zero parameters) as a function of the number of training structures. (c) CV-RMSE as a function of the number of features using a training set size of 563 structures.

1. The first step involves constructing a cluster space, which requires a prototype structure, a set of cutoff radii that define which clusters to include in the expansion, and a specification as to which species are allowed on each site. Internally, the SPGLIB library[49] is employed to find the symmetries of the underlying lattice.

2. Next, cluster vectors are computed according to Eq. (3) for all structures in the reference set and compiled into a structure container, which holds the sensing matrix $\mathbf{\Pi}$ as well as the vector of target values $\boldsymbol{Q}$.

3. Then linear regression techniques in combination with CV are employed to solve Eq. (2) and find an optimal parameter vector $\boldsymbol{J}$. Here, externally provided optimization algorithms from, e.g., SCIKIT-LEARN and SCIPY can be used.

4. Parameters and cluster space are subsequently combined to obtain the actual CE, which can be used to predict the property in question for arbitrary supercells of the prototype structure.

5. For efficient sampling one can also set up a CE calculator for a specific supercell to be used in MC simulations via the MCHAMMER module.

This workflow is supplemented by a number of tools, which enable one for example to enumerate structures,[44] extract the convex hull, map relaxed structures onto ideal lattices, or find ground states.[42]

A full description of the different entities involved in this process and the Python objects that describe them can be found in the ICET user guide.[50]

VI. APPLICATIONS

A. Phase diagram of the Ag–Pd system

1. Reference calculations

As a first example for the application of ICET, we describe the construction of a CE for the face-centered cubic (FCC) Ag–Pd alloy. To this end, a database of 631 reference structures corresponding to all distinct supercells with up to 8 atoms was set up using the structure enumeration feature of ICET. More refined structure selection approaches are available[43] but are not considered in this example.

Reference calculations were then carried out using DFT calculations using the projector augmented method[51,52] as implemented in VASP.[53,54] The van-der-Waals density functional method[55,56] with consistent exchange,[57] which has been shown to be very well suited for transition metals,[58] was employed to represent the exchange-correlation potential. The Brillouin zone was sampled using $\boldsymbol{k}$-point grids equivalent to a $18 \times 18 \times 18$-mesh for the primitive FCC unit cell and the plane wave cutoff energy was set to $384\,\mathrm{eV}$. Both the atomic positions and the cell metric were relaxed until residual forces and stress were less than $10\,\mathrm{meV/\mathring{A}}$ and $0.8\,\mathrm{GPa}$, respectively. Relaxations were carried out using first-order Methfessel-Paxton smearing with a width of $0.1\,\mathrm{eV}$. Final energy calculations were carried out using for the relaxed structures the tetrahedron method with Blöchl corrections using a smearing width of $0.05\,\mathrm{eV}$.

2. Construction of CE models

Next a cluster space was set up, including clusters up to fourth order with cutoffs of $3.3\,a_0$, $1.6\,a_0$, and $1.5\,a_0$ in units of the lattice parameter a_0 for pairs, triplets, and quadruplets, respectively. The resulting cluster space contained 81 parameters, including 1 zerolet, 1 singlet, 24 pairs, 20 triplets, and 35 quadruplets. In principle, ICET does not impose limits on cluster order or size.

In order to study the convergence with respect to the number of training structures we computed the learning curve using the RMSE score averaged over the validation set using the CV estimator functionality in ICET. The latter generates an estimate for the CV-RMSE score by using the shuffle-split
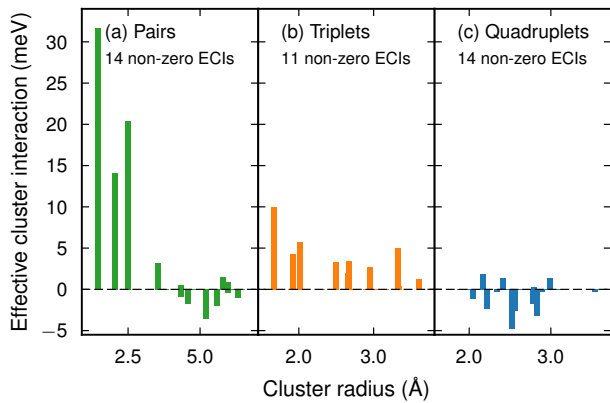
FIG. 5. ECIs for the Ag–Pd system obtained using the ARDR method with all 631 structures in the training set. The ECIs illustrate the sparseness of the solution and reflect the short-ranged nature of the interactions. The cluster radius is defined as the mean distance from the center of the cluster to any of its points.

method with 50 splits, while the final CE is obtained by training against the complete data set. Four different optimization methods were compared, including singular value decomposition, LASSO, RFE based on ordinary least-squares (OLS), and ARDR (Fig. 4a).

In the overdetermined region all methods yield similar CV-RMSE scores; in the underdetermined region, though, LASSO and ARDR outperform the other two methods. While LASSO and ARDR have almost identical CV scores there is, however, a significant difference in the number of features, i.e. non-zero parameters (Fig. 4b). To study this behavior further we analyzed the CV-RMSE score as a function of the number of features in the solution (Fig. 4c). To this end, we used a training set of 563 structures and scanned the hyperparameters of the regression algorithms to control the sparsity of the solution.

This analysis reveals that ARDR and RFE with OLS yield CEs with a very low CV-RMSE of 2 meV/atom using only 30 features. LASSO reaches the same level but yields about 50 features. In this case, ARDR is thus the method of choice since it converges as quickly with the number of training structures as LASSO but yields a smaller number of orbits with non-zero ECIs (Fig. 5), which leads to a more transferable model and reduces the computational cost for sampling.

3. Phase diagram from MC simulations

To construct a phase diagram for the Ag–Pd system, we employed the MCHAMMER module for sampling a $5 \times 5 \times 5$ conventional FCC supercell (500 sites) in both the SGC and VCSGC ensembles. We carried out 200,000 MC trial steps with the SGC ensemble at 105 values of $\Delta\mu$ in the range $[-1.04, 1.04]$ and the same number of trial steps with the VCSGC ensemble at $\bar{\kappa} = 200$ and 105 values of $\bar{\phi}$ in the range $[-2.3, 0.3]$. Simulations were run at 25 K intervals between 100 and 900 K, corresponding in total to approximately
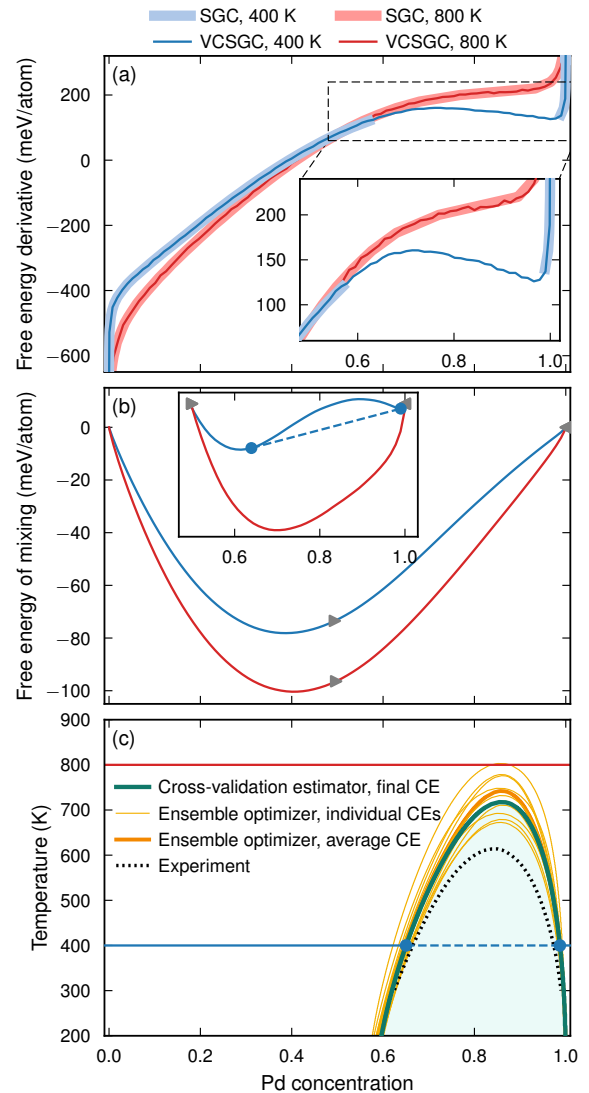

FIG. 6. Ag–Pd system. (a) Free energy derivative $\partial F/\partial c$ obtained by MC sampling in the SGC and VCSGC ensembles. The SGC ensemble cannot yield stable solutions in the two-phase region (inset). (b) Free energy of mixing obtained by numerical integration in the VCSGC ensemble. The inset indicates the behavior between $c = 0.5$ and 1.0 where the free energy is concave at low temperature. (c) Phase diagram constructed from the free energy landscape using CEs obtained using the CV estimator (green line) as well as the ensemble optimizer (orange lines). Experimental data (dotted black line) from Ref. 59 obtained via the CALPHAD method.

6.9×10^8 trial steps or 1.4×10^6 MC cycles.

Simulations were first carried out for the CE described above that was constructed using the CV estimator with ARDR and the full dataset of 631 structures (CV-RMSE 2 meV/atom).

The free energy derivatives extracted from the SGC and VCSGC simulations (using Eqs. (4) and (5)) coincide everywhere except for a region on the Pd-rich side at lower temperatures where the SGC simulations exhibit a discontinuity, which is the hallmark of a two-phase region (Fig. 6a). In this

situation, the full free energy can thus only be recovered from the VCSGC simulations (Fig. 6b). The two-phase region is manifested by a concave region in the free energy of mixing (inset in Fig. 6b).

To construct the phase diagram, we fitted the free energy of mixing Fig. 6b) to third-order Redlich–Kister polynomials at each temperature separately and then fitted the temperature dependence of the polynomial expansion coefficients to conventional third-order polynomials. This representation provides a continuous and smooth map of the free energy in both temperature and composition, akin to the CALPHAD approach.[59] The predicted phase diagram exhibits a pronounced miscibility gap on the Pd-rich side with a critical temperature T_c of 718 K. This is overall in good agreement with the experimental result,[59] except for an overestimation of T_c, which has been experimentally determined as approximately 610 K.

To illustrate the sensitivity of the phase diagram to variations in the training set, we also considered a set of ten CEs that were obtained using the ARDR method via the ensemble optimizer functionality of ICET. The latter approach enables one to generate a series of CEs that are trained in identical fashion but are based on different training sets. The latter are generated by selection with replacement (bagging) from the reference data set such that the number of structures in the training sets equals the number of reference structures. The thus obtained CEs are numerically very similar to the CE obtained using the CV estimator approach before (Fig. 4). Using this set of CEs enabled us to estimate not only the error in the mixing energies but its impact on the final phase diagram.

The CE obtained by averaging the ECIs over all CEs in the ensemble yields a T_c of 742 K that is only slightly higher than the CE generated using the CV estimator. The individual CEs yield, however, a larger variation, spanning the range from 673 to 803 K. This demonstrates that CV-RMSE alone is an insufficient measure of the quality of a CE. Rather a more careful assessment of the quantity of interest must be carried out if an accurate estimate is required.

B. Chemical ordering in an inorganic clathrate

1. Background and reference structures

The CE approach is not limited to metallic system and the prediction of phase diagrams. It is also very useful to model for example the degree of ordering as a function of temperature and composition. The latter can in turn can be experimentally assessed using diffraction techniques based on X-ray or neutron scattering.[60] Here, this possibility is illustrated by using ICET to predict the site occupancy factors (SOFs) in the inorganic clathrate $Ba_8Al_xSi_{46-x}$.

Clathrates are inclusion compounds with a lattice structure that can trap atomic or small molecular species. $Ba_8Al_xSi_{46-x}$ falls into the class of type-I clathrates, which belong to spacegroup $Pm\bar{3}n$.[61] In this case, the host lattice is made up of Al and Si atoms, which occupy Wyckoff sites $6c$, $16i$, and $24k$, whereas Ba atoms reside inside the cages for

charge balance. Al and Si do not occupy the framework sites randomly but exhibit some degree of ordering, which results from a delicate balance between energy and entropy and can be experimentally accessed via the SOFs. While for a completely random distribution at, e.g., $x = 12$ one would expect SOFs of 12/46=26% for all sites, in $Ba_8Al_xSi_{46-x}$ one observes values in the range from close to zero to 80%.[62] Detailed studies of this behavior, including careful comparison with experiment, have been published elsewhere.[25–27] Here, we focus on the construction and sampling of CEs for this system, in particular highlighting the analysis capabilities provided by ICET.

2. CE construction

The unit cell contains 46 framework sites, which prevents an enumeration approach for structure generation. Instead, 240 occupations of the primitive unit cell for $x = (13 \ldots 16)$ were produced by randomly distributing Al and Si atoms over the host lattice. The structures were relaxed using DFT calculations using a similar procedure as for the Ag–Pd structures described above. The PBE exchange-correlation functional was used[63] with a plane wave energy cutoff of 319 eV and a Γ-centered $3 \times 3 \times 3$ k-point mesh. The other parameters were identical to those given in Sect. VI A.

A cluster basis was constructed using a cutoff of $0.49\,a_0$ for both pairs and triplets, resulting in 13 and 23 symmetry inequivalent clusters, respectively. CEs were generated using ARDR, LASSO, and RFE with OLS in conjunction with the CV estimator functionality based on the shuffle-split method with 50 splits.

RFE with OLS yields both fast convergence with training set size and sparse solutions (Fig. 7a,b). LASSO and ARDR require almost twice as many training structures to achieve similar CV-RMSE values. ARDR provides sparse solutions that are similar to those obtained by RFE with OLS. The CE models obtained by LASSO, however, have a much larger number of features (Fig. 7c). These trends are similar to the case of Ag–Pd (Sect. VI A) except for the roles of ARDR and RFE with OLS being reversed.

3. CE sampling

To predict the SOFs as a function of temperature, we employed the MCHAMMER module for sampling a $2 \times 2 \times 2$ supercell (268 sites) in the canonical ensemble at $x = 12$, for which experimental data is available,[62] using a simulated annealing approach. The temperature was decreased from 1200 to 0 K at a rate of 100 K/10,000 MC cycles, corresponding to a total of almost 48 million trial steps. The SOFs were monitored during the simulation using the observer functionality of ICET, which enables one to compute various quantities of interest at specified intervals along the trajectory.

The simulations were first carried out using the CE obtained using RFE with OLS and a set of 240 structures. To estimate
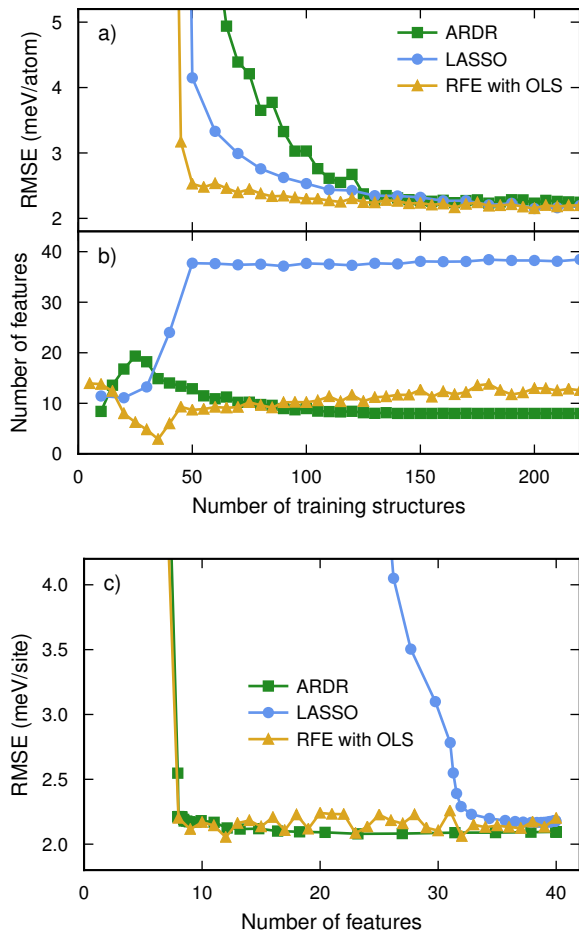
FIG. 7. Construction of CEs for $Ba_8Al_xSi_{46-x}$. (a) CV-RMSE and (b) number of features as a function of the number of training structures. (c) CV-RMSE as a function of the number of features obtained by varying the hyper-parameters of the respective methods using a fixed training set comprising 240 structures.

FIG. 8. Al SOFs as a function of temperature in $Ba_8Al_xSi_{46-x}$ with $x = 12$ from simulated annealing. Simulations were carried out for an ensemble of CE models obtained by different 90-10 splits of the reference data set (thin shaded lines) as well as the CE obtained by averaging over the ensemble (bold line). Experimental data from Ref. 62.

tively, in agreement with earlier simulations.[26,27]

The calculated SOFs compare very well with the available experimental data in the temperature interval between 800 and 900 K.[62] This is very reasonable since it is likely that kinetic factors prevent full ordering in the experiments. A more extensive analysis of the SOFs as both a function of temperature and composition[26] shows very good comparison with experiment over the entire composition range,[60,62] which further validates the present approach. The ground state configurations obtained in the zero-temperature limit are valuable in themselves as they provide ordered prototypes, which can be used for further analyses of e.g., electrical[25] and thermal transport properties.[64]

VII. CONCLUSIONS

In the present paper, we have introduced the ICET Python package for the construction and sampling of alloy CEs. Thanks to its modular design it can be readily extended and combined with other Python packages to achieve complex functionalities. It thereby also provides an excellent environment for method development, including the adaptation of further machine learning techniques (e.g., genetic algorithms[65]) and the integration in extended workflows, e.g., in the context of high-throughput computing.[32,33] ICET also provides a number of supplementary features pertaining, e.g., to structure enumeration and mapping as well as data analysis.

ICET readily supports a variety of different methods for constructing CEs, as demonstrated specifically for the metallic alloy Ag–Pd and the inorganic clathrate $Ba_8Al_xSi_{46-x}$. Several different regression methods were compared, illustrating a balance between the sparsity of the solution and the accuracy of the final CE. While for Ag–Pd the ARDR method provided the best performance, yielding both low CV scores and a sparse solution, RFE based on OLS achieved the best results

the statistical reliability of the thus predicted SOFs, we furthermore ran the simulations for an ensemble of models that were generated from the available data (bagging), in almost identical fashion as for the Ag–Pd models described above.

At the composition of $x = 12$ the number of nearest neighbor Al–Al pairs is zero over the entire temperature range. This behavior is due to Al–Al bonds being energetically unfavorable, which is familiar from the Loewenstein rule for zeolites. While the Al–Al pair distribution is experimentally practically impossible to access, diffraction experiments can provide information about the SOFs, which are ultimately the result of the interplay of energy and entropy.[60] The SOFs obtained from MC simulations exhibit a systematic variation with temperature and strongly deviate from the random limit, which would correspond to approximately 35% (Fig. 8). At temperatures below approximately 600 K the SOFs converge and the structure adopts a well ordered configuration. The ground state at this stoichiometry corresponds to Al SOFs of 100%, 37.5%, and 0% for Wyckoff sites 6c, 16i, and 24k, respec-
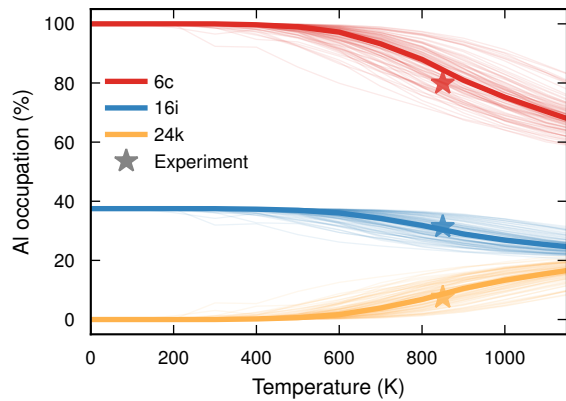
in the case of $Ba_8Al_xSi_{46-x}$. Regression using LASSO led to less optimal solutions in both situations, an observation that we have also made in the case of force constant models.[66,67]

ICET also allows construction of ensembles of CE models, which provides a powerful means to investigate the sensitivity of a prediction to variations in the underlying model. Specifically, this enables one to extrapolate the impact of the statistical uncertainty in CE models to complex observables such as a phase diagram or SOFs. In the case of the Ag–Pd system, this approach was for example adopted to demonstrate that a set of models with numerically similar CV scores can yield variations in the critical temperature on the order of $100\,K$, providing an estimate of average and standard deviation. The same approach was employed for the $Ba_8Al_xSi_{46-x}$ system to determine the statistical uncertainty of the predicted temperature dependence of the SOFs.

The MCHAMMER module of ICET includes functionality for extracting additional information from MC trajectories, as illustrated by the application to the inorganic clathrate $Ba_8Al_xSi_{46-x}$. This allows one to observe for example structural order parameters throughout a simulation, including SOFs, neighbor counts, or short-range order parameters.[68]

Overall ICET package is thus well suited for efficient construction and sampling of CEs, for example in high-throughput schemes. For such applications one must, however, not only consider the computational effort but also the amount of human intervention required. In the future, it is therefore desirable to set up protocols that further automatize the selection of e.g., regression method, hyper-parameters, and training set size.

ICET is provided under an open-source license. Its development is hosted on GITLAB to encourage community participation and the most recent released version can be conveniently installed via PYPI. A comprehensive user guide with an extensive tutorial section is available online.[50]

* erhart@chalmers.se

[1] A. Van der Ven, J. Thomas, B. Puchala, and A. Natarajan, Annual Reviews of Materials Research **48**, 27 (2018).

[2] J. M. Sanchez, F. Ducastelle, and D. Gratias, Physica **128**, 334 (1984).

[3] D. de Fontaine, in *Solid State Physics*, Vol. 47, edited by H. Ehrenreich and D. Turnbull (Academic Press, 1994) p. 33.

[4] M. Asta, R. McCormack, and D. de Fontaine, Phys. Rev. B **48**, 748 (1993).

[5] V. Ozoliņš, C. Wolverton, and A. Zunger, Physical Review B **57**, 6427 (1998).

[6] A. Van der Ven, M. K. Aydinol, G. Ceder, G. Kresse, and J. Hafner, Phys. Rev. B **58**, 2975 (1998).

[7] G. Ceder, A. Van der Ven, C. Marianetti, and D. Morgan, Modelling and Simulation in Materials Science and Engineering **8**, 311 (2000).

[8] F. Zhou, T. Maxisch, and G. Ceder, Physical Review Letters **97**, 155704 (2006).

[9] R. Drautz, H. Reichert, M. Fähnle, H. Dosch, and J. M. Sanchez, Physical Review Letters **87**, 236102 (2001).

[10] M. H. F. Sluiter and Y. Kawazoe, Phys. Rev. B **68**, 085410 (2003).

[11] P. Welker, O. Wieckhorst, T. C. Kerscher, and S. Müller, Journal of Physics: Condensed Matter **22**, 384203 (2010).

[12] J. A. Stephens, H. C. Ham, and G. S. Hwang, Journal of Physical Chemistry C **114**, 21516 (2010).

[13] W. Chen, D. Schmidt, W. F. Schneider, and C. Wolverton, The Journal of Physical Chemistry C **115**, 17915 (2011).

[14] L. Cao and T. Mueller, Journal of Physical Chemistry C **119**, 17735 (2015).

[15] L. M. Herder, J. M. Bray, and W. F. Schneider, Surface Science **640**, 104 (2015).

[16] T. Mueller and G. Ceder, ACS Nano **4**, 5647 (2010).

[17] R. V. Chepulskii, W. H. Butler, A. van de Walle, and S. Curtarolo, Scripta Materialia **62**, 179 (2010).

[18] K. Yuge, Physical Review B **84**, 1 (2011).

[19] T. Mueller, Physical Review B **86**, 144201 (2012).

[20] T. L. Tan, L.-L. Wang, D. D. Johnson, and K. Bai, Nano Letters **12**, 4875 (2012).

[21] L.-L. Wang, T. L. Tan, and D. D. Johnson, Nano Letters **14**, 7077 (2014).

[22] C. Li, D. Raciti, T. Pu, L. Cao, C. He, C. Wang, and T. Mueller, The Journal of Physical Chemistry C **122**, 18040 (2018).

[23] L. Cao, C. Li, and T. Mueller, Journal of Chemical Information and Modeling (2018), 10.1021/acs.jcim.8b00413.

[24] H. Kim, M. Kaviany, J. C. Thomas, A. Van der Ven, C. Uher, and B. Huang, Physical Review Letters **105**, 265901 (2010).

[25] M. Ångqvist, D. O. Lindroth, and P. Erhart, Chemistry of Materials **28**, 6877 (2016).

[26] M. Ångqvist and P. Erhart, Chemistry of Materials **29**, 7554 (2017).

[27] M. Troppenz, S. Rigamonti, and C. Draxl, Chemistry of Materials **29**, 2414 (2017).

[28] N. S. H. Gunda, B. Puchala, and A. Van der Ven, Physical Review Materials **2**, 033604 (2018).

[29] A. Van der Ven, G. Ceder, M. Asta, and P. D. Tepesch, Physical Review B **64**, 184307 (2001).

[30] D. Morgan, A. van de Walle, G. Ceder, J. D. Althoff, and D. de Fontaine, Modelling and Simulation in Materials Science and Engineering **8**, 295 (2000).

[31] A. van de Walle and G. Ceder, Review of Modern Physics **74**, 11 (2002).

[32] R. H. Taylor, F. Rose, C. Toher, O. Levy, K. Yang, M. B. Nardelli, and S. Curtarolo, Computational Materials Science **93**, 178 (2014).

[33] G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, and B. Kozinsky, Computational Materials Science **111**, 218 (2016).

[34] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, Marcin Dułak, J. Friis, M. N. Groves, B. Ham-

mer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, Kristen Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, Andrew Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, Tejs Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, Journal of Physics: Condensed Matter **29**, 273002 (2017).

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Journal of Machine Learning Research **12**, 2825 (2011).

[36] A. van de Walle and G. Ceder, Journal of Phase Equilibria **23**, 348 (2002).

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

[38] A. van de Walle and M. Asta, Modelling and Simulation in Materials Science and Engineering **10**, 521 (2002).

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

[40] "CASM code," https://github.com/prisms-center/CASMcode (Accessed: 2019-01-10).

[41] W. Huang, D. A. Kitchaev, S. T. Dacek, Z. Rong, A. Urban, S. Cao, C. Luo, and G. Ceder, Physical Review B **94**, 134424 (2016).

[42] P. M. Larsen, K. W. Jacobsen, and J. Schiøtz, Physical Review Letters **120**, 256101 (2018).

[43] L. J. Nelson, G. L. W. Hart, F. Zhou, and V. Ozoliņš, Physical Review B **87**, 035125 (2013).

[44] G. L. W. Hart and R. W. Forcade, Physical Review B **77**, 224115 (2008).

[45] G. L. W. Hart and R. W. Forcade, Physical Review B **80**, 014120 (2009).

[46] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, ACM Transactions in Mathematical Software **22**, 469 (1996).

[47] L. J. Nelson, V. Ozoliņš, C. S. Reese, F. Zhou, and G. L. W. Hart, Physical Review B **88**, 155105 (2013).

[48] B. Sadigh and P. Erhart, Physical Review B **86**, 134204 (2012).

[49] A. Togo and I. Tanaka, "SPGLIB: a software library for crystal symmetry search," (2018), arxiv:1808.01590.

[50] https://icet.materialsmodeling.org/ (Accessed: 2019-01-10).

[51] P. E. Blöchl, Physical Review B **50**, 17953 (1994).

[52] G. Kresse and D. Joubert, Physical Review B **59**, 1758 (1999).

[53] G. Kresse and J. Furthmüller, Physical Review B **54**, 11169 (1996).

[54] G. Kresse and J. Furthmüller, Computational Materials Science **6**, 15 (1996).

[55] M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, and B. I. Lundqvist, Physical Review Letters **92**, 246401 (2004).

[56] J. Klimeš, D. R. Bowler, and A. Michaelides, Physical Review B **83**, 195131 (2011).

[57] K. Berland and P. Hyldgaard, Phys. Rev. B **89**, 035412 (2014).

[58] L. Gharaee, P. Erhart, and P. Hyldgaard, Physical Review B **95**, 085147 (2017).

[59] A. Dinsdale, A. Watson, A. Kroupa, J. Vrestal, A. Zemanova, and J. Vizdal, *Atlas of phase diagrams for lead-free soldering, COST 531 lead-free solders*, Vol. 1 (European Science Foundation, 2008).

[60] M. Christensen, S. Johnsen, and B. B. Iversen, Dalton Transactions **39**, 978 (2010).

[61] A. V. Shevelkov and K. Kovnir, in *Zintl phases*, Structure and Bonding, edited by T. F. Fässler (Springer, Heidelberg, 2011) p. 97.

[62] J. H. Roudebush, C. de la Cruz, B. C. Chakoumakos, and S. M. Kauzlarich, Inorganic Chemistry **51**, 1805 (2012).

[63] J. P. Perdew, K. Burke, and M. Ernzerhof, Physical Review Letters **77**, 3865 (1996), erratum, *ibid.* **78**, 1396(E) (1997).

[64] D. O. Lindroth, J. Brorsson, E. Fransson, F. Eriksson, A. Palmqvist, and P. Erhart, "Thermal conductivity in intermetallic clathrates: A first principles perspective," (2019), arxiv:1807.01502.

[65] V. Blum, G. L. W. Hart, M. J. Walorski, and A. Zunger, Phys. Rev. B **72**, 165113 (2005).

[66] F. Eriksson, E. Fransson, and P. Erhart, Advanced Simulation and Theory (2019), accepted.

[67] E. Fransson, F. Eriksson, and P. Erhart, "Efficient construction of high-order force constant models for low symmetry systems using machine learning," (2019), to be submitted.

[68] J. M. Cowley, Journal of Applied Physics **21**, 24 (1949).