Computational Materials Science 217 (2023) 111897

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

Full length article

CASM — A software package for first-principles based study of multicomponent crystalline solids

Brian Puchala [a,*,1], John C. Thomas [b,*,1], Anirudh Raju Natarajan [b], Jon Gabriel Goiri [b], Sesha Sai Behara [b], Jonas L. Kaufman [b], Anton Van der Ven [b,*]

[a] Department of Materials Science and Engineering, University of Michigan, Ann Arbor, United States of America
[b] Materials Department, University of California, Santa Barbara, United States of America

ARTICLE INFO

Keywords:
Software
Casm
Ab initio
First principles
Open source
Thermodynamics
Kinetics
Alloy
Crystal structure
Symmetry
Crystal structure enumeration
Cluster expansion
Effective Hamiltonian
Monte Carlo
Kinetic Monte Carlo

ABSTRACT

CASM is a software package that enables first-principles based studies of crystalline materials. It has been designed to treat coupled chemical, mechanical, vibrational, and magnetic degrees of freedom to determine ground state and finite temperature properties of crystals. The symmetry of the underlying parent crystal structure is used to enumerate perturbations of the parent crystal structure and generate derivative structures which can be input to first-principles calculations in order to explore the ground state energy landscape. CASM algorithmically constructs cluster expansions that fully couple discrete and continuous degrees of freedom, and generates highly efficient code to evaluate the cluster expansion basis functions. Widely used machine learning methods are integrated for fitting expansion coefficients to first-principle calculations. The fully parameterized cluster expansions can be combined with (kinetic) Monte Carlo methods to calculate finite temperature thermodynamic and kinetic properties. CASM Alloy Manager identifies distinct parent crystal structures in an alloy system, creating individual projects for each, and integrating the results. The integrated infrastructure facilitates the linkage between first-principles statistical mechanics predictions with higher length scale computational methods, such as phase field simulations. CASM projects are designed to be easy to share in repositories, to re-use, and to extend to include additional chemical species or types of degrees of freedom.

1. Introduction

First-principles electronic structure methods as applied to materials have reached a remarkable level of sophistication and accuracy. These methods are capable of predicting important properties at zero Kelvin, including lattice parameters, elastic moduli, the relative stability between different crystal structures, and a range of electronic properties [1]. For most applications, however, further analysis is necessary to account for the role of temperature [2,3]. The atoms and electrons of a solid evolve dynamically over time at finite temperature, even in thermodynamic equilibrium, with a solid visiting large numbers of microstates during the course of a measurement of common thermodynamic and mechanical properties. Materials properties at finite temperature are in essence thermal averages over the visited microstates. A first-principles prediction of materials properties at elevated temperature, therefore, requires a statistical mechanics approach.

Multicomponent crystals can exhibit many different degrees of freedom at the atomic and electronic length scales that lead to excited states at finite temperature [3,4]. For example, the atoms of a crystal vibrate around their equilibrium sites. Vibrational contributions to thermodynamic properties are often treated within the (quasi) harmonic approximation [5,6], with anharmonicity becoming important at high temperatures [7–12]. Alloys exhibit additional degrees of freedom due to the disorder that emerges when several chemical species share sites in a crystal [2]. Many materials for functional applications exhibit localized electronic states, such as local magnetic moments or localized charge states, that can disorder spatially over the sites of their crystal [13–16]. There is also an increased interest in molecular crystals, with molecular building blocks adopting multiple orientations, leading to other forms of disorder [17,18]. The excitation of atomic and electronic degrees of freedom leads to entropy and a temperature dependence of thermodynamic and mechanical properties more generally.

Monte Carlo techniques play a central role in modern statistical mechanics approaches to first-principles thermodynamics and kinetics. They explicitly sample microstates within finite-sized super cells of a

* Corresponding authors.
E-mail addresses: bpuchala@umich.edu (B. Puchala), johnct@umich.edu (J.C. Thomas), avdv@ucsb.edu (A. Van der Ven).
[1] Authors contributed equally.

https://doi.org/10.1016/j.commatsci.2022.111897
Received 4 June 2022; Received in revised form 26 October 2022; Accepted 2 November 2022
Available online 21 November 2022

crystal in order to calculate thermal averages [19–22] or to evaluate Green–Kubo expressions of transport coefficients [23–28]. A crucial ingredient for Monte Carlo techniques is a way to calculate the energy of each sampled microstate. First-principles electronic structure methods are too costly to enable a direct calculation of the energies for all the microstates sampled in Monte Carlo simulations. Instead, surrogate models are necessary that can interpolate and extrapolate a robust training set of first-principles energies.

The cluster expansion approach, first introduced by Sanchez et al. [29], serves as a rigorous mathematical framework with which to construct surrogate models for the energy of a crystal as a function of atomic and electronic degrees of freedom. Cluster expansions have been formulated to treat chemical disorder [2,29–32], magnetic disorder of non-collinear magnetic moments [33] and orientational disorder among molecules in a molecular crystal [18]. The traditional lattice dynamical Hamiltonians for harmonic and anharmonic vibrational excitations also have a mathematical form that is similar to chemical and magnetic cluster expansions in that they are expressed as polynomials of atomic degrees of freedom associated with clusters of sites [5,6]. The similarity among the different *effective* Hamiltonians suggests a natural generalization of the cluster expansion approach to a wide variety of atomic and localized electronic degrees of freedom within a crystal [3,34].

This article describes the Clusters Approach to Statistical Mechanics (CASM) software package, a toolset for automating the construction and parameterization of generalized cluster expansions and utilizing them within (kinetic) Monte Carlo simulations to predict thermodynamic and kinetic properties. CASM has been designed to treat a multitude of atomic and electronic degrees of freedom simultaneously, exploiting symmetry and group theoretical techniques to algorithmically generate generalized cluster expansions. The software provides powerful tools with which to systematically enumerate mechanical, vibrational, chemical and magnetic configurations within a parent crystal structure. This not only enables systematic exploration of the energy landscape of a particular crystal, but also generates the training data needed to parameterize cluster expansions of the energy.

In studying the thermodynamics of a system that contains multiple components, a distinction needs to be made between excitations that determine the thermodynamic properties of a parent crystal structure and global phase stability between different parent crystal structures [35]. The thermodynamic properties of a particular phase arise from symmetry breaking perturbations of a higher symmetry reference parent crystal structure. Ordering reactions to form compounds and/or symmetry breaking shuffles and strains may occur on a parent crystal structure through group/subgroup phase transition mechanisms [3]. These phenomena can be described with a single CASM project and are the topic of Section 2. In general, though, multiple parent crystal structures, and their derived group/subgroup ordered and shuffled phases, compete for global phase stability. Section 3 summarizes CASM Alloy Manager, which supervises multiple CASM projects and assembles a comprehensive description of thermodynamic properties within a composition space to determine global phase stability. Section 4 describes how properties calculated with CASM can interface with higher length scale models of thermo-mechano-kinetic phase evolution. It also describes how data generated by CASM can be stored within repositories such as Materials Commons [36] and how it can be accessed and built upon by other users.

2. A CASM project

Fig. 1 schematically illustrates the capabilities and associated workflow of a CASM project. A CASM project is restricted to excitations on a particular parent crystal structure. It starts with the specification of the parent crystal structure and a set of degrees of freedom that are deemed to be significant for predicting properties of interest. CASM provides methods to enumerate distinct chemical and crystallographic perturbations to the parent crystal structure, exploiting the symmetry of the underlying parent crystal structure to create a database of symmetrically unique derived crystal structures. The derivative structures can then be fed into a first-principles electronic structure code, such as VASP [37–39], QUANTUM ESPRESSO [40,41], or FHI-AIMS [42], to calculate their energies and other electronic-structure related properties. Independent of dataset construction, CASM algorithmically constructs cluster expansion basis functions for the specified degrees of freedom, fully compliant with the symmetry constraints of the parent crystal structure, and generates tailor-made code optimized to efficiently evaluate the basis functions. Regression analysis can be used to parameterize the coefficients of the cluster expansion using the first-principles energies for the derivative structures as training data. The parameterized cluster expansion can then be used in highly efficient Monte Carlo calculations of thermodynamic and kinetic quantities. Each component of the CASM workflow is described in more detail in the following subsections.

2.1. Parent crystal structure

A CASM project requires the specification of a parent crystal structure. The parent crystal structure serves as the reference to which chemical and structural perturbations are applied. Usually the parent crystal structure is a high symmetry phase, making it possible to study group/subgroup symmetry breaking ordering and structural transitions. A crystal is determined by a unit cell and a collection of basis sites. The unit cell vectors, $\vec{l}_1$, $\vec{l}_2$ and $\vec{l}_3$ generate a lattice. The basis sites have coordinates $\vec{b}_1$, $\vec{b}_2$, ... and represent a motif within the unit cell that is translated to all the lattice sites to generate the full crystal. A site i in the full crystal with coordinates $\vec{r}_i = n_1\vec{l}_1 + n_2\vec{l}_2 + n_3\vec{l}_3 + \vec{b}_b$, is completely specified by the integers n_1, n_2 and n_3, which determine the lattice site, and b, which is an index that specifies the basis site.

2.2. Degrees of freedom

Most crystalline materials of technological interest exhibit a variety of site specific atomic and electronic degrees of freedom. These can be described with variables assigned to each site i of a parent crystal structure that measure the state at site i. For example, atoms vibrate around high symmetry sites and may undergo collective shuffles to lower symmetry configurations at low temperature. The displacement of each atom away from its position in the reference parent crystal structure can be tracked with a displacement vector $\vec{d}_i = (d_x^{(i)}, d_y^{(i)}, d_z^{(i)})$. The displacements at all sites can be collected in a *state* vector $\vec{\mathbf{d}} = (\vec{d}_1, \ldots, \vec{d}_i, \ldots, \vec{d}_N)$.

It is common that some sublattices of a parent crystal structure accommodate multiple chemical species. These may order at low temperature, but become disordered at elevated temperature. The chemical occupant of a crystal site i can be tracked with occupation variables, s_i, that identify the species at site i in a particular configuration [2]. If, for example, the crystal sublattice with index b hosts three chemical species, A, B and C, then the occupation variables s_i can have one of three values. In CASM, the values of s_i would be 0, 1 or 2, depending on whether A, B or C occupies site i. Other occupation variables are often also used, such as $\sigma_i = \phi_b(s_i)$, where the site function ϕ_b returns other values depending on the occupant at site i. A common set of values for a ternary alloy is $\sigma_i = [-1, 0, 1]$ [2,29]. The chemical arrangement over the sites of the crystal can be tracked with the state vector $\vec{\mathbf{s}} = (s_1, \ldots, s_i, \ldots, s_N)$.

Many compounds and alloys are magnetic and have localized non-collinear magnetic moments at a subset of sites or clusters of sites of the crystal [15,16,33]. Similar to the other degrees of freedom, the local magnetic moment at a site i can be represented with the vector $\vec{m}_i = (m_x^{(i)}, m_y^{(i)}, m_z^{(i)})$. Usually only the orientation of the magnetic moment is

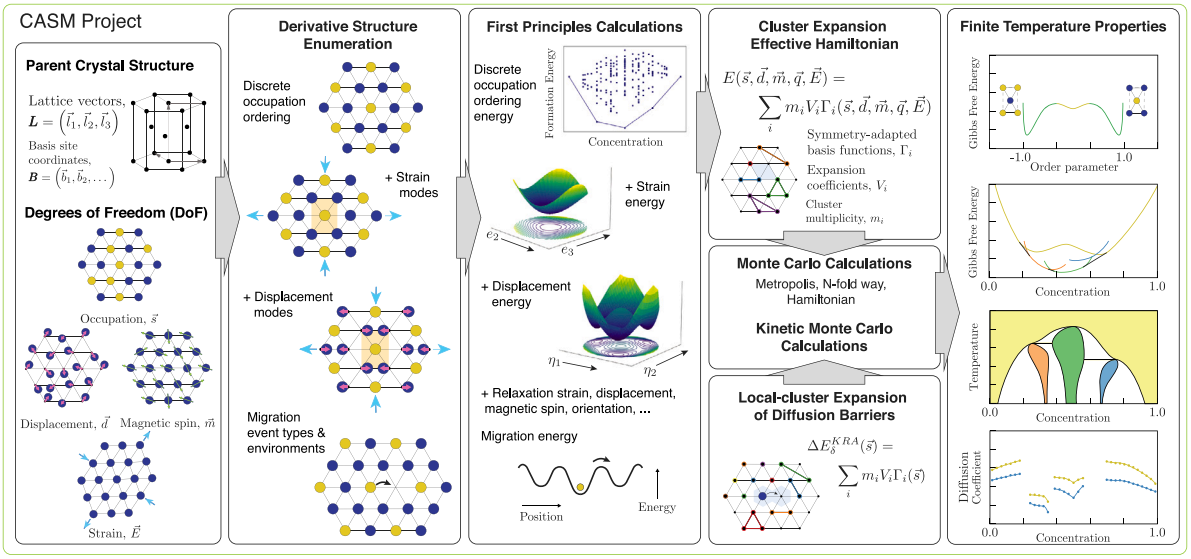


Fig. 1. Schematic of a CASM project, from specification of the parent crystal structure and allowed degrees of freedom, through derivative structure enumeration, first principles calculation, and (local-) cluster expansion construction and fitting, to calculation of finite temperature thermodynamic and kinetic properties.

important and the $\vec{m}_i$ can be treated as unit vectors. The magnetic state of the crystal is specified by the state vector $\vec{\mathbf{m}} = (\vec{m}_1, \dots \vec{m}_i, \dots)$.

There is an increased interest in molecular crystals and hybrid organic–inorganic crystals in which an inorganic crystal can host complex organic molecules [43,44]. The molecules of these crystals can often adopt different orientations [17], opening up the possibility of orientational order–disorder phenomena [18]. Many surface catalysis problems also involve the adsorption of molecular species at regular sites of a crystalline substrate. The adsorbed molecules can often also adopt different orientations. The orientation of a molecule at a particular site can be tracked by a quaternion vector, $\vec{q} = (q_0, q_1, q_2, q_3)$, which is a unit vector in a four dimensional space that can specify the orientation of a molecule relative to a reference orientation [18].

More abstract site degrees of freedom are also possible. These tend to be discrete degrees of freedom. For example, some transition metal oxides may exhibit orbital ordering, in which electrons can occupy one of several possible d-orbitals centered on the transition metal [14,45]. Other discrete site degrees of freedom include collinear magnetic states and a discrete set of molecular orientations on a surface or within a crystal. Similar to chemical degrees of freedom, these discrete states can be tracked with occupation variables s_i assigned to each site i [13,15]. Additional subtleties emerge since the application of symmetry to a particular discrete state may map the state into another discrete state. This contrasts with chemical occupants, which are invariant to symmetry operations of the crystal. CASM has been designed to accurately treat the effect of symmetry on more abstracted discrete degrees of freedom.

A final degree of freedom is macroscopic in nature and is related to homogeneous deformations of the crystal [46]. The deformation of a crystal can be described by a deformation gradient $\mathbf{F}$, which is a 3×3 matrix that relates the lattice vectors of the deformed crystal, $\mathbf{L}' = (\vec{l}'_1, \vec{l}'_2, \vec{l}'_3)$ to the lattice vectors of the reference parent crystal structure $\mathbf{L} = (\vec{l}_1, \vec{l}_2, \vec{l}_3)$ according to $\mathbf{L}' = \mathbf{FL}$. The deformation gradient $\mathbf{F}$ contains not only information about the deformation of the crystal, but also about any rigid rotation of the crystal. It can be factored into a product of a symmetric stretch tensor $\mathbf{U}$, which describes the deformation of the unit cell, with a rotation matrix $\mathbf{R}$ according to $\mathbf{F} = \mathbf{RU}$. Since the $\mathbf{R}^\top\mathbf{R} = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix, it is possible to define metrics of deformation that are independent of $\mathbf{R}$ in terms of $\mathbf{F}^\top\mathbf{F} = \mathbf{U}^\top\mathbf{U}$. Two useful metrics of strain are the Green–Lagrange strain $\mathbf{E} = 1/2(\mathbf{F}^\top\mathbf{F} - \mathbf{I})$ and the Hencky strain $\mathbf{E} = 1/2 \ln(\mathbf{F}^\top\mathbf{F})$. The strain tensors $\mathbf{E}$ are symmetric 3×3 matrices, containing six independent components. The state of strain of the crystal can therefore be uniquely specified with a vector of independent strain components $\vec{E} = (E_{11}, E_{22}, E_{33}, \sqrt{2}E_{23}, \sqrt{2}E_{13}, \sqrt{2}E_{12})$ [46].

Table 1
A summary of the types of parent crystal systems for which CASM can construct (local-) cluster expansions, within the limitations of computational complexity.

Parent crystal systems treated by CASM
Any choice of lattice vectors and any number of sublattices
Any number of occupants, specified per sublattice
Occupants may represent atomic type and discrete values of charge state, spin state, or molecular orientation
Displacement or magnetic spin continuous site degrees of freedom, specified per sublattice
Strain continuous degrees of freedom
Continuous degrees of freedom may be restricted to arbitrary subspaces
Coupled cluster expansions
Local-cluster expansions

CASM offers users a menu of degrees of freedom that can be specified to be active at different sublattices (basis sites) of the crystal. Furthermore, the strain of the crystal can also be treated as a degree of freedom within CASM. Degrees of freedom may also be restricted to user-specified subspaces, allowing, for example, only volumetric strains, or only displacements in two-dimensional subspaces. A summary of the types of parent crystal systems that can be treated by CASM is included in Table 1. CASM can be extended to treat additional degrees of freedom through the definition of traits that specify its dimension, how it transforms under the application of a symmetry operation, whether it is a site or global property, and how basis functions should be generated.

2.3. Generalized cluster expansions

The cluster expansion formalism, as introduced by Sanchez et al. [29,32], is a mathematical approach with which to systematically construct expansions to describe crystal properties as a function of site degrees of freedom. The simplest example is a cluster expansion for a binary A–B alloy having a parent crystal structure that forms a lattice such as fcc or bcc (i.e. only one basis site). By assigning occupation variables σ_i to each site i that are +1 (−1) if occupied by B (A), it is possible to show that the energy of the crystal as a function of the arrangement of A and B atoms $\vec{\sigma} = (\sigma_1, \dots, \sigma_i, \dots)$ can be expressed

as [29]

$$E(\vec{\sigma}) = \sum_{\alpha} V_{\alpha} \Phi_{\alpha}(\vec{\sigma}) \tag{1}$$

where α indexes clusters of sites such as points, pairs, triplets, etc. The Φ_{α}, referred to as cluster functions, are defined as

$$\Phi_{\alpha}(\vec{\sigma}) = \prod_{i \in \alpha} \sigma_i. \tag{2}$$

For a binary alloy the cluster functions are simply products of the occupation variables of the sites belonging to the cluster α. The V_{α} are expansion coefficients whose values are determined by the specific chemical properties of the alloy. The cluster functions Φ_{α} of Eq. (2) can be shown to form a complete basis for any scalar function of the configuration, $\vec{\sigma}$, of the crystal [2,29,32]. For a binary alloy on a parent crystal structure with N sites, there are exactly 2^N configurations $\vec{\sigma}$, and there are an equal number of cluster functions Φ_{α}, one for each distinct cluster of sites within the crystal. While the cluster expansion, Eq. (1), extends over all 2^N clusters, in practice it must be truncated beyond some maximally sized cluster. The expansion coefficients of a truncated cluster expansion can then be trained to a large data set of energies for different configurations as calculated with a first-principles electronic structure method.

Cluster expansions of the energy of more complicated alloys that involve more than two species [2,29,30,47], possibly distributed over multiple sublattices [48–51], have a similar form as Eq. (1). The energy can again be expressed as a sum of chemistry dependent coefficients that multiply basis functions associated with clusters of sites α. When more than two species share a sublattice, however, the basis functions are more complex and there are multiple polynomial basis functions for each cluster α. Similar types of cluster expansions have been derived in terms of magnetic moments $\vec{m}_i$ of magnetic solids [16,33] and in terms of quaternions $\vec{q}_i$ that track the orientation of molecules in molecular crystals [18].

Lattice dynamical Hamiltonians that describe the energy of a crystal as a function of atomic displacements away from their sites in a high symmetry crystal also have a form that is similar to that of cluster expansions [5,6]. These Hamiltonians are traditionally motivated by expressing the energy of the crystal as a multi-variable Taylor expansion in terms of the displacement $\vec{d}_i$ at each site i and take the form

$$E(\vec{\mathbf{d}}) = E_0 + \sum_{i,j} \sum_{\alpha,\beta} \Lambda^{i,j}_{\alpha,\beta} d^{(i)}_{\alpha} d^{(j)}_{\beta} + \sum_{i,j,k} \sum_{\alpha,\beta,\gamma} \Lambda^{i,j,k}_{\alpha,\beta,\gamma} d^{(i)}_{\alpha} d^{(j)}_{\beta} d^{(k)}_{\gamma} + \cdots. \tag{3}$$

where the indices i, j and k represent sites in the crystal and where the Greek indices represent Cartesian components of the displacement vectors. The coefficients $\Lambda^{i,j}_{\alpha,\beta}$, $\Lambda^{i,j,k}_{\alpha,\beta,\gamma}$, etc. are again chemistry dependent. The expansion coefficients of a truncated lattice dynamical Hamiltonian can be parameterized by training to a data set of first-principles calculated energies and forces [5,52,53]. As is clear from Eq. (3), the energy of the crystal as a function of atomic displacement can be expressed as a sum over functions of the displacement variables belonging to clusters of sites. The harmonic approximation is adopted if only terms up to second order in the expansion are kept (i.e. corresponding to point and pair clusters). Higher order terms in the expansion describe anharmonicity in the energy of the crystal as a function of the displacement degrees of freedom $\vec{\mathbf{d}}$. Translational and rotational invariance as well as the symmetry of the crystal imposes constraints on the expansion coefficients $\Lambda^{i,j}_{\alpha,\beta}$, $\Lambda^{i,j,k}_{\alpha,\beta,\gamma}$, etc., thereby reducing the number of independent Hamiltonian parameters that need to be estimated.

The above examples motivate a generalization of the cluster expansion to treat a wide variety of local and global degrees of freedom simultaneously [34]. This becomes important when the sites of a crystal host chemical, displacement and/or magnetic degrees of freedom and when strain is treated as an explicit degree of freedom. As with the alloy and magnetic cluster expansions and similar to the lattice dynamical Hamiltonians, the first step is to generate cluster functions that are polynomials of the multiple site degrees of freedom belonging to a cluster α. The space group symmetry operations of the parent crystal structure can then be applied to the cluster functions to ensure that the resultant expansion of the energy is invariant to the symmetry of the parent crystal structure. The details of generating symmetry compliant basis functions are described elsewhere [34]. The resulting generalized cluster expansion can be expressed as

$$E(\vec{\mathbf{s}}, \vec{\mathbf{d}}, \vec{\mathbf{m}}, \vec{E}) = \sum_{\alpha,n} V^n_{\alpha} \Phi^n_{\alpha}(\vec{\mathbf{s}}, \vec{\mathbf{d}}, \vec{\mathbf{m}}, \vec{E}) \tag{4}$$

where Φ^n_{α} are the cluster functions that depend on the site degrees of freedom in addition to the global strains of the crystal. The sum extends over all possible clusters of sites α and all independent and symmetry allowed cluster functions, indexed with n, involving site degrees of freedom that belong to the cluster α [34].

By exploiting the underlying symmetry of the parent crystal structure, it is possible to reduce the number of unknown expansion coefficients V^n_{α} appearing in Eq. (4). All cluster functions that are related to a prototype cluster function Φ^n_{α} by a space group symmetry operation of the parent crystal structure belong to an orbit, denoted as Ω^n_{α}. The cluster functions of a particular orbit have the same numerical value for their corresponding expansion coefficient. Hence, Eq. (4) can be rewritten as a sum first over orbits of cluster functions followed by a sum over individual cluster functions within each orbit according to

$$E(\vec{\mathbf{s}}, \vec{\mathbf{d}}, \vec{\mathbf{m}}, \vec{E}) = \sum_{\Omega^n_{\alpha}} V^n_{\alpha} \left(\sum_{\Phi^m_{\beta} \in \Omega^n_{\alpha}} \Phi^m_{\beta}(\vec{\mathbf{s}}, \vec{\mathbf{d}}, \vec{\mathbf{m}}, \vec{E}) \right). \tag{5}$$

This refactorization, made possible by the underlying symmetry of the crystal, motivates the introduction of generalized correlation functions defined as

$$\Gamma^n_{\alpha}(\vec{\mathbf{s}}, \vec{\mathbf{d}}, \vec{\mathbf{m}}, \vec{E}) = \frac{1}{m^n_{\alpha} N_u} \left(\sum_{\Phi^m_{\beta} \in \Omega^n_{\alpha}} \Phi^m_{\beta}(\vec{\mathbf{s}}, \vec{\mathbf{d}}, \vec{\mathbf{m}}, \vec{E}) \right) \tag{6}$$

where N_u is the number of primitive unit cells in the crystal and where m^n_{α} is the multiplicity of symmetrically equivalent cluster functions per primitive cell. In terms of the correlation functions, the energy of the crystal, normalized by the number of primitive cells N_u, can then be written as

$$e(\vec{\mathbf{s}}, \vec{\mathbf{d}}, \vec{\mathbf{m}}, \vec{E}) = \sum_{\Omega^n_{\alpha}} m^n_{\alpha} V^n_{\alpha} \Gamma^n_{\alpha}(\vec{\mathbf{s}}, \vec{\mathbf{d}}, \vec{\mathbf{m}}, \vec{E}) \tag{7}$$

The generalized correlation functions as defined by Eq. (6) serve as descriptors of a particular state of the solid determined by the particular values of $\vec{\mathbf{s}}$, $\vec{\mathbf{d}}$, $\vec{\mathbf{m}}$ and $\vec{E}$. Since they are obtained by averaging over all cluster functions that are equivalent by symmetry, they have the same value within all states of the crystal that are symmetrically equivalent.

Cluster expansions can be used to represent any property of a crystal, not just the total energy. For example, when only explicitly treating chemical degrees of freedom, a cluster expansion can be used to represent the fully relaxed volume, or relaxation strains more generally [34,54], as a function of chemical ordering $\vec{\mathbf{s}}$.

Given site-to-site distance cutoffs for pair, triplet, etc. clusters, and a user-specified maximum polynomial order for basis functions of continuous degrees of freedom, CASM algorithmically constructs symmetry invariant cluster functions which fully couple all degrees of freedom present on sites in the clusters, and if present, strain degrees of freedom. CASM then generates optimized code which is compiled and used to evaluate the basis functions and calculate correlations for any configuration.

As an illustration, Fig. 2 shows plots of the number of basis functions associated with different prototype clusters when considering a variety of degrees of freedom on a parent fcc crystal. The basis functions were generated with CASM. Representative pair, triplet and quadruplet clusters were chosen with different cluster point group sizes to illustrate

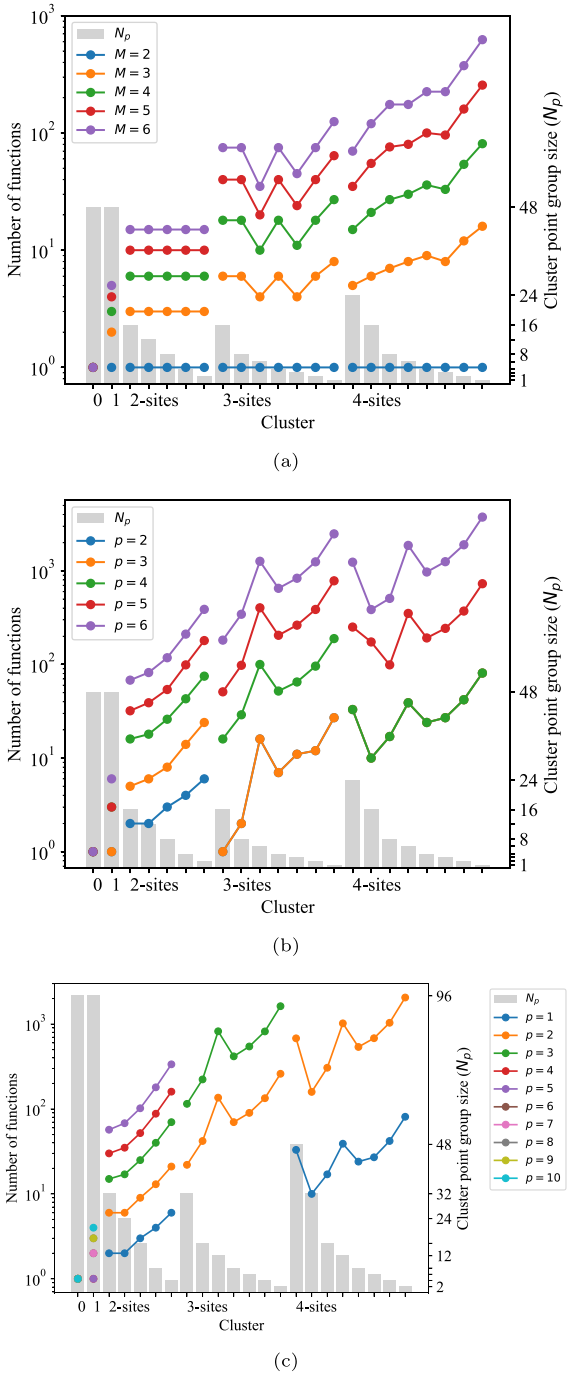


Fig. 2. Counts of prototype cluster functions, Φ^n_α, associated with clusters in an FCC crystal of varying number of sites and point group size. (a) Occupation basis functions, $\Phi^n_\alpha(\vec{\mathbf{s}})$, with number of alloying elements, M, varying from 2 to 6. (b) Displacement basis functions, $\Phi^n_\alpha(\vec{\mathbf{d}})$, with maximum polynomial order, p, varying from 2 to 6. (c) Prototype cluster functions of magnetic spin variables, $\Phi^n_\alpha(\vec{\mathbf{m}})$. Data was generated for clusters satisfying $p * N_s \leq 10$, where N_s is the number of sites in the cluster, p is the maximum polynomial order of harmonic site basis functions, and $p * N_s$ is the maximum polynomial order of the cluster functions.

how symmetry affects the number of symmetrically distinct cluster functions.

Fig. 2(a) shows the number of prototype cluster functions, $\Phi^n_\alpha(\vec{\mathbf{s}})$, when treating only chemical degrees of freedom. While in a binary alloy ($M = 2$) there is only one basis function per cluster (2-site, 3-site, 4-site), that number increases rapidly with the number of chemical components $M > 2$. Fig. 2(a) also shows that the number of cluster basis

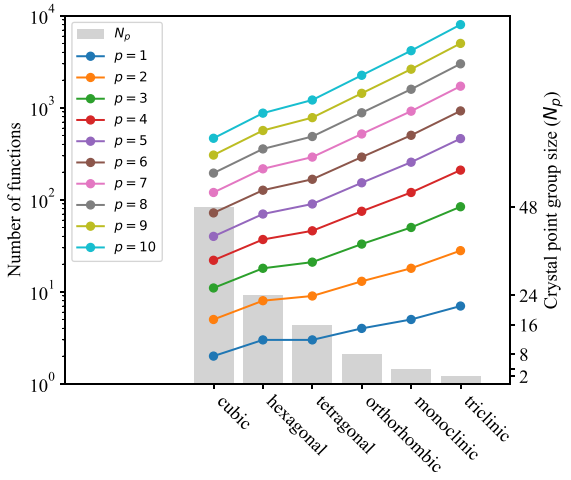


Fig. 3. Count of strain basis functions, $\Phi^n_\alpha(\vec{E})$, using the Hencky strain metric, for representative structures from each crystal family, with maximum polynomial order, p, varying from 1 to 10.

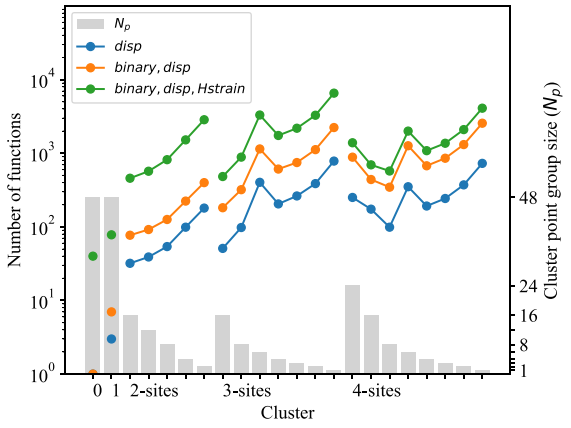


Fig. 4. Counts of prototype cluster functions for a system that couples chemical occupation, displacement, and Hencky strain degrees of freedom, $\Phi^n_\alpha(\vec{\mathbf{s}}, \vec{\mathbf{d}}, \vec{E})$, with maximum polynomial order 5.

functions for 3-site and 4-site clusters in fcc can be sensitive to the point group of the cluster. Fig. 2(b) shows the number of prototype cluster functions when considering only displacement degrees of freedom on each site of an fcc crystal. Each curve corresponds to a maximum polynomial order of displacement variables. Cluster functions with polynomial order of two form the basis of an harmonic lattice dynamical Hamiltonian, while those with polynomial order above two serve to capture anharmonicity in the energy surface. Fig. 2(c) shows the number of prototype cluster functions of magnetic spin degrees of freedom, $\Phi^n_\alpha(\vec{\mathbf{m}})$, for varying maximum polynomial order. These consist of sums of products of cubic harmonics, each cubic harmonic being a function of a site magnetic moment of the cluster. The linear combinations of products of cubic harmonics ensure that the resulting basis function is invariant to the point group symmetry of the cluster [34]. A general trend in all three plots is that clusters having a higher symmetry (as measured by the number of cluster point group sizes) tend to have fewer basis functions for the same cluster size, although the type of symmetry operations within the cluster point group also has an effect.

Fig. 3 shows the number of symmetry invariant strain functions, $\Phi^n_\alpha(\vec{E})$, for representative crystal structures from each crystal family. These basis functions are polynomials of the components of the 6-dimensional strain vector $\vec{E} = (E_{11}, E_{22}, E_{33}, \sqrt{2}E_{23}, \sqrt{2}E_{13}, \sqrt{2}E_{12})$ and serve as a basis with which to expand the energy of a crystal as

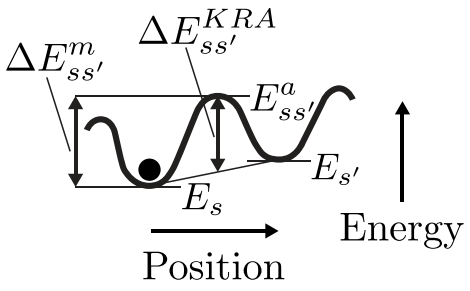


Fig. 5. Schematic showing the kinetically resolved activation energy, $\Delta E^{KRA}_{ss'}$, for a diffusion event, in terms of the energies of the initial state E_s, final state $E_{s'}$, and activated state $E^a_{ss'}$. Expressing the migration energy, $\Delta E^m_{ss'}$, for an event in terms of the kinetically resolved activation energy, $\Delta E^m_{ss'} = \Delta E^{KRA}_{ss'} + (E_s + E_{s'})/2$, ensures the forward and reverse event rates are consistent.

a function of strain [46]. The number of symmetry invariant strain polynomial basis functions increases as the symmetry decreases.

As a final example, Fig. 4 shows the number of prototype symmetry invariant cluster functions for a system that couples binary chemical occupational degrees of freedom with displacement and strain degrees of freedom, with cluster functions of maximum polynomial order 5. The number of basis functions per cluster is clearly much larger in a system that couples multiple degrees of freedom than systems where only one degree of freedom is treated at a time. It should be recognized though that since cluster expansions are always truncated in practice, the number of terms in a fully converged cluster expansion may be similar when explicitly treating multiple degrees of freedom versus treating only one degree of freedom. For example, when constructing a cluster expansion for an alloy that only explicitly treats chemical occupational degrees of freedom, it is common to train the expansion coefficients to fully relaxed energies. Such a cluster expansion then implicitly accounts for the effects of strain and displacement. As a result, while most chemical interactions tend to be short-ranged in the absence of relaxations, relaxations will introduce long-range effective interactions that require large numbers of basis functions belonging to clusters that extend over large distances in a cluster expansion that only treats chemical degrees of freedom. This has for example been addressed with a mixed-basis cluster expansion to describe long-range elastic interactions in alloys [55]. A cluster expansion that treats chemical occupational, displacement and strain degrees of freedom simultaneously, while having many more basis functions per cluster, is expected to only require small prototype clusters in the expansion due to the short-ranged nature of most chemical interactions.

The number of calculations needed to train the expansion coefficients of different cluster expansions is also not necessarily very different. For example, the expansion coefficients of a cluster expansion that only explicitly treats chemical occupational degrees of freedom are trained to fully relaxed energies, which each typically require 10–50 static first-principles electronic structure calculations. The training of the expansion coefficients of a coupled cluster expansion that combines chemical occupational, displacement and strain degrees of freedom may require many more configurations, however, each of these will only require one static calculation.

2.4. Local-cluster expansion

Cluster expansions can also be used to parameterize local properties as a function of the surrounding environment. These are referred to as local-cluster expansions. Local-cluster expansions are especially useful in describing the local environment dependence of the migration barrier for diffusion [23,27,56] and the dependence of generalized stacking fault energies on chemical disorder around the fault plane [57].

Diffusion in concentrated multicomponent solids arises from large numbers of elementary atomic hops that connect one microstate s, corresponding to a particular chemical decoration of the crystal $\vec{\mathbf{s}}$, to a neighboring microstate, s'. According to transition state theory within the harmonic approximation, the rate at which a system transitions from s to s' is [58]

$$R_{ss'} = \nu^*_{ss'} \exp\left(-\frac{\Delta E^m_{ss'}}{k_b T}\right), \tag{8}$$

where $\nu^*_{ss'}$ is a vibrational prefactor for the event, and $\Delta E^m_{ss'}$ is the migration energy, as shown schematically in Fig. 5.

While the migration energy depends on the direction of the hop, the energy of the activated state at the saddle point between the initial and final states, $E^a_{ss'}$, must be directionally independent: $E^a_{ss'} == E^a_{s's}$. The kinetically resolved activation energy, defined as [23]

$$\Delta E^{KRA}_{ss'} = E^a_{ss'} - \frac{E_s + E_{s'}}{2}, \tag{9}$$

is a property of the local environment that is independent of hop direction and that can be used to calculate $E^a_{ss'}$ and $\Delta E^m_{ss'}$ consistently for both the forward and reverse events.

Since $\Delta E^{KRA}_{ss'}$ is a property of the local environment, it can be parameterized using a local-cluster expansion according to [23]

$$\Delta E^{KRA}_{ss'} = \Delta E^{KRA}_{\delta}(\vec{\mathbf{s}}) \tag{10}$$

$$= \sum_{\Omega^n_\alpha} m^n_\alpha V^n_\alpha \Gamma^n_\alpha(\vec{\mathbf{s}}), \tag{11}$$

where δ indicates the type of hop event connecting s to s', and vice versa, $\vec{\mathbf{s}}$ is the collection of occupation variables associated with state s, Ω^n_α is an orbit of symmetrically equivalent local-cluster functions, determined by the point group symmetry of δ, m^n_α is the total number of symmetrically equivalent cluster functions, V^n_α are expansion coefficients, and $\Gamma^n_\alpha(\vec{\mathbf{s}}) = \frac{1}{m^n_\alpha}\sum_{\Phi_\beta \in \Omega^n_\alpha} \Phi^n_\beta(\vec{\mathbf{s}})$ are the local correlation functions. By combining a local-cluster expansion for $\Delta E^{KRA}_{ss'}$ with a cluster expansion for the configurational energy of the crystal (i.e. the end states E_s and $E_{s'}$), it is possible to describe the configuration dependence of the migration barriers according to

$$\Delta E^m_{ss'} = \Delta E^{KRA}_{ss'} + \frac{E_{s'} - E_s}{2}. \tag{12}$$

The type of hop event, δ, specifies the cluster of sites on which atomic exchanges occur and how the occupants permute among those sites. In CASM, the hop type δ also specifies the occupants that exchange sites. For example, in an A–B binary alloy in which diffusion is mediated by nearest neighbor atom–vacancy exchanges, there is one local-cluster expansion for A-vacancy exchanges and another for B-vacancy exchanges [27]. It is also possible that the type of occupants exchanging sites is not specified by δ. Then a single local-cluster expansion must be parameterized that explicitly depends on occupation variables belonging to the hop cluster to be able to distinguish hops performed by different chemical species.

Besides diffusion barriers, there are a variety of other local properties within a crystal that depend on local surroundings and that can therefore be described with a local-cluster expansion. CASM can algorithmically generate a local-cluster expansion about any cluster of sites given (i) the point group symmetry of the property associated with those sites, (ii) a cutoff distance to specify sites that should be included in the local environment, (iii) site-to-site distance cutoffs for pair, triplet, etc. clusters that are to form the truncated cluster expansion, and (iv) user-specified maximum polynomial order for basis functions of continuous degrees of freedom. The point group symmetry of the local property and the space group symmetry of the parent crystal structure is then used to generate equivalent local-cluster expansions for each symmetry equivalent but distinct local property in the primitive cell of the parent crystal structure. As with the periodic cluster expansions, the generated basis functions can couple chemical, displacement

Table 2
A summary of the types of derivative structure enumeration methods implemented in CASM.

Enumeration methods in CASM
Symmetrically unique supercells
Symmetrically unique orderings of discrete occupants in 1d-, 2d-, and 3d-periodic structures
Symmetrically unique orderings on a subset of sites within a structure, specified individually or by sublattice, cluster, or local-cluster expansion neighborhood
Symmetrically unique super-configuration orderings comprised of combinations of user-specified sub-orderings
Stochastic enumeration via Monte Carlo calculation (simulated annealing)
Stochastic generation of special quasi-random structures (SQS)
Enumeration of displacements, as excitations of symmetry-adapted collective modes
Enumeration of strain states, as points on a grid in an irreducible wedge of strain space
Filtered enumeration, using any of the above methods and filtering structures by composition, correlations, symmetry, etc.

and magnetic degrees of freedom in the local environment, and if present, strain degrees of freedom. For each equivalent-but-distinct local-cluster expansion CASM generates and compiles optimized code to evaluate the local-cluster expansion basis functions.

2.5. Calculating properties of derivative structures

CASM includes several methods to systematically enumerate, or stochastically generate, perturbations to a parent crystal structure for different degrees of freedom. A primary purpose of these methods is to construct a data set of first-principles calculated energies of the derivative structures in order to train cluster expansion coefficients. The calculation of the energy of derivative structures can also provide important information about ground states and structural instabilities that lead to group/subgroup symmetry breaking. Table 2 has a summary of the enumeration methods implemented in CASM and discussed in more detail in this section.

While CASM implements specialized algorithms to enumerate perturbations to a parent crystal structure for different types of degrees of freedom, each algorithm relies on the use of periodic supercells within which perturbations are enumerated or sampled. CASM can systematically enumerate supercells of the primitive unit cell of the parent crystal structure using well-established algorithms that rely on the Hermite normal form of a lattice [59]. The point group symmetry of the parent crystal structure is used to maintain a list of the symmetrically unique supercells.

2.5.1. Exhaustive enumeration of discrete degrees of freedom

Given a particular supercell, symmetrically unique configurations of the values of the degrees of freedom on the sites within the supercell may be enumerated. For discrete site degrees of freedom, such as chemical occupancy, an exhaustive enumeration of different orderings is possible. The space group symmetry of the parent crystal structure is again used to maintain a list of only symmetrically distinct configurations within a particular supercell.

The number of possible configurations that can be enumerated grows very rapidly with the number of sites within a supercell. In large supercells, it may be desirable to enumerate site degrees of freedom over only a subset of sites. CASM implements an exhaustive enumeration approach that can take as input a configuration with a particular background ordering and a selection of sites on which to perform enumeration. By default all sites are included, but in large supercells it is possible to specify a subset of sites, either directly, by sublattice, or as symmetrically distinct clusters. This approach can, for example, be used to enumerate anti-site defects of various cluster sizes in a particular ordering that may be a ground state in a multicomponent alloy.

CASM enables the enumeration of local configurations around the sites that participate in a diffusion event. In typical multicomponent systems, diffusion occurs both in disordered solid solutions and in well-ordered compounds. Local hop environments can be sampled within large supercells that have a background configuration that mimics the disordered state and within low-energy ordered phases. In many solids, diffusion involves only two sites, with an atom performing a direct exchange with a vacancy on a neighboring site. There are many important examples, however, where a diffusion event may involve three or more sites, due for example to cyclic hop events [60–62]. A diffusion event is, therefore, defined by the cluster of sites involved in the hop and the change in decoration of the cluster upon transitioning from the initial to the final state.

To enumerate local environments around the sites involved in a diffusion event a background configuration is first chosen and anti-site configurations on the sites within a user-specified range of the hop cluster can be exhaustively enumerated. The background configuration, the sites involved in the event, and the point group symmetry of the event provide all the necessary symmetry information to identify symmetrically distinct environments. The range of sites included in the local environment is specified by a cutoff radius, and enumeration of occupations can be performed exhaustively, or for a subset of sites specified directly, by sublattice, or as symmetrically distinct clusters.

2.5.2. Enumeration of continuous degrees of freedom

For continuous degrees of freedom, such as atomic displacements and the orientations of non-collinear magnetic moments, there are an infinite number of possible perturbations, but enumeration may be done when restricted to a discrete grid. A supercell with N sites has $3N$ displacement degrees of freedom in three dimensions, 3 of which correspond to rigid translations. A basis rotation can be used to align axes that span the remaining $3N - 3$ degrees of freedom in different directions, corresponding to different choices of collective displacement modes. When the axes are chosen to align with certain high-symmetry directions, they may be called symmetry-adapted axes and the corresponding collective modes are called symmetry-adapted collective modes. Any displacement field within the same supercell can be expressed as a linear combination of symmetry-adapted collective displacement modes, and the amplitudes of the contributions form symmetry-adapted displacement order parameters.

CASM algorithms use group theory to identify irreducible displacement subspaces [34,46], the lowest dimension subspaces which are not mixed upon application of operations in the symmetry group of the parent crystal structure. Within an irreducible subspace, high symmetry directions correspond to collective modes that are invariant to some parent crystal structure symmetry operations. CASM finds these directions so that they can be used to form a grid on which to enumerate displacement fields as linear combinations of symmetry-adapted collective modes.

As an illustration, consider an enumeration of displacements within a $\sqrt{3}a \times \sqrt{3}a$ supercell of a two-dimensional triangular lattice. Fig. 6 shows symmetry adapted collective displacements within the $\sqrt{3}a \times \sqrt{3}a$ supercell. In two dimensions there are a total of six displacement degrees of freedom within the $\sqrt{3}a \times \sqrt{3}a$ supercell as it contains three sites. Any displacement pattern in the supercell can be expressed as a linear combination of the six collective displacement modes of Fig. 6. The symmetry adapted collective displacement modes of Fig. 6 can be divided into three groups. The first two sets (Figs. 6(a) and (b)), consisting of two collective displacement modes each, span irreducible subspaces. Any point in those subspaces, corresponding to a collective displacement pattern that can be represented as a linear combination of the spanning symmetry-adapted collective modes, maps onto another point in the same subspace upon the application of a symmetry operation of the crystal that maps the $\sqrt{3}a \times \sqrt{3}a$ supercell onto itself. The

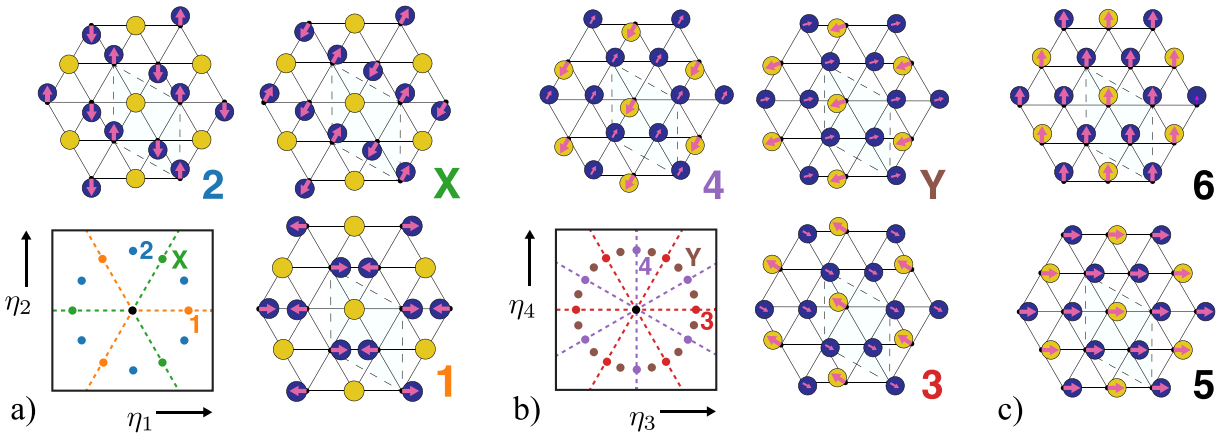


Fig. 6. Symmetry-adapted displacement modes in an ordered A–B 2d hexagonal crystal, with unit cell shown in light blue. The irreducible subspace (η_1, η_2), which is spanned by the collective modes numbered 1 and 2, is shown in (a), and the irreducible subspace (η_3, η_4), which is spanned by the collective modes numbered 3 and 4, is shown in (b). Dotted lines in the irreducible subspaces indicate high-symmetry directions in displacement space. The collective mode X is a linear combinations of modes 1 and 2 which lies on a high-symmetry direction. The collective mode Y is a linear combination of modes 3 and 4 which does not lie on a high-symmetry direction. The two homogeneous displacement modes, numbered 5 and 6, which do not affect the energy and are excluded from enumeration of derivative structures are shown in (c).

third set (Fig. 6(c)) simply describes rigid translations, which do not affect the energy of the crystal.

The appeal of the symmetry adapted collective displacement modes is that they enable a clear depiction of high symmetry collective displacements as well as all symmetrically equivalent displacement modes. For example, displacement mode 1, with amplitude η_1, and displacement mode 2, with amplitude η_2, can be linearly combined to generate new displacement modes that also reside in the same subspace. Displacement mode X in Fig. 6(a) is one such example and resides on a high symmetry subspace. Any linear combination of mode 1 and mode 2 generates a displacement pattern that has several symmetrically equivalent displacement patterns. Since modes 1 and 2 span an irreducible subspace, these symmetrically equivalent displacement patterns will also reside in the same subspace spanned by modes 1 and 2. Collective displacement patterns can therefore be enumerated systematically by gridding the symmetrically distinct wedges within the irreducible subspaces spanned by symmetry adapted collective displacement modes.

The ability to construct and grid up irreducible subspaces of collective displacement modes is invaluable in the analysis of dynamical instabilities of a crystal. Many high temperature phases become dynamically unstable upon cooling with respect to a symmetry breaking shuffle that resides in an irreducible subspace [9–11,63–66]. Notable examples are the cubic fluorite crystal structures of ZrO_2 and HfO_2 that exhibit low temperature instabilities with respect to a collective oxygen shuffle [67–69]. The collective oxygen shuffle resides in a three-dimensional irreducible subspace of oxygen and metal displacements within the cubic unit cell of the fluorite crystal structure. These are shown in Fig. 7(a). Fig. 7(b) shows the energy surface for HfO_2 as calculated over a grid spanning a two-dimensional slice of the three dimensional irreducible subspace. The energy surface exhibits a local maximum at the origin, corresponding to the cubic crystal, surrounded by multiple local minima, corresponding to oxygen shuffles that generate the tetragonal form of HfO_2. CASM is capable of identifying an irreducible wedge of symmetrically distinct displacements, thereby minimizing the number of static first-principles electronic structure calculations that need to be performed to construct an energy surface as depicted in Fig. 7(b).

Similar approaches can be applied to magnetic degrees of freedom [70], or strain degrees of freedom [46]. Fig. 8 illustrates symmetry-adapted strain modes for a two-dimensional triangular lattice. The three Cartesian strains, E_{xx}, E_{yy} and E_{xy}, in two-dimensions can be combined to generate symmetry adapted coordinates according to $e_1 = 1/\sqrt{2}(E_{xx} + E_{yy})$, $e_2 = 1/\sqrt{2}(E_{xx} - E_{yy})$ and $e_3 = \sqrt{2}E_{xy}$. The first strain order parameter, e_1, describes symmetry preserving deformations, while the two other order parameters, e_2 and e_3, span a two-dimensional irreducible subspace.

Local perturbations of continuous degrees of freedom within large supercells can also be enumerated with CASM. As with enumeration of occupation degrees of freedom, sites within a supercell can be directly specified or enumerated as symmetrically distinct point, pair, triplet, etc. clusters of sites in any previously generated structure. Degrees of freedom associated with the selected sites form a subspace in which local perturbations can be enumerated on a grid in the same manner as collective modes can be enumerated within an entire supercell. The capability to systematically enumerate displacements of multiple sites belonging to different clusters provide the ability to generate training data to fit lattice-dynamical Hamiltonians with anharmonic interactions.

2.5.3. Stochastic enumeration

In some cases, it is useful to generate derivative structures stochastically rather than through exhaustive enumeration. CASM can generate configurations with random occupation or local continuous degrees of freedom values, but typically more useful are configurations generated with Monte Carlo simulated annealing techniques. The Monte Carlo code implemented in CASM includes an option to periodically save sampled configurations when particular user-specified criteria are satisfied. A typical use case is to identify new configurations that a preliminary cluster expansion predicts as having low energies. The energies of the new configurations can then be calculated with a first-principles electronic structure method and added to the training dataset to re-fit the cluster expansion in an iterative manner (see Section 2.7).

CASM can also be used to stochastically generate so-called special quasi-random structures (SQS) [71], used to predict properties of disordered alloys. An SQS is a supercell configuration chosen such that, for a given number of sites, its correlations optimally match those of the disordered alloy with the same composition in the large size limit. For this reason, calculated properties tend to converge to those of the true disordered alloy more quickly when using an SQS. The SQS generating method in CASM, which is similar to the approach of van de Walle et al. [72], is a simulated annealing method that searches for configurations with correlations as close as possible to user specified target correlations.

2.5.4. Interfacing with first-principles electronic structure codes

Derivative structures enumerated by CASM may be output using a CASM crystal structure format [73] that can include any continuous

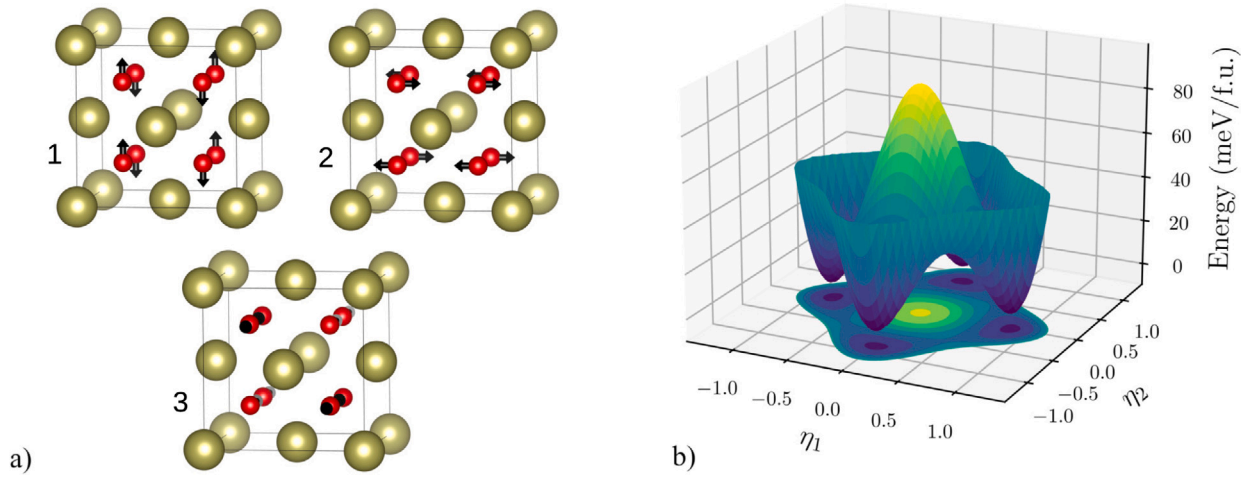


Fig. 7. (a) Symmetry-adapted displacement modes of cubic fluorite that convert the crystal to tetragonal symmetry. (b) Energy of HfO_2 as calculated over a grid within a two-dimensional slice of the three dimensional irreducible subspace spanned by the collective displacements of (a). The variables η_1 and η_2 refer to the amplitudes of the symmetry-adapted collective displacement modes 1 and 2.

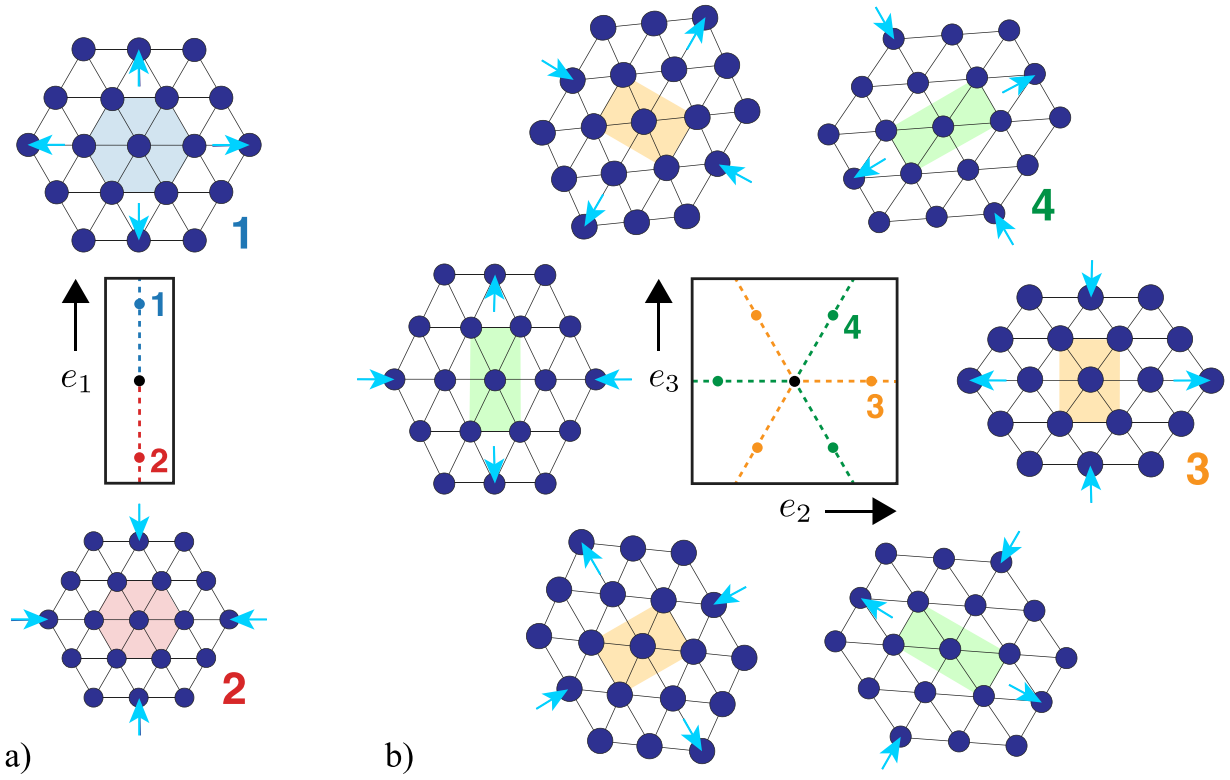


Fig. 8. Symmetry-adapted strain modes in a 2d hexagonal crystal with no occupational ordering. There are two irreducible subspaces, (e_1), shown in (a), corresponding to symmetry-preserving volumetric strain, and (e_2, e_3), shown in (b), corresponding to shape-changing strain. Dotted lines in the irreducible subspaces indicate high-symmetry directions in strain space.

properties that CASM knows how to transform under application of symmetry operations. CASM includes a Python package to construct VASP input files and to parse VASP output files. VASP output files are parsed to construct a representation of the calculation results in the CASM crystal structure format, including the calculated energy, lattice vectors, atomic coordinates, etc. Calculations using other first-principles and empirical atomistic codes can be used with CASM in a similar fashion, by converting between the CASM structure format. CASM supports conversion to many crystal structure formats through the use of the Atomic Simulation Environment (ASE) Python package [74].

When a calculated structure is imported into CASM to be represented as a configuration of degree of freedom values, a mapping algorithm [75] is used to assign atoms to crystal sites and to determine strains, displacements, and any other values such as magnetic spin. Mapping of calculated structures is crucial, even if the input structure was generated from a known configuration, because it is not uncommon for large relaxations to occur, effectively moving occupants among crystal sites. It is also not uncommon for some orderings to relax to entirely

different parent structures, such as an hcp-derived ordering relaxing to a structure that maps most closely to an fcc-derived structure [76–79].

The CASM mapping method systematically considers and scores mappings of the structure's lattice to the parent crystal structure lattice, and then, in the context of a particular lattice mapping, the assignment of atoms to crystal sites. Symmetry information is used to reduce the number of mappings that must be considered, and to construct mapping scores that depend only on symmetry-breaking strains or displacements. A weighted combination of the lattice and displacement mapping scores is used to score total structure mappings and find the configuration that a structure is most similar to or to reject the import of a structure altogether [75].

2.6. Managing configurations and calculated properties

It is often useful to work with a particular subset of enumerated configurations, for instance, when selecting which first-principles calculations to perform, analyzing results, or selecting configurations to use when fitting expansion coefficients. To make this easier, CASM includes features for querying properties of configurations and using those properties to filter and save selections of configurations.

Properties that can be queried include (i) direct functions of the configuration such as supercell size, composition, point group, and multiplicity of symmetrically equivalent configurations, and correlations, and (ii) first-principles calculated properties and derived quantities such as total energy, relaxation volume, mapping scores for the lattice and atomic deformation, and distance from the composition-energy convex hull. It is also possible to write plugins that evaluate custom properties of configurations.

2.7. Fitting expansion coefficients

The coefficients V^n_α of a truncated cluster expansion, Eq. (4), can be trained to a dataset of energies for configurations $\mathbb{C} = (\vec{\mathbf{s}}, \vec{\mathbf{d}}, \vec{\mathbf{m}}, \vec{E})$ that have been calculated with a first-principles electronic structure method. Eq. (7), which relates the energy per primitive cell to the correlation functions serves as the key to training the expansion coefficients V^n_α. By calculating the energies of M configurations, $\mathbb{C}_1, \ldots, \mathbb{C}_I, \ldots, \mathbb{C}_M$, with a first principles method, Eq. (7) can be used to set up the following linear set of equations:

$$\begin{pmatrix} e(\mathbb{C}_1) \\ \cdot \\ \cdot \\ \cdot \\ e(\mathbb{C}_I) \\ \cdot \\ \cdot \\ \cdot \\ e(\mathbb{C}_M) \end{pmatrix} = \begin{pmatrix} \Gamma^1_\alpha(\mathbb{C}_1) & \ldots & \Gamma^n_\gamma(\mathbb{C}_1) & \ldots \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \Gamma^1_\alpha(\mathbb{C}_I) & \ldots & \Gamma^n_\gamma(\mathbb{C}_I) & \ldots \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \Gamma^1_\alpha(\mathbb{C}_M) & \ldots & \Gamma^n_\gamma(\mathbb{C}_M) & \ldots \end{pmatrix} \begin{pmatrix} m^1_\alpha V^1_\alpha \\ \cdot \\ \cdot \\ \cdot \\ m^n_\gamma V^n_\gamma \\ \cdot \\ \cdot \\ \cdot \end{pmatrix}, \tag{13}$$

where the $e(\mathbb{C}_I)$ are the energies per unit cell of configurations $\mathbb{C}_I$, the $\Gamma^n_\gamma(\mathbb{C}_I)$ are the correlation functions (Eq. (6)) evaluated in configuration $\mathbb{C}_I$ and the V^n_γ are the unknown expansion coefficients.

A variety of techniques exist to estimate the unknown expansion coefficients V^n_γ appearing in Eq. (13). In general it is desirable to parameterize a sparse cluster expansion model to minimize the computational cost of Monte Carlo simulations where large numbers of energy evaluations are made. Hence, many approaches of inverting Eq. (13) seek to identify a minimal set of non-zero expansion coefficients V^n_γ. Some of these approaches employ algorithms or heuristics for selecting only a subset of columns in Eq. (13) for training, and force the coefficients corresponding to the other columns to be zero. The algorithms developed to scan through different combinations of non-zero expansion coefficients select the optimal set by minimizing a cross-validation score (either leave one out or a k-fold cross-validation scores) [19]. CASM allows users to additionally penalize sets based on the number of total non-zero coefficients. One useful algorithm, recursive feature elimination, fits all coefficients, and then recursively sets the least significant coefficients to zero until a user-specified number remain [80]. Another approach relies on genetic algorithms to identify optimal sets of non-zero expansion coefficients [20,81,82].

Common regression methods are motivated by Bayesian inference approaches, such as LASSO [83,84], ridge regression [85] and the relevance vector machine (RVM) [86,87]. For particular assumptions about priors and hyper-priors, these approaches generate best estimates of the cluster expansion coefficients and their distributions for a given dataset. The LASSO method, which can be interpreted as a Bayesian approach that assumes a Laplacian prior distribution on the expansion coefficients, achieves sparsity by minimizing the L_1 norm of the expansion coefficient vector [83,84]. Ridge-regression techniques assume a Gaussian prior distribution on the expansion coefficients and is often used to rank the importance of different clusters that appear in a cluster expansion [85]. Bayesian approaches can also provide a way to build up cluster expansions for chemical degrees of freedom in high dimensional composition spaces by relying on optimized cluster expansions of the lower dimensional composition subspaces [88].

There are also Bayesian approaches of generating an ensemble of cluster expansion models for the purpose of quantifying uncertainty and to propagate errors in calculated thermodynamic and kinetic properties. Once the probability distributions for the expansion coefficients V^n_γ are known, a large number of cluster expansion Hamiltonians can be sampled, e.g. with Markov chain Monte Carlo simulations [89]. The sampled cluster expansion Hamiltonians can then be used to calculate properties such as free energies and atomic mobilities using Metropolis Monte Carlo and kinetic Monte Carlo, respectively, and thereby estimate errors on free energies and atomic mobilities [87,89].

The CASM software package sets up the regression problem by managing first-principles calculations of the energies of enumerated configurations and by assembling the matrix of correlations appearing in Eq. (13). CASM includes an interface with the widely-used Python package SCIKIT-LEARN [80], which offers users with the latest regression techniques, and correlations and calculated energies may be easily output for use in other fitting tools. Python scripts will also be made available to generate ensembles of cluster expansion models using Bayesian inference techniques.

The parameterization of a cluster expansion often proceeds in an iterative manner. For example, an important property of a cluster expansion that describes the energy as a function of chemical degrees of freedom is that it predicts the correct ground states. Accurate ground state prediction is often crucial to ensure qualitatively correct predictions of finite temperature thermodynamic properties such as phase diagrams. CASM facilitates an iterative workflow, whereby a cluster expansion fit to an initial set of training data is used to identify new configurations that have low energies to be included in the next round of fitting. Also of tremendous importance when parameterizing a cluster expansion for chemical degrees of freedom is that the energies used in the training set correspond to relaxed structures that still map onto the original parent crystal structure. Training configurations that are dynamically unstable on the parent crystal structure and that relax to another parent crystal structure can severely contaminate the fit of a cluster expansion of chemical degrees of freedom. These structures need to be eliminated from the training dataset. The CASM crystal mapping tools [75] are highly effective at identifying these structures.

2.8. Monte Carlo calculations

The link between a cluster expansion description of a crystal and its finite temperature thermodynamic and kinetic properties is made through the application of statistical mechanics methods, primarily Monte Carlo calculations. The thermodynamic properties of a system in equilibrium are determined by derivatives of the free energy, Φ,

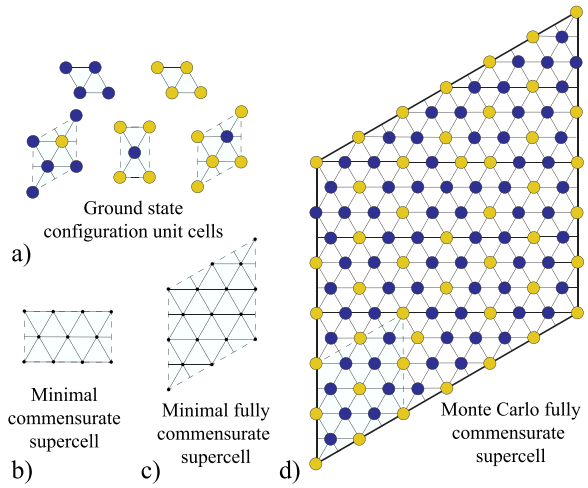


Fig. 9. Construction of a Monte Carlo cell that is fully commensurate with the ground state orderings. All ground state orderings, shown in (a), have at least one orientation which fills the minimal commensurate supercell, shown in (b). All orientations of the ground state orderings fill the minimal fully commensurate supercell, shown in (c). For best convergence, Monte Carlo calculations are typically performed in supercells of the minimal fully commensurate supercell, such as the example shown in (d).

with respect to the thermodynamic boundary conditions such as temperature, T, chemical potential, μ, pressure, P, etc. When a system is out of equilibrium, kinetic transport coefficients describe the time-dependent evolution of the state of the crystal. This section describes Monte Carlo calculations that rely on cluster expansions constructed and fit by CASM to calculate Φ and transport coefficients as a function of thermodynamic boundary conditions.

2.8.1. Free energy calculation

It is a fundamental result of statistical mechanics that the link between the spectrum of energies of an ensemble of microstates, s, to the free energy is through the partition function, Z, according to [90]

$$Z = \sum_s e^{-\beta \Omega_s} \tag{14}$$

$$\beta \Phi = -\ln Z. \tag{15}$$

Here, $\beta = 1/k_b T$ is the reciprocal temperature, and k_b is Boltzmann's constant. Ω_s is the potential energy of microstate s, which includes the energy typically calculated with a cluster expansion, E_s, and additional terms that account for the imposed thermodynamic boundary conditions. For example, under conditions of temperature and volume control, $\Omega_s = E_s$, while under conditions of temperature and pressure control, $\Omega_s = E_s + PV_s$.

For crystalline solids it is useful to consider the semi-grand canonical ensemble, in which temperature, pressure and chemical potential are controlled and the system has a fixed number of unit cells N_u. Under the constraint of a fixed number of unit cells, not all compositions and chemical potentials are independent. For an M-component alloy in which each chemical occupant is allowed on every crystal site, there are $k = M - 1$ composition degrees of freedom and $\Omega_s = E_s + PV_s - \sum_i^{M-1}(\mu_i - \mu_M)N_i$ [25,91,92].

In the general case, some occupants may be restricted to a subset of sublattices, and there will be $k <= M - 1$ independent composition axes. Then Ω_s can be written in terms of k parametric compositions, $\vec{x} = (x_1, x_2, \ldots, x_k)$, defined by a choice of composition axes that span the allowed composition space and normalized per unit cell, and exchange chemical potentials, $\vec{\tilde{\mu}}$, defined to be conjugate to $\vec{x}$, according to [93]:

$$\Omega_s = E_s + PV_s - N_u \sum_i^k \tilde{\mu}_i x_i. \tag{16}$$

Table 3
A summary of semi-grand canonical and canonical Monte Carlo capabilities implemented in CASM.

Equilibrium Monte Carlo capabilities in CASM
Semi-grand canonical and canonical ensembles
Metropolis and N-fold way algorithms
Discrete occupants may be atomic species, discrete spin states, or discrete molecular orientations
Find fully commensurate supercells of all ground states
Control temperature, chemical potential, magnetic field, and strain
Apply quadratic bias potentials for composition, order parameter, and magnetic spin
Customize properties to be sampled
Automatic convergence of any component of sampled quantities

CASM handles the general case, automatically calculating standard choices for the composition axes. Users can choose from one of the standard choices or define their own custom composition axes.

Eqs. (14) and (15) cannot be used to directly calculate Φ, except in limited cases, because the number of microstates grows very quickly with N. Instead, a tractable approach is to use Monte Carlo methods to calculate ensemble average properties of the system, $\langle X \rangle = \frac{1}{Z}\sum_s X_s \exp(-\beta\Omega_s)$. Taking derivatives of Eqs. (14) and (15), it can be shown that $\langle X \rangle$ are derivatives of Φ with respect to the thermodynamic boundary conditions being controlled [90]. For example, in the semi-grand canonical ensemble, taking derivatives of the free energy, Φ_{SG}, with respect to T and $\tilde{\mu}_i$ yields:

$$\left(\frac{\partial \beta \Phi_{SG}}{\partial \beta}\right)_{P,\tilde{\mu}_i,N} = -\frac{1}{Z}\frac{\partial Z}{\partial \beta} = \langle \Omega \rangle, \tag{17}$$

$$\left(\frac{\partial \beta \Phi_{SG}}{\partial \tilde{\mu}_i}\right)_{\beta,P,\tilde{\mu}_{j\neq i},N} = -\frac{1}{Z}\frac{\partial Z}{\partial \tilde{\mu}_i} = -\beta N \langle x_i \rangle. \tag{18}$$

Then, Φ_{SG} can be calculated at other thermodynamic conditions by numerical integration starting from a reference state where it can be directly approximated, such as at very low temperatures [91,94]. Explicitly, the temperature dependence of Φ_{SG} is:

$$\beta'' \Phi(\beta'', \vec{\tilde{\mu}}) = \beta' \Phi(\beta', \vec{\tilde{\mu}}) + \int_{\beta'}^{\beta''} \langle \Omega \rangle d\beta, \tag{19}$$

where β' and β'' are the reciprocal temperatures at two states in thermodynamic state space, and all boundary conditions other than temperature are held constant. The chemical potential dependence of Φ_{SG} is:

$$\Phi(\beta, \vec{\tilde{\mu}}'') = \Phi(\beta, \vec{\tilde{\mu}}') - N \int_{\vec{\tilde{\mu}}'}^{\vec{\tilde{\mu}}''} \langle \vec{x} \rangle \cdot d\vec{\tilde{\mu}}, \tag{20}$$

where $\vec{\tilde{\mu}}'$ and $\vec{\tilde{\mu}}''$ are the chemical potentials conjugate to $\vec{x}$ for two states in thermodynamic state space, and the temperature and all other boundary conditions are held constant.

The above integral equations yield a general approach for calculating Φ from first-principles calculations regardless of degrees of freedom or boundary conditions: construct and fit a cluster expansion for E_s, determine the form of Ω_s, perform Monte Carlo calculations on a grid in thermodynamic space, and calculate Φ as a function of the relevant thermodynamic conditions by thermodynamic integration from a state in which Φ can be approximated.

2.8.2. Equilibrium Monte Carlo implementation

CASM includes Monte Carlo implementations (see Table 3) for the Metropolis algorithm and the N-fold way algorithm for systems with discrete occupation degrees of freedom in either the canonical or grand canonical ensemble. The implementation for the canonical ensemble method is optimized to efficiently track occupants, such that there is

no slow down attempting to find pairs that can swap at dilute concentrations. There is no limitation, other than computational expense, on the number of sublattices, or the number or type of occupants on each sublattice. Occupants may have anisotropic properties, such as magnetic spins, or represent discrete molecular orientations. The energy of the system may be parameterized by a cluster expansion of occupation degrees of freedom or a coupled cluster expansion of occupation and strain degrees of freedom under conditions of constant strain. Support for Hamiltonian Monte Carlo calculations [16,95], which efficiently sample over continuous degrees of freedom, is planned for a future release of CASM.

Linear and quadratic potentials can be applied to the composition, magnetic moment, and configurational order parameters. Quadratic potentials enable so-called "variance-constrained" or "umbrella sampling" Monte Carlo methods that maintain the system at chosen values of the composition or configurational order parameter in order to sample regions of phase space in which the free energy is concave. For a detailed description of these methods, see [22,96]. As described in Section 2.5, a correlation matching potential is also implemented in order to generate SQS, using the objective function proposed by Van de Walle et al. [72].

The Monte Carlo methods implemented in CASM offer significant control over the calculation, allowing users to specify the initial configuration, linear or custom paths through thermodynamic space, manual or automatic convergence of ensemble averages to desired precision levels, and the quantities to be sampled. There is also a capability to write extensions to sample custom quantities. Automatic convergence to a requested precision and estimation of the precision of the mean is calculated following the method of van de Walle [91]. Paths in thermodynamic space can be restarted if they stop before finishing.

Performing calculations in Monte Carlo supercells that are fully commensurate with all important configurations helps to ensure accurate sampling of the energetically favorable states that are preferred in the large size limit. Given a set of important ground states and metastable states, CASM can calculate the minimal volume supercell that is either (a) commensurate with one variant of each configuration in the set, or (b) fully commensurate with all symmetrically equivalent configurations in the set. This is shown schematically in Fig. 9.

The primary output of Monte Carlo calculations is a summary of the applied boundary conditions and the ensemble average value of sampled quantities $\langle \Omega/N_u \rangle$, $\langle N_i/N_u \rangle$, $\langle E/N_u \rangle$, $\langle \Gamma^n_\alpha \rangle$, etc., which are normalized by the total number of parent structure unit cells, N_u, in the Monte Carlo cell. CASM also calculates thermodynamic response functions, including the heat capacity and chemical susceptibilities, which are second derivatives of the free energy with respect to its natural variables. Along with configurational order parameters, and correlations, the thermodynamic response functions can be useful for identifying and characterizing phase transitions. As an additional option, individual observations and the configuration being sampled may be saved to disk to enable custom analysis as a post-processing step, or to visualize sampled microstates.

2.8.3. Calculation of kinetic coefficients

Diffusion in solids occurs through infrequent discrete events in which one or more chemical occupants change sites. The evolution of the system can be modeled using the master equation

$$\frac{\partial P_s(t)}{\partial t} = \sum_{s' \neq s} \left(R_{s's} P_{s'}(t) - R_{ss'} P_s(t) \right), \tag{21}$$

where $R_{ss'}$ is the rate at which the system transitions from state s to state s', and $P_s(t)$ is the probability of being in state s at time t. The rates, $R_{ss'}$, can be calculated when provided a catalog of possible events, a cluster expansion for E_s, and local-cluster expansions for $\Delta E^{KRA}_{ss'}$ and $\nu^*_{ss'}$ as described in Section 2.4. Then the kinetic Monte Carlo method [97] can be used to evolve the system over time and sample trajectories.

Table 4
A summary of kinetic Monte Carlo capabilities implemented in CASM.

Kinetic Monte Carlo capabilities in CASM
Substitutional and interstitial diffusion treated on equal footing
May include multi-site hops, discrete magnetic spin flips, or discrete molecular re-orientations
May include deposition- or dissolution-type events
Control temperature, composition, chemical potential, magnetic field, and strain

At the macroscopic scale, diffusional fluxes, $\vec{J}$, are driven by gradients in chemical potentials, $\nabla \mu_j$. For multicomponent isotropic diffusion, these can be related to first order through the Onsager kinetic coefficients, $\boldsymbol{L}$, according to

$$\vec{J} = -\boldsymbol{L} \nabla \vec{\mu}, \tag{22}$$

where element L_{ij} of the kinetic coefficient matrix, $\boldsymbol{L}$, gives the contribution of the gradient in chemical potential of species j, $\nabla \mu_j$, to the flux J_i of species i. The extension to anisotropic diffusion yields kinetic coefficients represented by a rank-four tensor, with element $L_{ij,ab}$ relating chemical potential gradients of species j along spatial direction a to the flux of species i along spatial direction b.

For comparison to experimental observations, it is useful to calculate the diffusion coefficients, $\boldsymbol{D}$, relating diffusional fluxes to gradients in composition per unit volume, $\vec{c}$. As with $\boldsymbol{L}$, for multicomponent, isotropic diffusion, $\vec{J} = -\boldsymbol{D} \nabla \vec{c}$, where $\boldsymbol{D}$ is a matrix, while in the general anisotropic case $\boldsymbol{D}$ is a rank-four tensor. Thermodynamic calculations, as described in the previous section, give the relationship between $\vec{\mu}$ and $\vec{c}$ and can be used to determine the thermodynamic factor, $\boldsymbol{\Theta}$, which relates $\boldsymbol{L}$ to $\boldsymbol{D}$ according to $\boldsymbol{D} = \boldsymbol{L\Theta}$ [25,27].

The link between diffusional hops at the atomic scale and the macroscopic transport behavior can be made using linear response methods that relate fluctuations in concentration that occur at equilibrium to transport coefficients that describe non-equilibrium transport. Kinetic Monte Carlo trajectories in relatively small systems at equilibrium can be used to calculate the kinetic coefficients according to [98,99]

$$L_{ij,ab} = \frac{\left\langle \left(\sum_\zeta \Delta \vec{R}^\zeta_{i,a} \right) \left(\sum_\zeta \Delta \vec{R}^\zeta_{j,b} \right) \right\rangle}{2tV k_b T}, \tag{23}$$

where $\Delta \vec{R}^\zeta_{i,a}$ is the a-Cartesian component of the vector connecting the beginning and ending points of the ζ-th atom of species i, V is the system volume, k_b is Boltzmann's constant, t is the observation time, and T is the temperature. Note that $\boldsymbol{L}$ is a function of the thermodynamic boundary conditions. It is typically calculated as a function of temperature and composition, but there may also be significant effects due to stress or electromagnetic fields.

2.8.4. Kinetic Monte Carlo implementation

CASM includes a kinetic Monte Carlo implementation (see Table 4) that uses local-cluster expansions for $\Delta E^{KRA}_{ss'}$ and $\nu^*_{ss'}$ to calculate event rates, $R_{ss'}$. The neighborhood of the local environments is used to determine which event rates must be updated at each step in the algorithm.

CASM treats substitutional [25,27,100], interstitial [23,24,26,28], and mixed diffusion on an equal footing, with interstitial sites represented as a sublattice in the parent crystal structure. Diffusion events can involve the exchange of occupants amongst any number of sites [61,62], limited only by computational expense. Events may consist of occupant changes on a single site, to treat rotations of discrete molecular orientation and magnetic spin flips. Events that do not preserve composition are also possible for simulations of processes such as deposition or dissolution.

End state energies are calculated using a cluster expansion in the same manner as the equilibrium Monte Carlo implementation. This means there is no limitation, other than computational expense, on

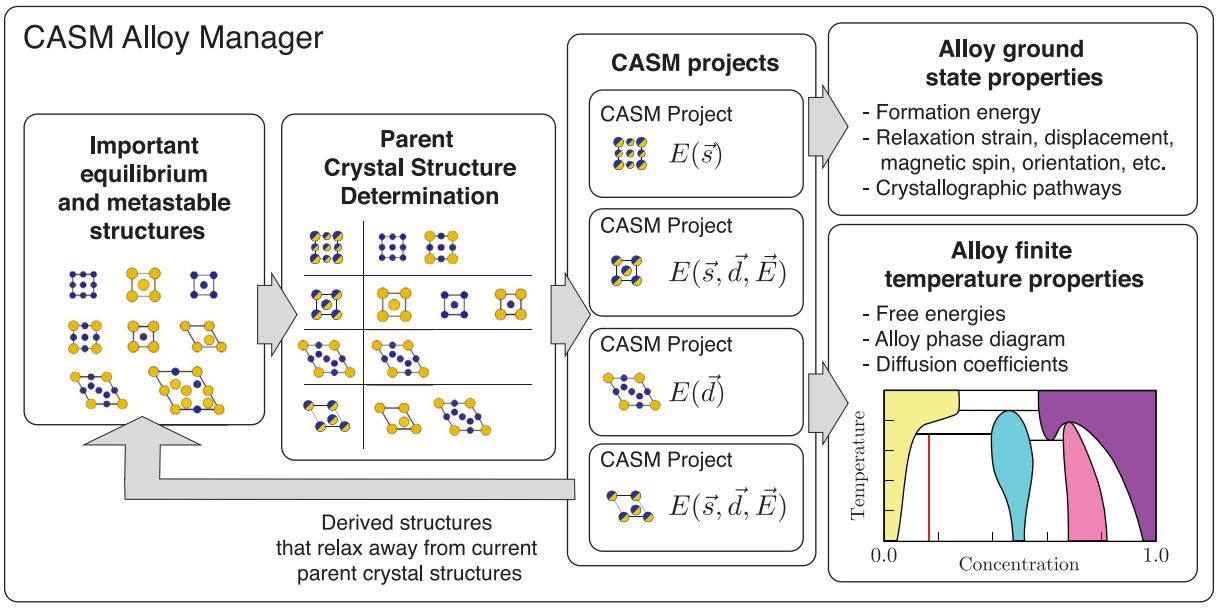


Fig. 10. Schematic showing the CASM Alloy Manager workflow.

the number of sublattices, or the number or type of occupants on each sublattice. Occupants may have anisotropic properties, such as magnetic spins, or represent discrete molecular orientations. The end state energies of the system may be parameterized by a cluster expansion of occupation degrees of freedom or a coupled cluster expansion of occupation and strain degrees of freedom under conditions of constant strain.

As with the equilibrium Monte Carlo implementation, users are able to specify the initial configuration, linear or custom paths through state space, convergence criteria, and the quantities that are to be sampled. For diffusion calculations, atomic coordinates that take into account periodic boundary crossings are typically sampled at linearly spaced time intervals. For other types of simulations, it is possible to output the state of the configuration using logarithmically spaced intervals, which can be useful for observing processes that happen over multiple length-scales while maintaining reasonable data storage requirements. CASM includes tools to analyze and automate convergence, reducing overall computational expense.

3. CASM alloy manager

The previous section, describing an individual CASM project, focuses on the capabilities of CASM with respect to a single parent crystal structure. Each parent crystal structure can form a solid solution at finite temperatures along with a variety of chemical and/or shuffle ordered compounds that have a group/subgroup relationship with the parent crystal structure. Given a parent crystal structure CASM enables the prediction of:

- ground state chemical, magnetic and/or shuffle orderings;
- finite temperature free energies of solid solutions and ordered compounds;
- kinetic transport coefficients.

The phase diagrams of many, if not most, multicomponent systems, however, show numerous stable phases that are derived from more than one parent crystal structure [35]. This occurs in many metallic alloy systems and within the multicomponent composition spaces of most functional materials. The Ni–Al alloy system, for example, favors fcc at Al- or Ni-rich compositions, but forms the B2 intermetallic compound, an ordering on bcc, at equiatomic compositions [101]. Similarly, Ti-rich compositions in the Ti–Al binary system favor an hcp solid solution and the hcp based DO_{19} ordering, while fcc derived ordered compounds, including $L1_0$ and $L1_2$, are favored at the more Al-rich compositions [102]. Phase diagrams involving oxides, nitrides and carbides also show a rich variety of stable phases that can be viewed as particular chemical orderings or symmetry breaking shuffles of more than one parent crystal structure [20,103,104].

Global phase stability is determined by the common tangent construction applied to the free energies of the solid solutions and derivative chemical and/or shuffle ordered compounds of competing parent crystal structures. A first step to establish global phase stability is, therefore, to identify likely parent crystal structures [35,75]. Once an exhaustive list of parent crystal structures has been established, their thermodynamic properties can be calculated using CASM. The task of CASM Alloy Manager (Fig. 10) is to (i) identify a minimal set of parent crystal structures for a particular chemical composition space, (ii) manage multiple CASM projects, one for each parent crystal structure, and (iii) integrate the thermodynamic predictions from each CASM project to calculate global phase stability.

The approach to identifying a minimal set of parent crystal structures within a given composition space starts with all known solid solution and ordered phases from experimental and first-principles databases [35]. The CASM mapping algorithm [75] is then applied to identify the smallest set of high symmetry parent crystal structures from which all known compounds of the system can be derived through a chemical ordering or through a group/subgroup structural distortion [35]. To expedite the search, the algorithm relies on a growing database of "usual suspect" parent crystal structures ranked by their frequency of occurrence [35]. Once a minimal set of parent crystal structures has been identified, CASM Alloy Manager spawns off CASM projects for each parent crystal structure and subsequently assembles calculated thermodynamic properties, including zero-Kelvin ground states and finite temperature free energies, to construct global phase diagrams. It is common that several enumerated structures of one parent crystal structure are dynamically unstable and relax to structures that map onto other parent crystal structure [76,79]. If such a structure is found to map onto the parent crystal structure of an existing CASM project, CASM Alloy Manager will import the calculation into that project. Otherwise, it will initialize a new CASM project including the calculated structure.

4. Interfacing with large length-scale methods and materials databases

The ground state and finite temperature properties generated by CASM can be used to parameterize higher length-scale methods. The phase field method [105–107] and other approaches [108,109] for modeling microstructural evolution depend on a number of parameters that CASM can calculate directly, including the homogeneous free energy for each phase being modeled, the stress-free relaxation strain, elastic constants, and kinetic coefficients. In some cases these properties can be directly parameterized by CASM, while in other cases it is necessary to sample the results generated by CASM and re-fit for the approximation level of the method being used [110–113].

CASM projects have a well-defined directory structure, which makes them easy to share, integrate with other codes, re-use, and extend to additional chemical species or types of degrees of freedom. To make results available to collaborators and the general community, CASM projects can be easily uploaded to materials repositories, such as Materials Commons [36]. With Materials Commons it is possible to upload a CASM project which can then be shared privately with collaborators until results are ready for publication. Materials Commons is capable of publishing large datasets, including all the first-principles and Monte Carlo data that might exist in a CASM project. It is also possible to publish a subset of a CASM project, containing just the parent crystal structure, cluster expansion, and fitted coefficients, and summarized data for fast download and application.

CASM Alloy Manager makes it possible to jump-start CASM projects by interfacing with external databases to import known structures. Structures created using PYMATGEN [114], downloaded from Materials Project [115], or converted from other crystal structure formats can be mapped to parent crystal structures in the alloy system. For any valid mapping, the structure and calculated properties can be imported into the corresponding CASM project. The lattice vectors and atomic coordinates of the imported structures can then be used as a starting point for first-principles calculations that re-relax the structure to ensure consistency of calculated properties. Future support for import from additional databases is planned via the OPTIMADE API [116,117].

5. Obtaining CASM

Core CASM components, for defining crystal structures, analyzing symmetry, enumerating configurations, and executing Monte Carlo calculations are implemented in a C++ library. Higher level usage of the core components, and methods for interfacing with other software programs, are implemented in a Python package, CASM-PYTHON. This includes interfacing with machine-learning packages for fitting coefficients, in particular SCIKIT-LEARN, managing first-principles calculations, and converting amongst different structure formats. The most commonly used methods for constructing, fitting, and applying cluster expansion Hamiltonians are available both via a command line program and CASM-PYTHON.

Open source versions of the CASM C++ library, Python package, and command line program, containing many of the features described in this paper, are available for download from GitHub. Open source release of the CASM kinetic Monte Carlo implementation, local environment enumeration, N-fold way Monte Carlo algorithm, and CASM Alloy Manager is planned for spring 2023. The CASM C++ library and command line program can be installed from source, or precompiled versions can be downloaded as a Conda package or Docker container. CASM-PYTHON is available for download from PyPI. Methods implemented in CASM can be used to power other programs that focus on particular applications, such as constructing structures with surfaces and interfaces [118] or fitting neural network Hamiltonians [119].

Detailed usage instructions for CASM, including tutorial applications are available online [73].

6. Conclusion

The Clusters Approach to Statistical Mechanics (CASM) software package enables first-principles statistical mechanics calculations of thermodynamic and kinetic properties of multi-component crystals. It is designed to algorithmically construct generalized cluster expansion Hamiltonians to extrapolate first-principles electronic structure calculations within (kinetic) Monte Carlo simulations. CASM is capable of treating multiple local site degrees of freedom within a crystal, including chemical, displacement, magnetic and other localized electronic excitations along with macroscopic strain. The CASM package also contains a variety of group theoretical tools with which to identify symmetry adapted order parameters and to systematically enumerate symmetry breaking perturbations to high symmetry reference crystals. The code base offers unique crystal mapping capabilities to quantify the similarity between different crystals and to identify crystallographic pathways between pairs of crystal structures. An alloy manager tool manages multiple CASM projects, by identifying distinct parent crystal structures within a multi-component composition space and by aggregating free energies to determine global phase stability. CASM is an open source software package consisting of C++ libraries and Python scripts.

CRediT authorship contribution statement

Brian Puchala: Conceptualization, Methodology, Software, Co-Lead, Visualization, Writing – original draft. **John C. Thomas:** Conceptualization, Methodology, Software, Co-Lead, Writing – review & editing. **Anirudh Raju Natarajan:** Conceptualization, Methodology, Software, Writing – review & editing. **Jon Gabriel Goiri:** Conceptualization, Methodology, Software, Writing – review & editing. **Sesha Sai Behara:** Software, Visualization. **Jonas L. Kaufman:** Software. **Anton Van der Ven:** Supervision, Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the U.S. Department of Energy, Office of Basic Energy Sciences, Division of Materials Sciences and Engineering, United States of America under Award #DE-SC0008637 as part of the Center for Predictive Integrated Structural Materials Science (PRISMS Center) at University of Michigan. We are also grateful for computing resources from the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231 using NERSC award BES-ERCAP0020609.

References

[1] K. Lejaeghere, G. Bihlmayer, T. Björkman, P. Blaha, S. Blügel, V. Blum, D. Caliste, I.E. Castelli, S.J. Clark, A. Dal Corso, et al., Reproducibility in density functional theory calculations of solids, Science 351 (6280) (2016) aad3000.

[2] D. De Fontaine, Cluster approach to order-disorder transformations in alloys, in: Solid State Physics, Vol. 47, Elsevier, 1994, pp. 33–176.

[3] A. Van der Ven, J.C. Thomas, B. Puchala, A.R. Natarajan, First-principles statistical mechanics of multicomponent crystals, Annu. Rev. Mater. Res. 48 (2018) 27–55.

[4] G. Ceder, A derivation of the Ising model for the computation of phase diagrams, Comput. Mater. Sci. 1 (2) (1993) 144–150.

[5] A. Van De Walle, G. Ceder, The effect of lattice vibrations on substitutional alloy thermodynamics, Rev. Modern Phys. 74 (1) (2002) 11.

[6] B. Fultz, Vibrational thermodynamics of materials, Prog. Mater. Sci. 55 (4) (2010) 247–352.

[7] W. Zhong, D. Vanderbilt, K.M. Rabe, First-principles theory of ferroelectric phase transitions for perovskites: The case of BaTiO 3, Phys. Rev. B 52 (9) (1995) 6301.

[8] J. Bhattacharya, A. Van der Ven, Mechanical instabilities and structural phase transitions: The cubic to tetragonal transformation, Acta Mater. 56 (16) (2008) 4226–4232.

[9] P. Souvatzis, O. Eriksson, M. Katsnelson, S. Rudin, Entropy driven stabilization of energetically unstable crystal structures explained from first principles theory, Phys. Rev. Lett. 100 (9) (2008) 095901.

[10] J.C. Thomas, A. Van der Ven, Finite-temperature properties of strongly anharmonic and mechanically unstable crystal phases from first principles, Phys. Rev. B 88 (21) (2013) 214111.

[11] J.S. Bechtel, J.C. Thomas, A. Van der Ven, Finite-temperature simulation of anharmonicity and octahedral tilting transitions in halide perovskites, Phys. Rev. Mater. 3 (11) (2019) 113605.

[12] J.C. Thomas, A. Van der Ven, Elastic properties and stress-temperature phase diagrams of high-temperature phases with low-temperature lattice instabilities, Phys. Rev. B 90 (2014) 224105, http://dx.doi.org/10.1103/PhysRevB.90.224105.

[13] F. Zhou, T. Maxisch, G. Ceder, Configurational electronic entropy and the phase diagram of mixed-valence oxides: The case of LixFePO4, Phys. Rev. Lett. 97 (15) (2006) 155704.

[14] M.D. Radin, A. Van der Ven, Simulating charge, spin, and orbital ordering: Application to Jahn–Teller distortions in layered transition-metal oxides, Chem. Mater. 30 (3) (2018) 607–618.

[15] E. Decolvenaere, E. Levin, R. Seshadri, A. Van der Ven, Modeling magnetic evolution and exchange hardening in disordered magnets: The example of Mn1–xFexRu2Sn Heusler alloys, Phys. Rev. Mater. 3 (10) (2019) 104411.

[16] D.A. Kitchaev, E.C. Schueller, A. Van der Ven, Mapping skyrmion stability in uniaxial lacunar spinel magnets from first principles, Phys. Rev. B 101 (5) (2020) 054409.

[17] J.S. Bechtel, R. Seshadri, A. Van der Ven, Energy landscape of molecular motion in cubic methylammonium lead iodide from first-principles, J. Phys. Chem. C 120 (23) (2016) 12403–12410.

[18] J.C. Thomas, J.S. Bechtel, A. Van der Ven, Hamiltonians and order parameters for crystals of orientable molecules, Phys. Rev. B 98 (9) (2018) 094105.

[19] A. van de Walle, G. Ceder, Automating first-principles phase diagram calculations, J. Phase Equilib. 23 (4) (2002) 348–359.

[20] B. Puchala, A. Van der Ven, Thermodynamics of the Zr-O system from first-principles calculations, Phys. Rev. B 88 (2013) 094108, http://dx.doi.org/10.1103/PhysRevB.88.094108.

[21] J. Vinckeviciute, M.D. Radin, A. Van der Ven, Stacking-sequence changes and Na ordering in layered intercalation materials, Chem. Mater. 28 (23) (2016) 8640–8650.

[22] A.R. Natarajan, J.C. Thomas, B. Puchala, A. Van der Ven, Symmetry-adapted order parameters and free energies for solids undergoing order-disorder phase transitions, Phys. Rev. B 96 (13) (2017) 134204.

[23] A. Van der Ven, G. Ceder, M. Asta, P.D. Tepesch, First-principles theory of ionic diffusion with nondilute carriers, Phys. Rev. B 64 (18) (2001) 184307.

[24] A. Van der Ven, J.C. Thomas, Q. Xu, B. Swoboda, D. Morgan, Nondilute diffusion from first principles: Li diffusion in LixTiS2, Phys. Rev. B 78 (10) (2008) 104306.

[25] A. Van der Ven, H.-C. Yu, G. Ceder, K. Thornton, Vacancy mediated substitutional diffusion in binary crystalline solids, Prog. Mater. Sci. 55 (2) (2010) 61–105.

[26] J. Bhattacharya, A. Van der Ven, First-principles study of competing mechanisms of nondilute Li diffusion in spinel LixTiS2, Phys. Rev. B 83 (14) (2011) 144302.

[27] J.G. Goiri, S.K. Kolli, A. Van der Ven, Role of short-and long-range ordering on diffusion in Ni-Al alloys, Phys. Rev. Mater. 3 (9) (2019) 093402.

[28] S.K. Kolli, A. Van der Ven, Elucidating the factors that cause cation diffusion shutdown in spinel-based electrodes, Chem. Mater. 33 (16) (2021) 6421–6432.

[29] J.M. Sanchez, F. Ducastelle, D. Gratias, Generalized cluster description of multicomponent systems, Physica A 128 (1–2) (1984) 334–350.

[30] M. Asta, C. Wolverton, D. De Fontaine, H. Dreyssé, Effective cluster interactions from cluster-variation formalism. I, Phys. Rev. B 44 (10) (1991) 4907.

[31] G. Inden, Atomic ordering, Phase Transform. Mater. (2001) 519–581.

[32] J.M. Sanchez, Cluster expansion and the configurational theory of alloys, Phys. Rev. B 81 (22) (2010) 224202.

[33] R. Drautz, M. Fähnle, Spin-cluster expansion: Parametrization of the general adiabatic magnetic energy surface with ab initio accuracy, Phys. Rev. B 69 (10) (2004) 104404.

[34] J.C. Thomas, B. Puchala, A. Van der Ven, Generalized cluster expansion Hamiltonians, 2022 (in preparation).

[35] S.K. Kolli, A.R. Natarajan, J.C. Thomas, T.M. Pollock, A. Van der Ven, Discovering hierarchies among intermetallic crystal structures, Phys. Rev. Mater. 4 (11) (2020) 113604.

[36] B. Puchala, G. Tarcea, E.A. Marquis, M. Hedstrom, H.V. Jagadish, J.E. Allison, The materials commons: A collaboration platform and information repository for the global materials community, JOM 68 (8) (2016) 2035–2044.

[37] G. Kresse, J. Hafner, Ab initio molecular-dynamics simulation of the liquid-metal-amorphous-semiconductor transition in germanium, Phys. Rev. B 49 (1994) 14251.

[38] 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 (1996) 15.

[39] 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.

[40] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G.L. Chiarotti, M. Cococcioni, I. Dabo, A.D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A.P. Seitsonen, A. Smogunov, P. Umari, R.M. Wentzcovitch, QUANTUM ESPRESSO: A modular and open-source software project for quantum simulations of materials, J. Phys.: Condens. Matter 21 (39) (2009) 395502, http://dx.doi.org/10.1088/0953-8984/21/39/395502.

[41] P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M.B. Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni, N. Colonna, I. Carnimeo, A.D. Corso, S. de Gironcoli, P. Delugas, R.A. DiStasio, A. Ferretti, A. Floris, G. Fratesi, G. Fugallo, R. Gebauer, U. Gerstmann, F. Giustino, T. Gorni, J. Jia, M. Kawamura, H.-Y. Ko, A. Kokalj, E. Küçükbenli, M. Lazzeri, M. Marsili, N. Marzari, F. Mauri, N.L. Nguyen, H.-V. Nguyen, A.O. de-la Roza, L. Paulatto, S. Poncé, D. Rocca, R. Sabatini, B. Santra, M. Schlipf, A.P. Seitsonen, A. Smogunov, I. Timrov, T. Thonhauser, P. Umari, N. Vast, X. Wu, S. Baroni, Advanced capabilities for materials modelling with quantum ESPRESSO, J. Phys.: Condens. Matter 29 (46) (2017) 465901, http://dx.doi.org/10.1088/1361-648x/aa8f79.

[42] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, M. Scheffler, Ab initio molecular simulations with numeric atom-centered orbitals, Comput. Phys. Comm. 180 (11) (2009) 2175–2196, http://dx.doi.org/10.1016/j.cpc.2009.06.022.

[43] C.C. Stoumpos, M.G. Kanatzidis, The renaissance of halide perovskites and their evolution as emerging semiconductors, Acc. Chem. Res. 48 (10) (2015) 2791–2802.

[44] D.H. Fabini, J.G. Labram, A.J. Lehner, J.S. Bechtel, H.A. Evans, A. Van der Ven, F. Wudl, M.L. Chabinyc, R. Seshadri, Main-group halide semiconductors derived from perovskite: Distinguishing chemical, structural, and electronic aspects, Inorg. Chem. 56 (1) (2017) 11–25.

[45] Y. Tokura, N. Nagaosa, Orbital physics in transition-metal oxides, Science 288 (5465) (2000) 462–468.

[46] J.C. Thomas, A. Van der Ven, The exploration of nonlinear elasticity and its efficient parameterization for crystalline materials, J. Mech. Phys. Solids 107 (2017) 76–95.

[47] J.M. Sanchez, Cluster expansions and the configurational energy of alloys, Phys. Rev. B 48 (18) (1993) 14013.

[48] P.D. Tepesch, G.D. Garbulsky, G. Ceder, Model for configurational thermodynamics in ionic systems, Phys. Rev. Lett. 74 (12) (1995) 2272.

[49] B. Han, A. Van der Ven, G. Ceder, B.-J. Hwang, Surface segregation and ordering of alloy surfaces in the presence of adsorbates, Phys. Rev. B 72 (20) (2005) 205409.

[50] J.C. Thomas, J.M. Millunchick, N.A. Modine, A. Van der Ven, Surface atomic order of compound III-V semiconductor alloys at finite temperature, Phys. Rev. B 80 (12) (2009) 125315.

[51] N.S.H. Gunda, A. Van der Ven, Understanding the interactions between interstitial and substitutional solutes in refractory alloys: The case of Ti-Al-O, Acta Mater. 191 (2020) 149–157.

[52] D. Alfè, PHON: A program to calculate phonons using the small displacement method, Comput. Phys. Comm. 180 (12) (2009) 2622–2633.

[53] A. Togo, I. Tanaka, First principles phonon calculations in materials science, Scr. Mater. 108 (2015) 1–5.

[54] A. Van de Walle, A complete representation of structure–property relationships in crystals, Nature Mater. 7 (6) (2008) 455–458.

[55] D.B. Laks, L.G. Ferreira, S. Froyen, A. Zunger, Efficient cluster expansion for substitutional systems, Phys. Rev. B 46 (19) (1992) 12587.

[56] 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.

[57] A.R. Natarajan, A. Van der Ven, Linking electronic structure calculations to generalized stacking fault energies in multicomponent alloys, NPJ Comput. Mater. 6 (1) (2020) 1–9.

[58] G.H. Vineyard, Frequency factors and isotope effects in solid state rate processes, J. Phys. Chem. Solids 3 (1–2) (1957) 121–127.

[59] G.L.W. Hart, R.W. Forcade, Algorithm for generating derivative structures, Phys. Rev. B 77 (2008) 224115, http://dx.doi.org/10.1103/PhysRevB.77.224115.

[60] Y. Mishin, A.Y. Lozovoi, A. Alavi, Evaluation of diffusion mechanisms in NiAl by embedded-atom and first-principles calculations, Phys. Rev. B 67 (1) (2003) 014201.

[61] Q. Xu, A. Van der Ven, First-principles investigation of migration barriers and point defect complexes in B2–NiAl, Intermetallics 17 (5) (2009) 319–329.

[62] Q. Xu, A. Van der Ven, Atomic transport in ordered compounds mediated by local disorder: Diffusion in B2-Nix Al1–x, Phys. Rev. B 81 (6) (2010) 064303.

[63] V. Ozolins, First-principles calculations of free energies of unstable phases: The case of fcc W, Phys. Rev. Lett. 102 (6) (2009) 065702.

[64] G. Grimvall, B. Magyari-Köpe, V. Ozoliņš, K.A. Persson, Lattice instabilities in metallic elements, Rev. Modern Phys. 84 (2) (2012) 945.

[65] M.-H. Chen, A. Emly, A. Van der Ven, Anharmonicity and phase stability of antiperovskite Li3 OCl, Phys. Rev. B 91 (21) (2015) 214306.

[66] M.D. Radin, J.C. Thomas, A. Van der Ven, Order-disorder versus displacive transitions in jahn-teller active layered materials, Phys. Rev. Mater. 4 (4) (2020) 043601.

[67] C. Carbogno, C.G. Levi, C.G. Van de Walle, M. Scheffler, Ferroelastic switching of doped zirconia: Modeling and understanding from first principles, Phys. Rev. B 90 (14) (2014) 144109.

[68] M.-H. Chen, J.C. Thomas, A.R. Natarajan, A. Van der Ven, Effects of strain on the stability of tetragonal ZrO 2, Phys. Rev. B 94 (5) (2016) 054108.

[69] S.S. Behara, A. Van der Ven, Ferroelectric HfO 2 and the importance of strain, Phys. Rev. Mater. 6 (5) (2022) 054403.

[70] F. Walsh, A.R. Natarajan, A. Van der Ven, Order parameters for antiferromagnetic structures: A first-principles study of iridium manganese, Phys. Rev. Mater. 6 (4) (2022) 044402.

[71] A. Zunger, S.-H. Wei, L.G. Ferreira, J.E. Bernard, Special quasirandom structures, Phys. Rev. Lett. 65 (1990) 353–356.

[72] A. van de Walle, P. Tiwary, M. de Jong, D.L. Olmsted, M. Asta, A. Dick, D. Shin, Y. Wang, L.-Q. Chen, Z.-K. Liu, Efficient stochastic generation of special quasirandom structures, CALPHAD 42 (2013) 13–18.

[73] CASM website, 2022, URL https://prisms-center.github.io/CASMcode_docs/ (Accessed 18 May, 2022).

[74] A.H. Larsen, J.J. Mortensen, J. Blomqvist, I.E. Castelli, R. Christensen, M. Dułak, J. Friis, M.N. Groves, B. Hammer, C. Hargus, E.D. Hermes, P.C. Jennings, P.B. Jensen, J. Kermode, J.R. Kitchin, E.L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J.B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K.S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, K.W. Jacobsen, The atomic simulation environment—A Python library for working with atoms, J. Phys.: Condens. Matter 29 (27) (2017) 273002.

[75] J.C. Thomas, A.R. Natarajan, A. Van der Ven, Comparing crystal structures with symmetry and geometry, NPJ Comput. Mater. 7 (1) (2021) 1–11.

[76] A.R. Natarajan, A. Van der Ven, Connecting the simpler structures to topologically close-packed phases, Phys. Rev. Lett. 121 (25) (2018) 255701.

[77] A. Raju Natarajan, A. Van der Ven, Toward an understanding of deformation mechanisms in metallic lithium and sodium from first-principles, Chem. Mater. 31 (19) (2019) 8222–8229.

[78] A.R. Natarajan, P. Dolin, A. Van der Ven, Crystallography, thermodynamics and phase transitions in refractory binary alloys, Acta Mater. 200 (2020) 171–186.

[79] S.K. Kolli, A.R. Natarajan, A. Van der Ven, Six new transformation pathways connecting simple crystal structures and common intermetallic crystal structures, Acta Mater. 221 (2021) 117429.

[80] 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, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[81] G.L. Hart, V. Blum, M.J. Walorski, A. Zunger, Evolutionary approach for determining first-principles hamiltonians, Nature Mater. 4 (5) (2005) 391–394.

[82] V. Blum, G.L. Hart, M.J. Walorski, A. Zunger, Using genetic algorithms to map first-principles results to model Hamiltonians: Application to the generalized Ising model for alloys, Phys. Rev. B 72 (16) (2005) 165113.

[83] L.J. Nelson, G.L. Hart, F. Zhou, V. Ozolinš, et al., Compressive sensing as a paradigm for building physics models, Phys. Rev. B 87 (3) (2013) 035125.

[84] L.J. Nelson, V. Ozolinš, C.S. Reese, F. Zhou, G.L. Hart, Cluster expansion made easy with Bayesian compressive sensing, Phys. Rev. B 88 (15) (2013) 155105.

[85] T. Mueller, G. Ceder, Bayesian approach to cluster expansions, Phys. Rev. B 80 (2) (2009) 024103.

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

[87] M. Aldegunde, N. Zabaras, J. Kristensen, Quantifying uncertainties in first-principles alloy thermodynamics using cluster expansions, J. Comput. Phys. 323 (2016) 17–44.

[88] J.G. Goiri, A. Van der Ven, Recursive alloy Hamiltonian construction and its application to the Ni-Al-Cr system, Acta Mater. 159 (2018) 257–265.

[89] J. Kristensen, N.J. Zabaras, Bayesian uncertainty quantification in the evaluation of alloy properties with the cluster expansion method, Comput. Phys. Comm. 185 (11) (2014) 2885–2892.

[90] E.T. Jaynes, Information theory and statistical mechanics, Phys. Rev. 106 (4) (1957) 620.

[91] A. van de Walle, M. Asta, Self-driven lattice-model Monte Carlo simulations of alloy thermodynamic properties and phase diagrams, Modelling Simulation Mater. Sci. Eng. 10 (5) (2002) 521.

[92] A.A. Belak, A. Van der Ven, Effect of disorder on the dilute equilibrium vacancy concentrations of multicomponent crystalline solids, Phys. Rev. B 91 (22) (2015) 224109.

[93] B. Puchala, J.C. Thomas, A. Van der Ven, Monte Carlo calculations of the thermodynamic properties of complex multicomponent crystals, 2022 (in preparation).

[94] A.F. Kohan, P.D. Tepesch, G. Ceder, C. Wolverton, Computation of alloy phase diagrams at low temperatures, Comput. Mater. Sci. 9 (3–4) (1998) 389–396.

[95] M. Betancourt, A conceptual introduction to Hamiltonian Monte Carlo, 2017, arXiv preprint arXiv:1701.02434.

[96] B. Sadigh, P. Erhart, Calculation of excess free energies of precipitates via direct thermodynamic integration across phase boundaries, Phys. Rev. B 86 (2012) 134204, http://dx.doi.org/10.1103/PhysRevB.86.134204.

[97] A.B. Bortz, M.H. Kalos, J.L. Lebowitz, A new algorithm for Monte Carlo simulation of Ising spin systems, J. Comput. Phys. 17 (1) (1975) 10–18.

[98] A.R. Allnatt, Theory of phenomenological coefficients in solid-state diffusion. I. general expressions, J. Chem. Phys. 43 (6) (1965) 1855–1863.

[99] A.R. Allnatt, E.L. Allnatt, Computer simulation of phenomenological coefficients for atom transport in a random alloy, Phil. Mag. A 49 (5) (1984) 625–635, https://doi.org/10.1080/01418618408233291.

[100] A. Van der Ven, G. Ceder, First principles calculation of the interdiffusion coefficient in binary alloys, Phys. Rev. Lett. 94 (4) (2005) 045901.

[101] J.G. Goiri, A. Van der Ven, Phase and structural stability in Ni-Al systems from first principles, Phys. Rev. B 94 (9) (2016) 094111.

[102] J.C. Schuster, M. Palm, Reassessment of the binary aluminum-titanium phase diagram, J. Phase Equilibria Diffus. 27 (3) (2006) 255–277.

[103] N.S.H. Gunda, B. Puchala, A. Van der Ven, Resolving phase stability in the Ti-O binary with first-principles statistical mechanics methods, Phys. Rev. Mater. 2 (3) (2018) 033604.

[104] N.S.H. Gunda, A. Van der Ven, First-principles insights on phase stability of titanium interstitial alloys, Phys. Rev. Mater. 2 (8) (2018) 083602.

[105] L.-Q. Chen, Phase-field models for microstructure evolution, Annu. Rev. Mater. Res. 32 (1) (2002) 113–140.

[106] H.-C. Yu, C. Ling, J. Bhattacharya, J.C. Thomas, K. Thornton, A. Van der Ven, Designing the next generation high capacity battery electrodes, Energy Environ. Sci. 7 (5) (2014) 1760–1768.

[107] S. Rudraraju, A. Van der Ven, K. Garikipati, Mechanochemical spinodal decomposition: A phenomenological theory of phase transformations in multi-component, crystalline solids, NPJ Comput. Mater. 2 (1) (2016) 1–9.

[108] H.-C. Yu, A. Van der Ven, K. Thornton, Theory of grain boundary diffusion induced by the Kirkendall effect, Appl. Phys. Lett. 93 (9) (2008) 091908.

[109] H.-C. Yu, H.-Y. Chen, K. Thornton, Extended smoothed boundary method for solving partial differential equations with general boundary conditions on complex boundaries, Modelling Simulation Mater. Sci. Eng. 20 (7) (2012) 075008.

[110] G.H. Teichert, N.H. Gunda, S. Rudraraju, A.R. Natarajan, B. Puchala, K. Garikipati, A. Van der Ven, A comparison of Redlich-Kister polynomial and cubic spline representations of the chemical potential in phase field computations, Comput. Mater. Sci. 128 (2017) 127–139.

[111] S. DeWitt, E.L. Solomon, A.R. Natarajan, V. Araullo-Peters, S. Rudraraju, L.K. Aagesen, B. Puchala, E.A. Marquis, A. Van Der Ven, K. Thornton, et al., Misfit-driven β''' precipitate composition and morphology in Mg-Nd alloys, Acta Mater. 136 (2017) 378–389.

[112] G.H. Teichert, A.R. Natarajan, A. Van der Ven, K. Garikipati, Machine learning materials physics: Integrable deep neural networks enable scale bridging by learning free energy functions, Comput. Methods Appl. Mech. Engrg. 353 (2019) 201–216.

[113] G.H. Teichert, A.R. Natarajan, A. Van der Ven, K. Garikipati, Scale bridging materials physics: Active learning workflows and integrable deep neural networks for free energy function representations in alloys, Comput. Methods Appl. Mech. Engrg. 371 (2020) 113281.

[114] S.P. Ong, W.D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V.L. Chevrier, K.A. Persson, G. Ceder, Python materials genomics (pymatgen): A robust, open-source python library for materials analysis, Comput. Mater. Sci. 68 (2013) 314–319, http://dx.doi.org/10.1016/j.commatsci.2012.10.028.

[115] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.a. Persson, The Materials Project: A materials genome approach to accelerating materials innovation, APL Mater. 1 (1) (2013) 011002, http://dx.doi.org/10.1063/1.4812323.

[116] C.W. Andersen, R. Armiento, E. Blokhin, G.J. Conduit, S. Dwaraknath, M.L. Evans, Á. Fekete, A. Gopakumar, S. Gražulis, A. Merkys, F. Mohamed, C. Oses, G. Pizzi, G.-M. Rignanese, M. Scheidgen, L. Talirz, C. Toher, D. Winston, R. Aversa, K. Choudhary, P. Colinet, S. Curtarolo, D. Di Stefano, C. Draxl, S. Er, M. Esters, M. Fornari, M. Giantomassi, M. Govoni, G. Hautier, V. Hegde, M.K. Horton, P. Huck, G. Huhs, J. Hummelshøj, A. Kariryaa, B. Kozinsky, S. Kumbhar, M. Liu, N. Marzari, A.J. Morris, A.A. Mostofi, K.A. Persson, G. Petretto, T. Purcell, F. Ricci, F. Rose, M. Scheffler, D. Speckhard, M. Uhrin, A. Vaitkus, P. Villars, D. Waroquiers, C. Wolverton, M. Wu, X. Yang, OPTIMADE, an API for exchanging materials data, Sci. Data 8 (1) (2021) 217.

[117] M.L. Evans, C.W. Andersen, S. Dwaraknath, M. Scheidgen, Á. Fekete, D. Winston, 'Optimade-python-tools': A Python library for serving and consuming materials data via OPTIMADE APIs, J. Open Source Softw. 6 (65) (2021) 3458, http://dx.doi.org/10.21105/joss.03458.

[118] J.G. Goiri, A. Van der Ven, MultiShifter: Software to generate structural models of extended two-dimensional defects in 3D and 2D crystals, Comput. Mater. Sci. 191 (2021) 110310, http://dx.doi.org/10.1016/j.commatsci.2021.110310.

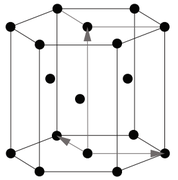
[119] A.R. Natarajan, A. Van der Ven, Machine-learning the configurational energy of multicomponent crystalline solids, NPJ Comput. Mater. 4 (1) (2018) 56.

CASM Project

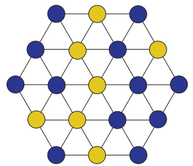
Parent Crystal Structure

Lattice vectors, $\boldsymbol{L} = \left(\vec{l}_1, \vec{l}_2, \vec{l}_3\right)$

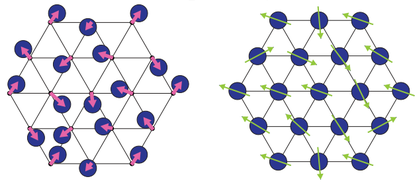
Basis site coordinates, $\boldsymbol{B} = \left(\vec{b}_1, \vec{b}_2, \ldots\right)$



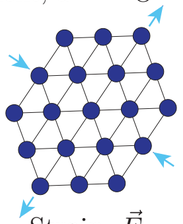
Degrees of Freedom (DoF)



Occupation, $\vec{s}$



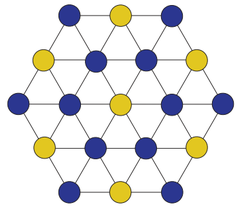
Displacement, $\vec{d}$ Magnetic spin, $\vec{m}$

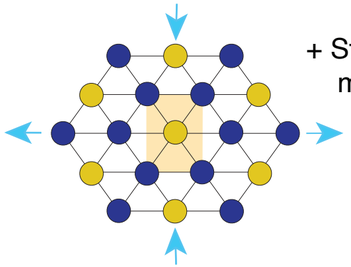


Strain, $\vec{E}$

Derivative Structure Enumeration

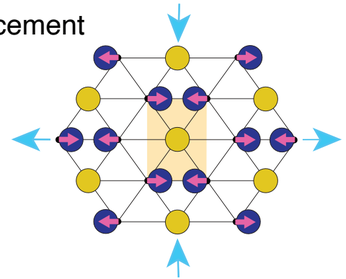
Discrete occupation ordering



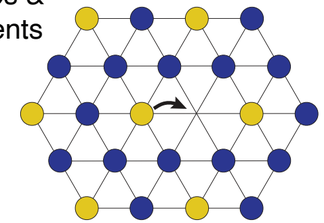


+ Strain modes

+ Displacement modes

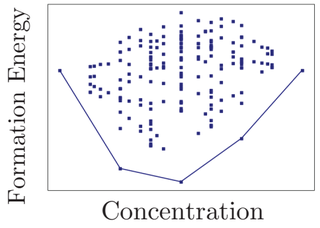


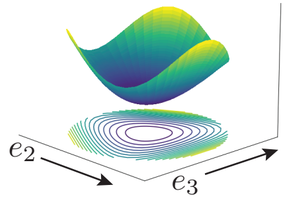
Migration event types & environments



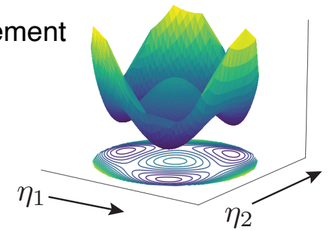
First Principles Calculations

Discrete occupation ordering energy





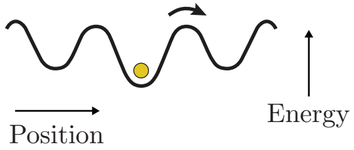
+ Strain energy



+ Displacement energy

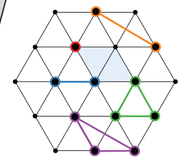
+ Relaxation strain, displacement, magnetic spin, orientation, ...

Migration energy



Cluster Expansion Effective Hamiltonian

$$E(\vec{s}, \vec{d}, \vec{m}, \vec{q}, \vec{E}) = \sum_i m_i V_i \Gamma_i(\vec{s}, \vec{d}, \vec{m}, \vec{q}, \vec{E})$$



Symmetry-adapted basis functions, Γ_i

Expansion coefficients, V_i

Cluster multiplicity, m_i

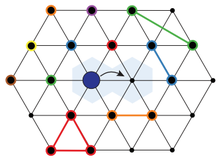
Monte Carlo Calculations

Metropolis, N-fold way, Hamiltonian

Kinetic Monte Carlo Calculations

Local-cluster Expansion of Diffusion Barriers

$$\Delta E_\delta^{KRA}(\vec{s}) = \sum_i m_i V_i \Gamma_i(\vec{s})$$



Finite Temperature Properties