Machine learning the DFT potential energy surface for inorganic halide perovskite $CsPbBr_3$

John C. Thomas,* Jonathon S. Bechtel, Anirudh Raju Natarajan, and Anton Van der Ven†
Materials Department, University of California, Santa Barbara, Santa Barbara, CA 93106
(Dated: February 10, 2020)

Structural phase transitions as a function of temperature dictate the structure–functionality relationships in many technologically important materials. Harmonic Hamiltonians have proven successful in predicting the vibrational properties of many materials. However, they are inadequate for modeling structural phase transitions in crystals with potential energy surfaces that are either strongly anharmonic or non-convex with respect to collective atomic displacements or homogeneous strains. In this paper we develop a framework to express highly anharmonic first-principles potential energy surfaces as polynomials of collective cluster deformations. We further adapt the approach to a nonlinear extension of the cluster expansion formalism through the use of an artificial neural net model. The machine learning models are trained on a large database of first-principles calculations and are shown to reproduce the potential energy surface with low error.

I. INTRODUCTION

Structural phase transitions are widespread among technologically important materials. Statistical mechanics approaches based on the quasi-harmonic approximation are well suited to describe the finite temperature thermodynamic properties of phases that reside at a local minimum in the potential energy surface (PES) of a particular compound. The harmonic approximation, however, breaks down for high temperature phases whose symmetry coincides with a saddle point on the PES. These phases only emerge at elevated temperature due to anharmonic vibrational excitations. A wide variety of high temperature phases fall in this category. These include the bcc forms of Ti, Zr and Hf[1–3], the high temperature cubic form of ZrO_2[4–6], hydrides such as TiH_2[7] and ZrH_2[8] as well as many cubic perovskite phases, including halide perovskites.[9–11]

While direct *ab initio* molecular dynamics simulations can be used to study the elevated temperature properties of anharmonically stabilized phases[12], the computational cost of density functional theory (DFT) calculations often makes such an approach intractable. An alternative is to rely on a model that is capable of accurately interpolating and extrapolating a limited number of DFT calculated energies within Monte Carlo or molecular dynamics simulations.

Several methods have been developed to extrapolate the first-principles PES of a compound for the purpose of studying group/subgroup structural phase transitions [7, 8, 13–24]. In the study of group/subgourp structural transitions, the PES is typically expressed as a function of descriptors of local atomic structure. It is often convenient to formulate these descriptors as nonlinear functions of atomic displacements measured relative to the highest symmetry phase participating in the transition. However, a challenge of this approach is to determine how the PES depends on these descriptors. In particular, the descriptors must be invariant to rigid translations and rotations of the crystal, which comprise nontrivial and highly nonlinear constraints.

Traditional approaches are based upon a Taylor expansion of the PES in terms of the Cartesian components of atomic displacement vectors. Each term of the Taylor expansion consists of a constant, that can be treated as a chemistry dependent adjustable parameter, multiplied by a polynomial of the Cartesian components of the displacement vectors belonging to clusters of sites in the crystal. The harmonic approximation emerges as the lowest order truncation of the Taylor expansion and consists exclusively of terms corresponding to point and pair clusters of sites. Since polynomials of the components of displacement vectors are not invariant to rigid translations and rotations of the crystal, constraints must be imposed on the expansion coefficients, which become increasingly onerous as the order of truncation of the Taylor expansion increases.

The anharmonic cluster expansion [8] follows a similar approach in that it expresses the PES as a sum of terms that depend on deformations of clusters of sites. However, instead of depending directly on the Cartesian components of displacement vectors, each term corresponding to a cluster of sites is expressed as a function of collective cluster deformation coordinates that are formulated to be invariant to translations and rotations of the cluster from the outset. The PES is then represented as a linear expansion of terms consisting of adjustable parameters multiplied by polynomials of the collective cluster deformation variables. In the original formulation of the method, a pre-rotation step was required that relies on the computationally expensive Kabsch algorithm to determine the collective cluster deformation variables from the individual displacement vectors of the sites belonging to each cluster.

The aims of this contribution are two fold. First we introduce new descriptors of cluster deformations that are invariant to rigid translations and rotations of clusters of

* johnct@ucsb.edu
† avdv@engineering.ucsb.edu

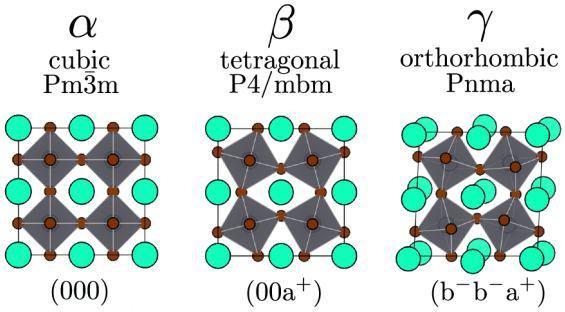

FIG. 1. High temperature cubic α-phase, intermediate temperature tetragonal β-phase, and low temperature orthorhombic γ-phase of the $CsPbBr_3$ perovskite which is the same phase sequence for many inorganic halide Cs-based perovskites.

sites and that can be evaluated rapidly without the need to resort to the Kabsch algorithm. Secondly we explore the use of neural networks to identify the optimal functional dependence of the PES on the collective cluster deformation variables. As a model system, we focus on the halide perovskite $CsPbBr_3$, a representative compound from a very promising family of chemistries for photovoltaic applications. The perovskite form of $CsPbBr_3$ undergoes a series of group/subgroup structural phase transitions involving octahedral tilts upon cooling, adopting a cubic symmetry at high temperature, transforming upon cooling to tetragonal symmetry at 403 K, and transforming upon further cooling to its groundstate orthorhombic phase at 361 K[25]. Since the different phases of inorganic halide perovskites can be connected by symmetry-lowering displacement modes from the high temperature cubic phase, it is convenient to parameterize the energy landscape in terms of distortions of the cubic reference.

II. METHOD

We start with the anharmonic cluster expansion approach to representing the PES of a crystal. Within this approach, the energy of a crystal as a function of atomic displacements, $\vec{u}_i$, relative their sites, i, in a high symmetry reference crystal is expressed as

$$E\left(\ldots,\vec{u}_i,\ldots\right) = E_o + \sum_{\alpha} \Phi^{\alpha}\left(q_1^{\alpha},\ldots,q_{N_\alpha}^{\alpha}\right) \qquad (1)$$

where E_o is the energy of the reference crystal and the $\Phi^{\alpha}\left(q_1^{\alpha},\ldots,q_{N_\alpha}^{\alpha}\right)$ are functions associated with clusters of sites α. The variables $\vec{Q}^{\alpha} = (q_1^{\alpha},\ldots,q_{N_\alpha}^{\alpha})$ are functions of the displacements, $\vec{u}_i$, of the sites of the cluster α, and uniquely describe the degree with which cluster α is distorted relative to its state in the reference crystal. The deformation variables, $\vec{Q}^{\alpha}$, must be invariant to rigid translations and rotations of the cluster to ensure that the energy of the crystal is itself invariant to rigid translations and rotations. Both the deformation variables, $\vec{Q}^{\alpha}$, and the cluster interaction functions, Φ^{α}, are zero in the reference crystal. The clusters that appear in Eq. (1) usually consist of compact, non-overlapping multi-body clusters such as tetrahedra or octahedra as well as terms for a number of longer-range pair clusters.

In the anharmonic cluster expansion of ref [8], the cluster interaction functions Φ^{α} are expressed as an expansion of cluster basis functions according to

$$\Phi^{\alpha}\left(q_1^{\alpha},\ldots,q_{N_\alpha}^{\alpha}\right) = \sum_{m} V_m^{\alpha}\phi_m^{\alpha}\left(q_1^{\alpha},\ldots,q_{N_\alpha}^{\alpha}\right) \qquad (2)$$

where the $\phi_m^{\alpha}\left(q_1^{\alpha},\ldots,q_{N_\alpha}^{\alpha}\right)$ are polynomials of the elements of $\vec{Q}^{\alpha}$ and are formulated to be invariant to symmetry operations that map the reference crystal onto itself. The expansion coefficients, V_m^{α}, are determined by the chemistry of the compound and can be treated as adjustable parameters to reproduce DFT energies calculated for a sufficiently large training set of vibrational excitations relative to the reference crystal. The requirement that the cluster basis functions, $\phi_m^{\alpha}\left(q_1^{\alpha},\ldots,q_{N_\alpha}^{\alpha}\right)$, are invariant to symmetries of the reference crystal ensures that the energy of any two distortion fields $(\ldots,\vec{u}_i,\ldots)$ and $(\ldots,\vec{u}_i',\ldots)$ that are related by a symmetry operation of the reference crystal have the same energy when evaluated with Eq. (1). Polynomial basis functions, ϕ_m^{α} extend to arbitrary order, but in practice only terms up to order 4 or 6 in terms of the elements of $\vec{Q}^{\alpha}$ are kept.

In the next sections, we introduce a new set of collective cluster deformation variables $\vec{Q}^{\alpha}$ that uniquely describe deformations of a cluster α and that are also invariant to any rigid translation or rotation of the cluster. We then introduce an approach that relies on neural networks to train cluster interaction functions $\Phi^{\alpha}\left(q_1^{\alpha},\ldots,q_{N_\alpha}^{\alpha}\right)$ that go beyond a linear expansion of cluster basis functions as in Eq. (2).

A. Collective cluster deformation variables and symmetry invariant descriptors of deformation

1. *Pair distances as measures of cluster deformations*

The starting ingredient to construct robust collective cluster deformation variables are the collection of all pair distances between the sites of a cluster α. This is motivated by the following property: Given the set of all distances between pairs of atoms in a particular deformed cluster α, it is possible to exactly reconstruct the full geometry of α, to within a rigid rotation and translation[26]. We introduce $\vec{D}^{\alpha} = (d_1,\ldots,d_l,\ldots,d_{N_\alpha})$ as comprising the pair distances d_l between sites of a n_α-point cluster where l indexes unique i,j pairs of the cluster and where N_α is the number of unique pairs in a n_α-point cluster (i.e. $N_\alpha = n_\alpha(n_\alpha-1)/2$).

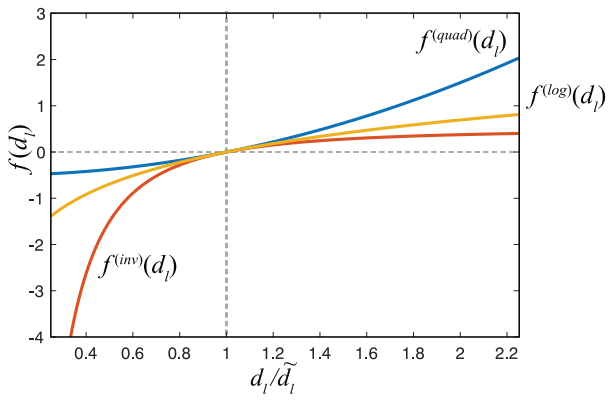

FIG. 2. Illustration of difference deformation metric functions (Figure will change to match text)

We are not limited to pair distances in constructing rotationally- and translationally-invariant deformation metrics for clusters of atoms. Any smooth monotonic function of the pair distances that can be inverted to obtain pair distances can also be used to define deformation metrics. We can thus fully specify the cluster geometry via the vector $\vec{F}^\alpha = (f_1, \ldots, f_l, \ldots, f_{N_\alpha})$, where $f_l = f(d_l)$. A simple choice for a deformation metric is the linear function $f^{(lin)}(d_l) = (d_l/\tilde{d}_l - 1)$, where $\tilde{d}_l$ is the length of the pair l in the reference crystal, though other functional forms have their own advantages, such as

$$f^{(quad)}(d_l) = (d_l^2/\tilde{d}_l^2 - 1)/2, \quad (3)$$

$$f^{(log)}(d_l) = \ln(d_l/\tilde{d}_l), \text{ and} \quad (4)$$

$$f^{(inv)}(d_l) = (1 - \tilde{d}_l^2/d_l^2)/2. \quad (5)$$

These functions, which are depicted in Fig. 2, all become equal to zero in the reference state (i.e. when $d_l = \tilde{d}_l$) and have identical slopes in the vicinity of the reference distance, thereby being equivalent for very small deformations. However, the behavior of each function is quite distinct at large deformations, as $d_l \to 0$ or $d_l \to \infty$.

2. *Collective cluster deformation variables*

While the vector $\vec{F}^\alpha$ fully determines the deformation state of the cluster, any linear transformation

$$\vec{Q}^\alpha = \mathbf{U}\,\vec{F}^\alpha, \quad (6)$$

where $\mathbf{U}$ is a full-rank $N_\alpha \times N_\alpha$ matrix, yields a vector $\vec{Q}^\alpha$ that also fully describes the cluster deformation state. A suitable choice for $\mathbf{U}$ is motivated by symmetry considerations.

The symmetries of a cluster are described by the cluster point group, which are the set of rotations and/or reflections, centered at the cluster, that map the cluster onto itself. [27] For a cluster embedded in a crystal, the cluster point group must also leave the crystal unchanged, and so the cluster point group is a subgroup of the crystal space group.

Application of a cluster point group operation $\hat{c}$ to a cluster α may permute the sites of the cluster. Formally, if the reference coordinates of the cluster are columns of the $3 \times n_\alpha$ matrix $\mathbf{R}^\alpha = (\vec{r}_1^{\,\alpha} | \ldots | \vec{r}_{n_\alpha}^{\,\alpha})$, then action of $\hat{c}$ can be expressed as

$$\hat{c}\,[\mathbf{R}^\alpha] = \mathbf{S}(\hat{c})\,\mathbf{R}^\alpha + \mathbf{T}(\hat{c}) \quad (7)$$

where $\mathbf{S}(\hat{c})$ is an orthogonal 3×3 matrix (i.e., rotation, reflection, or rotoreflection) and $\mathbf{T}(\hat{c})$ is a $3 \times n_\alpha$ translation matrix. The effect of a symmetry operation of the cluster on its reference coordinates is simply to permute the coordinates. This can equivalently be represented as

$$\hat{c}\,[\mathbf{R}^\alpha] = \mathbf{R}^\alpha\,\mathbf{W}^\top(\hat{c}), \quad (8)$$

where $\mathbf{W}(\hat{c})$ is a $n_\alpha \times n_\alpha$ permutation matrix describing the permutation of the columns of $\mathbf{R}^\alpha$.

Just as a symmetry operation, $\hat{c}$, permutes the coordinates of the cluster, it also permutes the order of each distinct pair $l = (i, j)$. The application of a symmetry operation will therefore reorder the elements of $\vec{F}^\alpha$. This can expressed as

$$\vec{F}'^\alpha = \mathbf{M}^{(F)}(\hat{c})\;\vec{F}^\alpha \quad (9)$$

where $\vec{F}'^\alpha$ and $\vec{F}^\alpha$ represent two deformations of the reference cluster that are related to each other by the cluster point group operation $\hat{c}$. The elements of $\mathbf{M}^{(F)}(\hat{c})$ are given by

$$\mathbf{M}^{(F)}_{(ij),(kl)}(\hat{c}) = \mathbf{W}(\hat{c})_{ik}\mathbf{W}(\hat{c})_{jl}. \quad (10)$$

In the above equation we have used the compound indices (ij) and (kl) to indicate atomic pairs after and before application of symmetry, respectively. The symmetry representation $\mathbf{M}^{(F)}(\hat{c})$ is also a permutation matrix that describes the discrete exchange of atomic pairs due to application of symmetry.

By combining Eq. (6) and Eq. (9), we can determine the effect of the application of $\hat{c}$ on the collective cluster deformation variables, $\vec{Q}^\alpha$ according to

$$\hat{c}\left[\vec{Q}^\alpha\right] = \mathbf{U}\,\mathbf{M}^{(F)}(\hat{c})\;\mathbf{U}^{-1}\,\vec{Q}^\alpha = \mathbf{M}^{(Q)}(\hat{c})\;\vec{Q}^\alpha, \quad (11)$$

where $\mathbf{M}^{(Q)}(\hat{c})$ is the matrix representation describing the action of $\hat{c}$ on $\vec{Q}^\alpha$.

Equation (11) motivates a choice for the matrix $\mathbf{U}$ relating the sought after collective cluster deformation variables, $\vec{Q}^\alpha$, to the elements of $\vec{F}^\alpha$, which are each individually a function of a pair distance in the cluster. We will use the matrix $\mathbf{U}$ that simultaneously block diagonalizes all the symmetry matrices $\mathbf{M}^{(Q)}(\hat{c})$ of the cluster point group. This choice for $\mathbf{U}$ generates collective cluster deformation variables $\vec{Q}^\alpha$ that reside in subspaces that transform under symmetry according to the

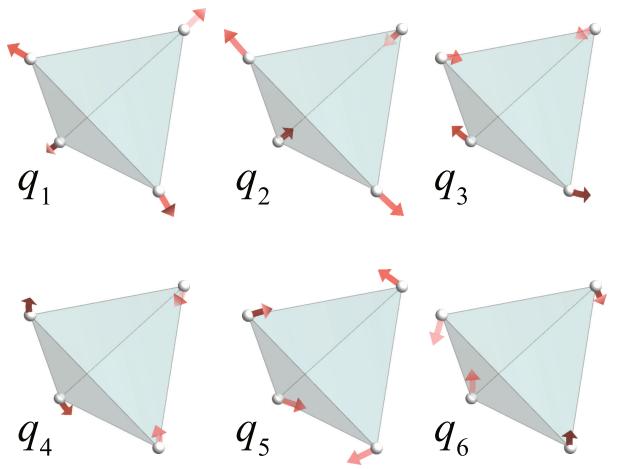

FIG. 3. Visualization of the six CCDs of a tetrahedron having T_d symmetry.

irreducible representations of the cluster point group[28]. Not only does this choice simplify the formulation of polynomials of the elements of $\vec{Q}^\alpha$ that are invariant to the symmetry of the crystal, but it also ensures that the $\vec{Q}^\alpha$ can serve as order-parameters with which to detect group/subgroup symmetry breaking transitions [28–30]. The elements of $\mathbf{U}$ for a tetrahedron cluster (assuming cluster point group T_d) and an octahedron cluster (assuming cluster point group O_h) are provided in the supporting information[31].

3. *Visualizing collective cluster deformations*

We can visualize the collective distortions that are activated upon independently varying a particular CCD component q_n^α by superimposing unit vectors proportional to $\partial \vec{r}_i^{\,\alpha}/\partial q_n^\alpha|_{q_{m\neq n}^\alpha=0}$ at each site i of the cluster. While these partial derivatives cannot be calculated directly, they can be obtained by inverting the Jacobian matrix whose elements are $J_{ij}(\vec{R}^\alpha) = \partial q_i^\alpha/\partial r_j^\alpha|_{\vec{R}^\alpha}$[32]. The inverse of the Jacobian has elements $[J^{-1}(\vec{R}^\alpha)]_{ij} = \partial \vec{r}_i^{\,\alpha}/\partial q_j^\alpha|_{\vec{Q}^\alpha}$.

Figure 3 shows the collective deformation modes corresponding to each element of $\vec{Q}^\alpha$ for a tetrahedron cluster. There are six such modes, with q_1^α corresponding to volumetric (i.e., symmetry-preserving) deformation. The modes corresponding to $(q_2^\alpha, q_3^\alpha, q_4^\alpha)$ belong to the T_2 irrep of T_d, and capture symmetry breaking to trigonal, orthorhombic, and monoclinic point groups. The modes corresponding (q_5^α, q_6^α) belong to the E irrep of T_d, and capture symmetry breaking to tetragonal and orthorhombic point groups.

Figure 4 shows the collective deformation modes corresponding to each element of $\vec{Q}^\alpha$ for a six-point octahedron cluster having O_h point symmetry in its reference state. There are 15 such modes, with q_1^α and q_{13}^α corresponding to volumetric (i.e., symmetry-preserving) deformation. The modes corresponding to $(q_2^\alpha, q_3^\alpha, q_4^\alpha)$ belong to the T_{1u} irrep of O_h. The modes corresponding to (q_5^α, q_6^α) and $(q_{14}^\alpha, q_{15}^\alpha)$ belong to the E_g irrep of O_h. The modes corresponding to $(q_7^\alpha, q_8^\alpha, q_9^\alpha)$ belong to the T_{2g} irrep of O_h, while the modes corresponding to $(q_{10}^\alpha, q_{11}^\alpha, q_{12}^\alpha)$ belong to the T_{2u} irrep.

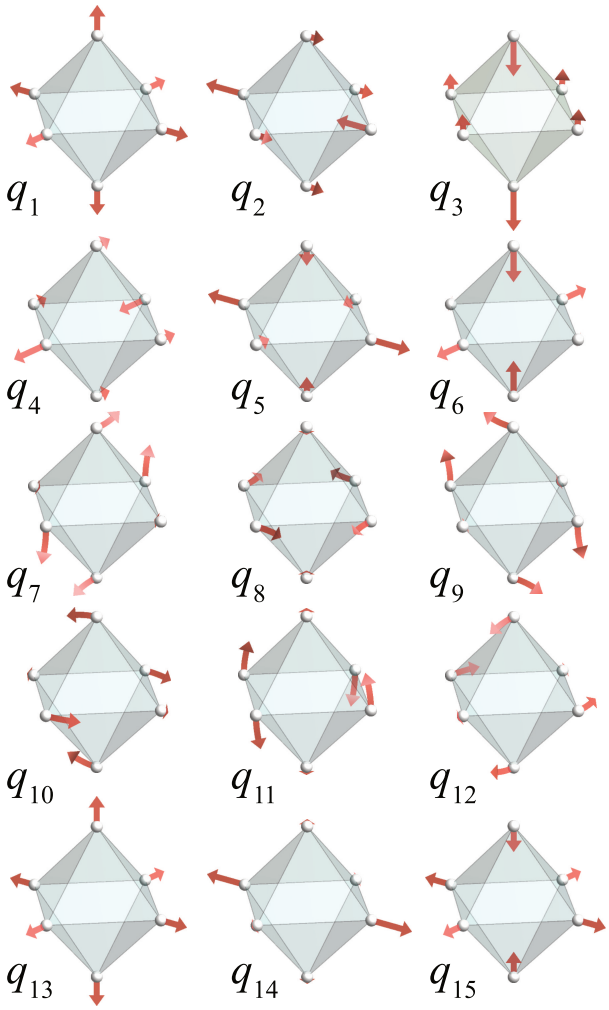

FIG. 4. Visualization of the 15 CCDs of a tetrahedron having O_h symmetry.

4. *Redundancy of cluster deformations parameters*

A non-planar cluster in three dimensions comprising n_α sites has $3n_\alpha - 6$ deformational degrees of freedom after removal of the six rigid translational and rotational degrees of freedom. The dimension of the CCD vector $\vec{Q}^\alpha$, in contrast, is $n_\alpha(n_\alpha - 1)/2$. This means that the number of CCD variables will be greater that the number of independent deformational degrees of freedom when n_α is greater 4. The realizable values of the CCDs then reside on a $3n_\alpha - 6$ dimensional surface (differentiable manifold) within the $n_\alpha(n_\alpha - 1)/2$ dimensional space

spanned by the CCDs. This suggests a degree of redundancy among the CCD variables whereby only $3n_\alpha - 6$ of the $n_\alpha(n_\alpha - 1)/2$ CCD variables are strictly necessary to characterize the deformation state of the cluster. While this is generally the case when the CCD variables are used to track deformations that preserve the topology of the reference cluster, there are situations where all CCD values are necessary to precisely reconstruct the geometry of the deformed cluster.

As an example, the octahedron cluster depicted in Fig. 4 has 15 distinct CCDs but only 12 degrees of freedom. The CCDs q_1^α and q_{13}^α are qualitatively similar, as are the pairs of CCDs (q_5^α, q_6^α) and $(q_{14}^\alpha, q_{15}^\alpha)$. In addition to having identical symmetry properties, these paired sets of CCDs also describe qualitatively identical deformation modes, as demonstrated by their visualized deformation trajectories in Fig. 4. The nature of redundant CCDs is described in more detail in the appendix, where a procedure is outlined to identify the most important CCD variables for topology preserving deformations.

5. *Symmetry invariant polynomials of the collective cluster deformation variables*

The collective cluster deformation variables, $\vec{Q}^\alpha$, are constructed to be invariant to rigid translations and rotations of the cluster and have been symmetry adapted such that they transform according to the irreducible representations of the cluster point group. The next task is to generate the polynomial basis functions, $\phi_m^\alpha(\vec{Q}^\alpha)$, appearing in Eq. (2). These functions are to be invariant to the point group symmetry of the cluster. Polynomial basis functions that are invariant to all symmetry operations of a point group that act on the arguments of the polynomial can be generated algorithmically using the Reynolds operator. This is described in [8, 29]. For the ϕ_m^α basis functions, the approach requires the symmetry representations, $\mathbf{M}^{(Q)}(\hat{c})$, of each cluster point group symmetry operation $\hat{c}$ that acts on $\vec{Q}^\alpha$. Symmetry-invariant CCD polynomials for the ideal 4-site tetrahedron cluster and the ideal 6-site octahedron cluster are provided in the supporting information.

B. Machine learning the potential energy landscape of a crystal

The anharmonic cluster expansion, Eqs. (1) and (2), can serve to interpolate and extrapolate the energies of a limited subset of first-principles calculations of different vibrational excitations of the reference crystal. The adjustable parameters V_m^α that appear in Eq. (2) can be fit to a training set of DFT energies using a variety of approaches that are commonly used to parameterize other lattice models such as alloy cluster expansions [33–36].

An alternative approach, that we pursue here, is to machine learn the PES as a function of descriptors that measure the distortion of the reference crystal. We rely on the anharmonic cluster expansion as a starting point. Descriptors of crystal distortions must satisfy several invariance relationships. First, they must be invariant to rigid translations and rotations of the crystal. Second, they must be invariant to the space-group symmetries of the reference crystal to ensure that symmetrically equivalent deformation states of the crystal evaluate to the same energy. If the descriptors do not satisfy these constraints, they would need to be learned, necessitating a much larger training set.

The anharmonic cluster expansion can guide the identification of suitable descriptors. While the collective cluster deformation variables $\vec{Q}^\alpha$ are invariant to rigid translations and rotations, they are not invariant to the symmetry operations of the crystal. The polynomial basis functions, $\phi_m^\alpha(\vec{Q}^\alpha)$, appearing in Eq. (2), however, are invariant to the symmetry of the crystal and evaluate to the same value for all symmetrically equivalent cluster deformations. Since they are a function of the CCDs, they are also invariant to rigid translations and rotations. A sufficient number of cluster basis functions, ϕ_m^α, can therefore serve as a finger print for each symmetrically distinct distortion state of a cluster.

The approach we follow to machine learn the PES will rely on Eq. 1, but will relax the linearity of the expansion in Eq. 2. Instead of expressing the cluster interaction functions Φ^α as a linear expansion of the cluster basis functions, ϕ_m^α, we will train a model that has a nonlinear dependence on the basis functions $\{\phi_m^\alpha\}$, which are themselves functions of the CCDs, $\vec{Q}^\alpha$. We will explore two architectures for this model: a cluster-centric architecture and a site-centric architecture, relying on artificial neural networks in either case to approximate the nonlinear functional form of Φ^α.

1. *Cluster-based neural net*

To set up a cluster-based neural net description of the PES, we first rewrite Eq. (1) in a manner that exploits the symmetries of the reference crystal. Many clusters of sites in the reference crystal are equivalent to each other by a space group operation of the reference crystal. For a cluster α, we denote the set of all symmetrically equivalent clusters by $\Omega(\alpha)$, referred to as the *orbit* of cluster α. By symmetry, all clusters belonging to a particular orbit $\Omega(\alpha)$ will have the same cluster interaction function $\Phi^{\Omega(\alpha)}$. The anharmonic cluster expansion can then be rewritten as

$$E\left(\ldots, \vec{u}_i, \ldots\right) = E_o + \sum_\alpha \Phi^{\Omega(\alpha)}\left(q_1^\alpha, \ldots, q_{N_\alpha}^\alpha\right). \quad (12)$$

Importantly, this expression indicates that although a particular cluster, such as a nearest-neighbor Pb–Pb pair, is repeated in all directions throughout the crystal, its pair interaction function can be reduced to a single func-

tional form, $\Phi^{\Omega(\alpha)}$, that is then evaluated locally for each equivalent cluster.

Instead of relying on the linear expansion for $\Phi^{\Omega(\alpha)}\left(q^\alpha_1,\ldots,q^\alpha_{N_\alpha}\right)$, we replace it with a neural net that has as inputs, not the CCDs, but rather a sufficiently large number of cluster basis functions $\{\phi^\alpha_m\}$. The energy expression can then be written as

$$E\left(\ldots,\vec{u}_i,\ldots\right) = E_o + \sum_\alpha \mathcal{N}^{\Omega(\alpha)}\left(\ldots,\phi^\alpha_m,\ldots\right) \quad (13)$$

where a separate neural net, $\mathcal{N}^{\Omega(\alpha)}$, approximates the energy contribution for each distinct cluster orbit. A visual interpretation of the computational graph for a cluster-based neural net model is depected in Figure 5(a).

2. *Site-based neural net*

An alternative approach to representing the PES is with a site-centric expression. To this end, we define a site-centric orbit $\Omega_i(\alpha)$ that contains all clusters β that are symmetrically equivalent to cluster α and that also contain site i. The orbit $\Omega_i(\alpha)$ then contains all clusters emanating from site i that are symmetrically equivalent to α. In terms of the site-centric orbits, we can rewrite the linear anharmonic cluster expansion as

$$E\left(\ldots,\vec{u}_j,\ldots\right) = E_o + \sum_i \sum_\alpha \sum_m \frac{1}{n_\alpha} V^\alpha_m \sum_{\beta\in\Omega_i(\alpha)} \phi^\alpha_m\left(\vec{Q}^\beta\right) \quad (14)$$

where the sum over α is restricted to include only one *cluster prototype* for each symmetrically distinct cluster orbit. The outer sum is over all sites i in the crystal, while the innermost sum accumulates the combined contribution from all clusters that are equivalent to α and that include site i. The factor of $1/n_\alpha$ corrects for overcounting due to the fact that the contribution for an individual cluster appears once for each of its constituent sites. Equation (14) emerges upon combining Eq. (1) and (2) and exploiting the linearity in Eq. (2).

Equation (14) motivates the introduction of site-centric correlation functions defined as

$$g^i_{\alpha,m} = \frac{1}{n_\alpha} \sum_{\beta\in\Omega_i(\alpha)} \phi^\alpha_m(\vec{Q}^\beta) \quad (15)$$

The sum extends over all clusters β that are equivalent to α by a crystal space group operation and that also include site i, ensuring that $g^i_{\alpha,m}$ is invariant to the subgroup of the space group that maps site i onto itself. This property guaranties that $g^i_{\alpha,m}$ evaluates to the same value for all distortion fields that are related to each other by a symmetry operation of the reference crystal. A feature vector $\vec{G}^i = \left(g^i_{\alpha,1},\ldots,g^i_{\alpha,m},\ldots,g^i_{\alpha',1},\ldots\right)$, formed by the site-centric correlation functions serves as an arbitrarily detailed descriptor of the local distortion in the vicinity of site i. The feature vector can be systematically improved by increasing the variety and cutoff range

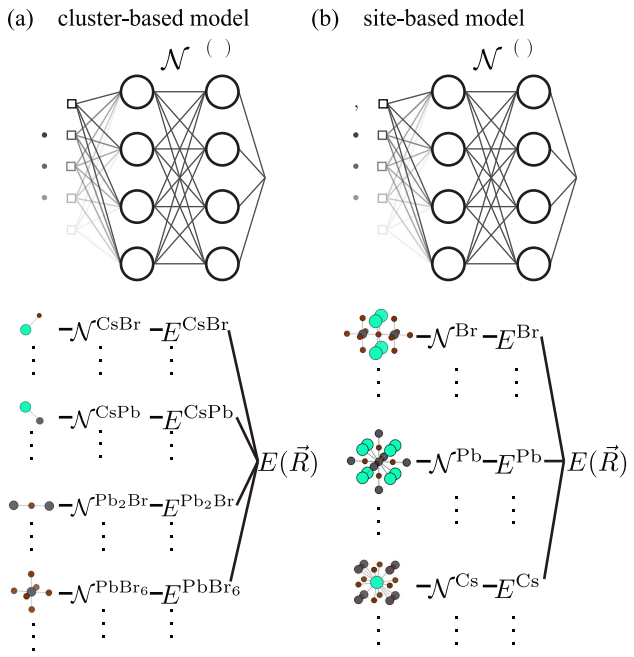

FIG. 5. Visualization of how (a) site-based and (b) cluster-based models incorporate site-averaged basis functions or cluster-based basis functions respectively.

of symmetrically distinct clusters, α, α' etc., constituting the descriptor, as well as the order of their corresponding basis functions.

A site-centric neural net description of the PES is thus formulated in terms of the feature vector $\vec{G}^i$ according to

$$E\left(\ldots,\vec{u}_j,\ldots\right) = E_o + \sum_i \mathcal{N}^{\eta(i)}(\vec{G}^i) \quad (16)$$

where the total energy of the crystal is a sum over contributions from each individual site i. $\eta(i)$ refers to the orbit of all sites equivalent to site i with respect to the symmetry of the reference crystal, such that there is a separate approximation function, $\mathcal{N}^{\eta(i)}$, for each symmetrically distinct site of the reference crystal. The site-based neural net model is summarized in Figure 5(b).

3. *Artificial Neural Network*

Whether working in the site-based or cluster-based cluster expansion, we make use of artificial neural network models that take as inputs $\vec{x}$ (where x_i could be either the local-orbit summed basis functions in the site-based model, or simply the evaluated basis functions in the cluster based model) and output an energy e. Artificial neural networks are hierarchical recursive functions made up of activation nodes f_i which represents a non-linear function f at node i. A one-layer neural net

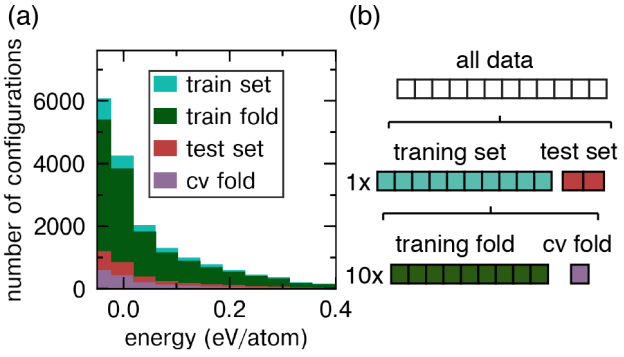

FIG. 6. (a) Distribution of energies for all configurations in the database. (b) All data is split into a training set and test set. The training set is further subdivided into 10 training folds and 10 validation folds for use in hyperparameter tuning.

produces output e from inputs $\mathbf{x}$ as follows:

$$e = b^{(1)} + \sum_j w_j^{(1)} f_j(b_j^{(0)} + \sum_k x_k W_{kj}^{(0)}) \qquad (17)$$

where $b^{(1)}$ is a bias term associated with the 1st layer, and $b_j^{(0)}$, are bias terms associated with the input layer into node j of the first hidden layer. $W_{kj}^{(0)}$ represents the weight matrix connecting the input layer to the first hidden layer, and $w_j^{(1)}$ is the weight matrix connecting the hidden layer to the output layer. The model variables are the weights and biases which are trained through optimization techniques described below. Layers having the general function form of Eq. (17) can applied in sequence to form an arbitrarily complex network function. Details of how various network architectures may be formulated for a crystalline system are provided in **??**. The activation function, f_j, can take several forms including the hyperbolic tangent, rectified linear unit, or logistic function. In this study we used the hyperbolic tangent exclusively.

4. Objective Function

In order to train the neural network model, we must minimize a convex objective function. Here we choose an objective function that penalizes the sum of the squares of the differences in model energies and those calculated with DFT for a large number of different vibrational excitations.

$$\Gamma = \sum_\sigma \left(E_{\mathrm{ANN}}(\sigma) - E_{\mathrm{DFT}}(\sigma)\right)^2 \qquad (18)$$

where σ denotes different vibrational excitations. The objective function is minimized with respect to the weights of the neural network. Many optimization algorithms exist to optimize the weights of the network function. We employed the Adam algorithm in this study[37, 38].

III. POTENTIAL ENERGY SURFACE OF HALIDE PEROVSKITES

In this section we develop a neural network model of the potential energy surface of $CsPbBr_3$, a compound belonging to a class of promising perovskite based materials for electronic and photovoltaic applications. $CsPbBr_3$ undergoes a series of group/subgroup structural phase transitions upon cooling. At high temperature, $CsPbBr_3$ is stable in a cubic perovskite crystal structure, but transitions to tetragonal and orthorhombic symmetries at lower temperatures due to tilting of the $PbBr_3$ octahedra. As with many halide perovskites, the cubic and tetragonal forms of perovskite $CsPbBr_3$ correspond to saddle points on the potential energy surface of the compound [39]. These phase only emerge at finite temperature due to large scale anharmonic vibrational excitations.

A. DFT

Density functional theory calculations were performed using the Vienna Ab Initio Simulation Package (VASP). [40, 41] A plane wave basis set with an energy cutoff of 400 eV was employed and projector augmented wave psuedopotentials (PAW). [40, 42] The GGA-PBEsol functional was used to approximate electron correlation and exchange. [43] Energies were converged to within 1 meV / atom with respect to k-point density and a 6×6×6 Γ-centered k-point mesh was used for the $CsPbBr_3$ unit cell. The VESTA program suite was used to visualize crystal structures.

B. Training Set

The training set is a critical component in a machine learning problem. The resulting model is only as good as the training set. Each element of the training set, corresponding to the energy of a particular state of strain and a particular set of atomic displacements relative to the high symmetry reference state, will be referred to as a configuration, σ. The most important regions of the PES include the potential energy wells in which the ground state structure resides. Therefore, much effort was made to sample configurations near the ground state structure along with the structures associated with the intermediate tetragonal phase and the high temperature cubic phase.

Sampling the PES was done in several ways. The starting point began with the geometric relaxation of the 15 tilt systems as previously described in [39]. For each of these relaxed structures, systematic displacement enumerations were made in terms of symmetry-adapted displacement modes, i.e. the displacement fields that block diagonalize the crystal symmetry representation. The same supercell (2×2×2) was used for all of the tilt systems to avoid numerical errors incurred when using differ-

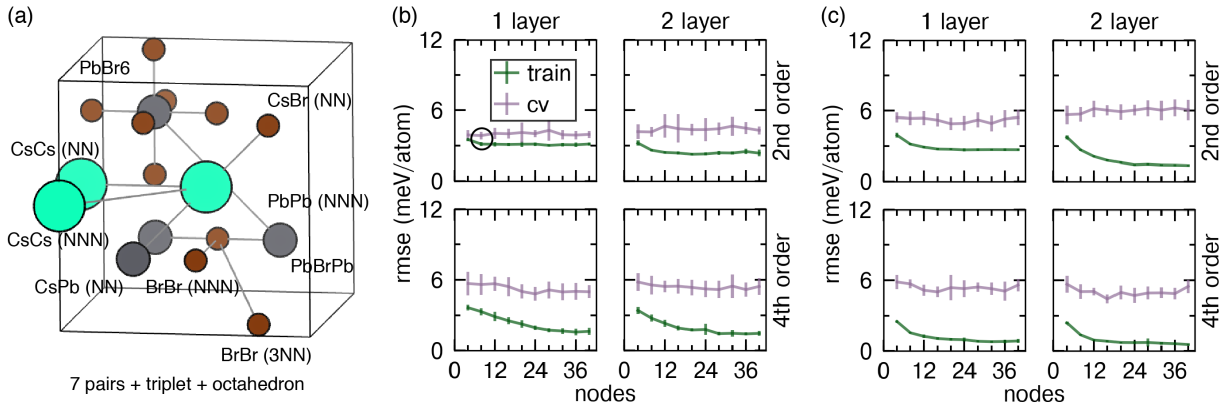

FIG. 7. (a) Clusters used in final model which includes 7 pairs, 1 triplet, and 1 octahedron. Results of 10-fold cross validation for (b) cluster-based model and (c) site-based model. Training and validation (cv) average RMSE is plotted with error bars of 1 standard deviation. In (b,c), left columns indicate 1 hidden layer while right columns indicate 2 hidden layers and top rows indicate 2nd order models while the bottom row indicates 4th order models.

ing k-point grids. Systematic strain enumerations were also included on the primitive perovskite structure, and the irreducible wedge of each subspace was sampled in the volume 1 cell. In addition to systematic enumerations, stochastic sampling of strains and displacements were made to generate more configurations. The strain and displacement fields were chosen at random from an n-sphere, and the correlations were compared to existing configurations to ensure uniqueness, i.e. that a very similar structure wasn't already included in the database. Also interpolations between structures were used for example between the three experimentally observed phases. In total, 31,000 configurations were calculated.

C. Hyperparameter tuning and model training

Training high-quality neural network models requires the selection of optimal model hyperparameters specifying the network architecture (i.e., number and connectivity of nodes) and number of input descriptors. We used k-fold cross validation to determine which set of hyperparameters best generalize to holdout sets of model validation data. This process is similar to model selection approaches in alloy cluster expansions, where a set of cluster basis functions are chosen to minimize a cross-validation metric. Due to the large number of input descriptors in an anharmonic cluster expansion, we restrict ourselves to five sets of clusters: (1) 4 pairs + 1 octahedron, (2) 5 pairs + 1 octahedron, (3) 8 pairs + 1 octahedron, (4) 4 pairs + 1 triplet + 1 octahedron, (5) 7 pairs + 1 triplet + 1 octahedron. The covalent bonding within the octahedra of $CsPbBr_3$ motivated the inclusion of an octahedral cluster. For each cluster in the model, we tested two groups of cluster basis functions to serve as input features: one included all cluster basis functions of the CCDs up to 2^{nd} order and another included all basis functions up to 4^{th} order. Additionally, we tested several network architectures by varying the number of hidden layers and the number of nodes per hidden layer, resulting in 400 unique hyperparameter sets.

Given a hyperparameter set, training the weights and biases in a neural net requires an optimization scheme[37]. We employed a batch training strategy with batch sizes of 2, 10, 100 and 1000 with at least 1000 training epochs per batch size. The Adam optimizer was used to update model weights and biases such that the least squares error of Eq. 18 was minimized.

Validation and training sets were used to find ANN hyperparameters that resulted in the most generalizable models with the smallest error. The total data set was split into a training set (80% of data), and a test set (20% of data). The test set was kept removed from any training iterations such that it remained an unbiased evaluator of model performance. K-fold cross validation with 10 folds was used to find the optimal hyperparameters (number of nodes, layers, and input features in the ANN model). A model was trained on 90% of the training dataset and a cross-validation error (CV) was evaluated on the remaining 10%. This procedure is repeated 10 times leaving a different fold of the training dataset out each time.

D. Optimal Hyperparameters

Figure 7 displays the training results for the best performing set of hyperparameters. This set consists of basis functions generated from 7 pairs, 1 triplet, and 1 octahedral cluster as pictured in Figure 7(a). Four other combinations of clusters were tested, but it was found that including more clusters, and especially including the triplet cluster resulted in more robust models. The neural net training results are displayed in Figures 7 for the cluster-based model (Figures 7 (b)) and the site-based model (Figures 7 (c)). For each model, input features up to order 2 or order 4 basis functions were tested [rows of Fig-

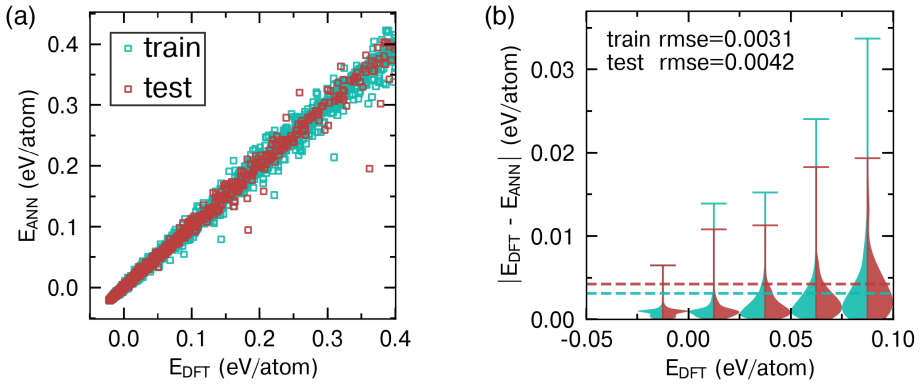

FIG. 8. Fitting statistics for 1 layer 2nd order cluster-based model with 8 hidden nodes per layer. (a) ANN energy vs DFT energy shows that both training and test set show similar average error. (b) Distribution of errors for lowest 125 meV configurations binned into 25 meV bins. The red and green dashed lines indicate the RMSE over the entire test and training set resepectively. Low energy configurations show very low error.

ures 7(b,c)] as well as number of hidden layers [columns of Figures 7(b,c)].

The site-based and cluster-based models perform similarly with several key differences. First, the cluster-based models generalize better to the validation set with smaller validation errors among all tested models. However, the site-based models achieve smaller errors on the training folds. Large differences between the training error and the validation error indicate that the models tend to overfit the training data and generalize poorly. The order 2 cluster-based model with 1 hidden layer performed the best in terms of generalizability with both the smallest validation error and the smallest difference between the training and validation errors. In particular the model with order 2 cluster-based model with 1 hidden layer and 8 nodes per hidden layer had the smallest validation error among all models and was therefore chosen as the best model according to the cross validation scheme.

E. ANN Fit Evaluation

After finding the optimal hyperparameters for our model (shown with the black circle on Figure 7(b) indicating the order 2 cluster-based model with 1 hidden layer of 8 nodes), we retrained the model on the full training set and calculated the error on the holdout set as shown in Figures 8(a,b). The training and test rmse were similar to those found in the hyperparameter tuning as expected. Additionally, we investigated the distribution of errors for different energy regions as shown in Figures 8(b). Interestingly, the model performs best for the lowest energy configurations, meaning that it faithfully reproduces the important ground state structures.

Figure 9 shows how the model reproduces the DFT potential energy surface along important paths in the space of atomic deformations. Figure 9 (a) shows the energy as a function of a linear interpolation between the cubic (α),

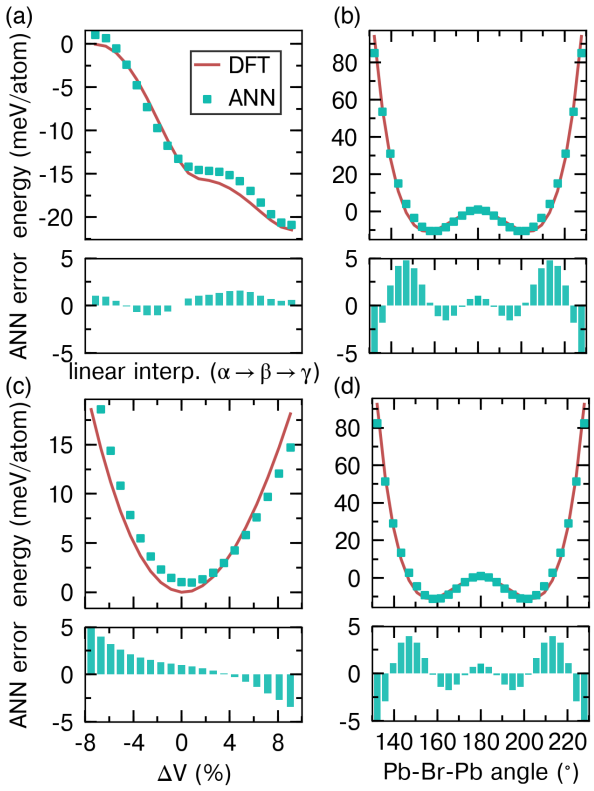

FIG. 9. Model and DFT energies as a function of (a) a linear interpolation experimentally observed phases, (b) in-phase tilts, (c) volume, and (d) anti-phase tilts. In all cases, the ANN PES aligns well with the DFT energy surface.

tetragonal (β) and orthorhombic (γ) phases of perovskite $CsPbBr_3$. Also shown is the energy of the crystal as a function of (b) in-phase octahedral tilt-mode amplitude, applied to the ideal cubic structure, (c) volumetric strain deformations of the crystal lattice, and (d) anti-phase octahedral tilt amplitude, applied to the ideal cubic struc-

ture. In all cases, the model predictions align well with the DFT energy surface. Furthermore, the model PES tends to be relatively smooth. The results for this simple model indicate that neural networks can reliably reproduce the potential energy surface of complex compounds such as $CsPbBr_3$, especially in the region of low energy configurations.

IV. DISCUSSION

A large variety of compounds adopt phases at high temperature that have symmetries coinciding with a saddle point on a zero Kelvin potential energy surface (PES). Crystal symmetries corresponding to a saddle point of the PES are dynamically unstable at low temperature but can become stable at high temperature through large anharmonic vibrational excitations. Phonon theories based on the harmonic approximation are unable to describe the high temperature thermodynamic properties of anharmonically stabilized phases. Instead, Monte Carlo or molecular dynamics simulations must be used to numerically perform thermodynamic averages over vibrational microstates sampled at high temperature. Direct simulation approaches, however, require a model of the energy of the crystal as a function of the atomic displacement and lattice strain degrees of freedom.

Mapping the atomic coordinates of a solid to reproduce a first-principles energy landscape is a challenging, high-dimensional supervised learning problem that requires careful consideration of many aspects of the machine learning pipeline, including feature engineering, model training, and model selection. High quality input features are an essential ingredient of any machine-learned model. In this study we have introduced collective cluster deformation (CCD) variables that uniquely describe the deformation of a particular cluster within the crystal relative to its geometry in a high symmetry reference state. Because the CCDs are symmetry-adapted functions of the set of all pair distances within the cluster, any model that is a function of these variables is inherently invariant to rigid-body rotation or translation of the crystal. Although they are defined relative to an undeformed high-symmetry crystal, the CCDs are not themselves invariant to the symmetry of this reference crystal. As such, the feature vector forming the input layer of the neural net is constructed from symmetry-invariant polynomials of the CCD variables, thus ensuring that the learned model is invariant to these additional crystal symmetries. Taken together, these properties specify features that are particularly well suited to machine-learning PES models for compounds that undergo group/subgroup structural transformations, in which the high symmetry phase is often stabilized at high temperature by large, anharmonic vibrational excitations. Moreover, the CCDs are themselves useful descriptors of local structure that have potential applications in high-throughput crystallographic data-mining frameworks, such as recently described workflows for characterizing local coordination environments[44].

The model selection methodology described here showed that low order basis functions tended to result in more generalizable models, with low error on holdout test sets. Higher order descriptor functions, as well as deeper (more layers) and wider (more nodes) neural nets, tended to overfit the training data resulting in poorly generalized models. The low training error of the more complex models indicates that the descriptors provide adequate information for models to learn the DFT energy surface, and motivate further studies focusing on reducing overfitting using techniques such as dropout or weight decay.

Two approaches were introduced in this study, one based on a *cluster*-centric neural net architecture and the other based on *site*-centric architecture. The cluster-based models are direct generalizations of previous anharmonic vibrational cluster expansion models as introduced by Thomas and Van der Ven. [8] The current work extends the linear models in [8] by allowing the functional form of the cluster energy to be learned by the machine learning model. The site based model using cluster basis functions is an extension to vibrational energy of the site-based neural-net approach introduced by Natarajan and Van der Ven for modeling configurational energy[45]. One benefit of the site-based model is that it allows interaction terms between basis functions from different clusters, which may explain the lower error achieved by the site-based model. The site-based model has some similarities to other descriptor-based machine learning approaches where descriptors are written in terms of exponentials of pair distances and bond angles[15, 46]. A key difference of the approach introduced here, however, is that it is specifically designed to represent the energy of a crystal relative to a high symmetry reference crystal, making it especially suited for studies of group/subgroup structural transitions and the thermodynamics of anharmonically stabilized phases. The reliance on descriptors that are invariant to the symmetries of the high symmetry reference phase ensures that symmetries are automatically satisfied. However, relative to more generic approaches, CCD descriptors that are measured relative to a high-symmetry reference crystal have much higher bias (in an information-theoretical context), so that significantly more information about the crystal deformation state can be encoded by fewer descriptors.

The models presented here differ from other recently described machine-learned potential models [15, 18–22, 24] in that they are built from neural networks and features that encode the connectivity of the reference crystal. For this reason, the CCD-based approach yields a very accurate anharmonic potential energy surface utilizing significantly fewer features and ANN weights than methods that are not referenced to a high-symmetry crystal structure. The CCD models presented here are well-suited to describing large but connectivity-preserving deformations of the reference crystal, such as occur dur-

ing the reversible phase transformations observed in $CsPbBr_3$ and other halide perovskites; however, the fact that the crystal connectivity is "built-in" to their functional form prevents these models from describing events that change the connectivity of the crystal, such as decohesion, mass transport, or plastic deformation. This distinction in transferability and complexity between the two approaches is a classic manifestation of the bias–variance tradeoff that is a universal challenge in developing predictive models[47].

Both the site-based and cluster-based models presented here were trained on a dataset of DFT-calculated deformation energies, yielding models that reproduce the energetics of crystal deformation with high accuracy. These models are thus well-adapted to simulation frameworks, such as traditional Monte Carlo methods, that predict thermodynamic equations of state by directly probing the energy density of states. However, models trained on energy data alone may underperform in predicting stresses and/or atomic forces, and are therefore less well-suited for use in simulation frameworks, such as molecular dynamics, that evolve the equations of motion of the crystal should incorporate force data in their training set.

Although the CCD-based models predict deformation energies with low validation error, the interaction range of the models is quite short, extending only to the third-nearest neighbor. This suggests that interatomic interactions in $CsPbBr_3$ are largely local in nature. Nevertheless, the CCD-based models cannot directly account for long-range coulomb effects, which can alter the energy of long-wavelength deformation modes near Γ. While the magnitude of this and other truncation effects would require a detailed comparison of model-derived phonon dispersion to the first-principles phonon dispersion, any discrepancies could be addressed either by adding CCDs for longer-range pair clusters to the feature vector or by correcting for dipole-dipole interactions, which follow well-known functional forms[48]. However, because long-range interactions influence the phonon dispersion in only a small neighborhood of the Brillouin zone near Γ, the differences in the phonon density of states (and, consequently, the vibrational free energies) will nevertheless be quite small.

V. CONCLUSIONS

The development of anharmonic vibrational hamiltonians is a challenging problem, however, by making use of machine learning techniques it is possible to capture a high degree of complexity that is present in the DFT energy landscape. We have presented a framework that utilizes neural-network models to reproduce the DFT energy landscape with high accuracy in the vicinity of a high-symmetry reference crystal. To construct features for the neural-network model we introduced collective-cluster-deformation variables, which are descriptive and easy-to-calculate functions of local geometry that are invariant to rigid-body transformations. The use of machine learning models is appealing because it removes much of the manual selection of terms in a Hamiltonian. Instead, the functional forms are learned through the training process. However, machine learning models, especially non-linear neural networks have a tendency to overfit the training data, and, therefore, hyperparameter tuning must be carefully considered. The next step in the progression of machine learning Hamiltonians is their use in finite temperature thermodynamics simulations which is a natural extension of the work presented here.

VI. ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation, Grant No. OAC-1642433. Computational resources provided by the National Energy Research Scientific Computing Center (NERSC), supported by the Office of Science and U.S. Department of Energy, under Contract DE-AC02-05CH11231, are gratefully acknowledged in addition to support from the Center for Scientific Computing from the CNSI, MRL: an NSF MRSEC (DMR-1720256).

Appendix A: Identifying deformation coordinates for dimensionality reduction

A n_α-atom non-planar cluster has $3n_\alpha - 6$ deformational degrees of freedom in three dimensions (after removal of rigid translation and rotation) and has $N_\alpha = n_\alpha(n_\alpha - 1)/2$ pair distances. For $n_\alpha > 4$, the number of pair distances exceeds the number of cluster degrees of freedom, such that any realizable deformation vector, $\vec{F}^\alpha$, must be confined to a $3n_\alpha - 6$-dimensional surface. In the vicinity of the undeformed cluster, coordinates on this cluster deformation surface can be projected uniquely into a $3n_\alpha - 6$-dimensional subspace, and a point in the subspace can be described by a truncated $3n_\alpha - 6$-element vector of optimized CCDs, which we denote $\vec{Q}^{\star\alpha}$.

A simple linear approximation of the cluster deformation surface can be computed from the matrix image of the Jacobian $\mathbf{J}_{\vec{F}}(\vec{R}^\alpha)$. However, a more robust set of linearized coordinates can be obtained by accounting for the fact that the pairwise deformation metrics are correlated for small deformations of the cluster. We define a correlation matrix $\mathbf{G}^{(F)}$ whose elements are the overlap, or similarity, between the deformation metrics of two pairs within the cluster. The elements of $\mathbf{G}^{(F)}$ are computed as inner products over the space of functions of the cluster coordinates, such that

$$G^{(F)}_{m,n} = \langle f_m, f_n \rangle = \int_{\vec{R}^\alpha} \mathrm{d}^{3N_\alpha}\vec{R}^\alpha \left[p(\vec{R}^\alpha) f_m\left(\vec{R}^\alpha\right) f_n\left(\vec{R}^\alpha\right)\right], \tag{A1}$$

where $p(\vec{R}^\alpha)$ is a probability density over all possible geometries of cluster α, and the integral is taken over the entire configuration space of $\vec{R}^\alpha$. A simple choice of $p(\vec{R}^\alpha)$ is a $3n_\alpha$-dimensional multivariate normal distribution, centered at the coordinates of the undeformed reference cluster and having an isotropic variance σ^2. This definition allows an analytic expression for Eq. (A1) for many choices of deformation metric. Physically motivated choices of the standard deviation σ are in the range of 10-25% of the nearest-neighbor pair distance for the crystal under consideration.

The correlation matrix $\mathbf{G}^{(F)}$ can be used to identify an optimized coordinate transformation from $\vec{F}^\alpha$ to $\vec{Q}^{\star\alpha}$. For a given change of basis $\vec{Q}^\alpha = \mathbf{U}\,\vec{F}^\alpha$, the corresponding transformation that takes $\mathbf{G}^{(F)}$ to $\mathbf{G}^{(Q)}$ is

$$\mathbf{G}^{(Q)} = \mathbf{U}^{-\top}\,\mathbf{G}^{(F)}\,\mathbf{U}^{-1}. \tag{A2}$$

The elements of $\mathbf{G}^{(Q)}$ measure the correlation between individual components of $\vec{Q}^\alpha$ over $p(\vec{R}^\alpha)$. If the transformation $\mathbf{U}$ is chosen appropriately, the correlation matrix $\mathbf{G}^{(Q)}$ will be the identity matrix, meaning that individual CCDs have unit variance and are uncorrelated over $p(\vec{R}^\alpha)$. This occurs when

$$\mathbf{U} = {\mathbf{G}^{(F)}}^{1/2} = \mathbf{\Lambda}^{1/2}\,\mathbf{V}^\top, \tag{A3}$$

where $\mathbf{\Lambda}$ is a diagonal matrix of the eigenvalues of $\mathbf{G}^{(F)}$ and $\mathbf{V}$ is an orthogonal matrix of the eigenvectors of $\mathbf{G}^{(F)}$. Each row i of $\mathbf{U}$ corresponds to a linear combination of the components of $\vec{F}^\alpha$ that yield a particular CCD value. The $\mathbf{U}$ transformation matrices are provided in the supplemental information for both a 4-point tetrahedron and 6-point octahedron cluster having maximal symmetry[31].

The transformation matrix $\mathbf{U}$, as defined in Eq. (A3) is optimal in several ways. First, the original correlation matrix $\mathbf{G}^{(F)}$ is invariant to symmetry in the sense that

$$\hat{c}\left[\mathbf{G}^{(F)}\right] = \mathbf{M}^{(F)}(\hat{c})\,\mathbf{G}^{(F)}\,\mathbf{M}^{(F)\top}(\hat{c}) = \mathbf{G}^{(F)}, \tag{A4}$$

where $\hat{c}$ is an operation in the point group of cluster α. This invariance relation is due to the fact that symmetrically-equivalent pairs of pair-deformation metrics have identical correlation. It is well known that if a symmetric matrix, such as $\mathbf{G}^{(F)}$, is invariant to a group representation (e.g, $\mathbf{M}^{(F)}(\hat{c})$), then any transformation that diagonalizes $\mathbf{G}^{(F)}$ also block-diagonalizes the symmetry matrices $\mathbf{M}^{(F)}(\hat{c})$. This means that the resulting CCD vector space, Q, is naturally separable into invariant subspaces, and that, under ideal circumstances, the invariant subspaces of Q will correspond to irreducible representations. To this end, eigenvectors of $\mathbf{G}^{(F)}$ having the same eigenvalue correspond to CCDs that are within the same invariant subspace. Moreover, if the eigenvalues of $\mathbf{G}^{(F)}$ are ordered from largest to smallest, there are spectral gaps between irreducible subspaces, with the largest gap occurring between the first $3n_\alpha - 6$ eigenvalues of $\mathbf{G}^{(F)}$ and the remaining eigenvalues. This gap occurs due to an underlying difference in behavior between directions with respect to the cluster deformation surface in the vicinity of the reference cluster. The first $3n_\alpha - 6$ directions approximately follow the cluster deformation surface, and so they have much larger variance than the remaining directions, which are nearly orthogonal to the cluster deformation surface. The truncated CCD vector, $\vec{Q}^{\star\alpha}$, then corresponds to the first $3n_\alpha - 6$ elements of the CCD vector obtained from the transformation matrix in Eq. A3.

[1] K. Persson, M. Ekman, and V. Ozoliņš, Phys. Rev. B **61**, 11221 (2000).
[2] P. Souvatzis, O. Eriksson, M. I. Katsnelson, and S. P. Rudin, Phys. Rev. Lett. **100**, 095901 (2008).
[3] G. Grimvall, B. Magyari-Köpe, V. Ozoliņš, and K. A. Persson, Rev. Mod. Phys. **84**, 945 (2012).
[4] K. Parlinski, Z. Q. Li, and Y. Kawazoe, Phys. Rev. Lett. **78**, 4063 (1997).
[5] S. Fabris, A. T. Paxton, and M. W. Finnis, Phys. Rev. B **63**, 094101 (2001).
[6] C. Carbogno, C. G. Levi, C. G. Van de Walle, and M. Scheffler, Phys. Rev. B **90**, 144109 (2014).
[7] J. Bhattacharya and A. Van der Ven, Acta Materialia **56**, 4226 (2008).
[8] J. C. Thomas and A. Van der Ven, Phys. Rev. B **88**, 214111 (2013).
[9] J. S. Bechtel, R. Seshadri, and A. Van der Ven, The Journal of Physical Chemistry C **120**, 12403 (2016), https://doi.org/10.1021/acs.jpcc.6b03570.
[10] R. X. Yang, J. M. Skelton, E. L. da Silva, J. M. Frost, and A. Walsh, The Journal of Physical Chemistry Letters **8**, 4720 (2017), pMID: 28903562, https://doi.org/10.1021/acs.jpclett.7b02423.
[11] A. Marronnier, H. Lee, B. Geffroy, J. Even, Y. Bonnassieux, and G. Roma, The Journal of Physical Chemistry Letters **8**, 2659 (2017), pMID: 28553717, https://doi.org/10.1021/acs.jpclett.7b00807.
[12] E. Cockayne, E. L. Shirley, B. Ravel, and J. C. Woicik, Phys. Rev. B **98**, 014111 (2018).
[13] D. Vanderbilt and W. Zhong, Ferroelectrics **206**, 181 (1998), https://doi.org/10.1080/00150199808009158.
[14] K. M. Rabe and U. V. Waghmare, Phys. Rev. B **52**, 13236 (1995), arXiv:9411006 [mtrl-th].
[15] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[16] J. C. Wojdeł, P. Hermet, M. P. Ljungberg, P. Ghosez, and J. Íñiguez, Journal of Physics: Condensed Matter **25**, 305401 (2013).
[17] F. Zhou, W. Nielson, Y. Xia, and V. Ozoliņš, Phys. Rev. Lett. **113**, 185501 (2014).
[18] N. Artrith and A. Urban, Comput. Mater. Sci. **114**, 135 (2016).

[19] N. Artrith, A. Urban, and G. Ceder, Phys. Rev. B **96**, 014112 (2017).

[20] A. Glielmo, P. Sollich, and A. De Vita, Phys. Rev. B **95**, 214302 (2017).

[21] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, Science Advances **3** (2017), 10.1126/sciadv.1701816, https://advances.sciencemag.org/content/3/12/e1701816.full.pdf.

[22] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, The Journal of Physical Chemistry C **121**, 511 (2017), https://doi.org/10.1021/acs.jpcc.6b10908.

[23] J. C. Thomas, J. S. Bechtel, and A. Van der Ven, Phys. Rev. B **98**, 094105 (2018).

[24] X.-G. Li, C. Hu, C. Chen, Z. Deng, J. Luo, and S. P. Ong, Phys. Rev. B **98**, 094104 (2018).

[25] M. Rodová, J. Brožek, K. Knížek, and K. Nitsch, Journal of Thermal Analysis and Calorimetry **71**, 667 (2003).

[26] T. F. Havel, I. D. Kuntz, and G. M. Crippen, Bulletin of Mathematical Biology **45**, 665 (1983).

[27] Physically, the atoms must be chemically identical and are thus indistinguishable, but we may assign each atom and position a distinguishing label in order to analyze the effect of symmetry.

[28] S. Dresselhaus, G. Dresselhaus, and A. Jorio, *Group Theory: Application to the Physics of Condensed Matter*, SpringerLink: Springer e-Books (Springer, 2008).

[29] J. C. Thomas and A. Van der Ven, Journal of the Mechanics and Physics of Solids **107**, 76 (2017).

[30] J. C. Thomas and A. Van der Ven, Phys. Rev. B **96**, 134121 (2017).

[31] See Supplemental Material at [URL will be inserted by publisher].,.

[32] Because the vector $\vec{Q}^{\alpha}$ is invariant to rigid rotation and translation of the cluster, the Jacobian is rank deficient and cannot be inverted in the conventional sense. We instead utilize the Moore-Penrose inverse, which ensures that the resulting basis vectors are orthogonal to the generators of rigid translation and rigid rotations of the cluster.

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

[34] T. Mueller and G. Ceder, Phys. Rev. B **80**, 024103 (2009).

[35] L. J. Nelson, G. L. W. Hart, F. Zhou, and V. Ozoliņš, Phys. Rev. B **87**, 035125 (2013).

[36] A. Van der Ven, J. Thomas, B. Puchala, and A. Natarajan, Annual Review of Materials Research **48**, 27 (2018), https://doi.org/10.1146/annurev-matsci-070317-124443.

[37] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (The MIT Press, 2016).

[38] D. P. Kingma and J. Ba, arXiv preprint arXiv:1412.6980 (2014).

[39] J. S. Bechtel and A. Van der Ven, Phys. Rev. Materials **2**, 025401 (2018).

[40] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).

[41] G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999).

[42] P. E. Blöchl, Phys. Rev. B. Condens. Matter **50**, 17953 (1994), arXiv:arXiv:1408.4701v2.

[43] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, Phys. Rev. Lett. **100**, 136406 (2008).

[44] D. Waroquiers, X. Gonze, G.-M. Rignanese, C. Welker-Nieuwoudt, F. Rosowski, M. Göbel, S. Schenk, P. Degelmann, R. André, R. Glaum, and G. Hautier, Chemistry of Materials **29**, 8346 (2017), https://doi.org/10.1021/acs.chemmater.7b02766.

[45] A. R. Natarajan and A. Van der Ven, npj Computational Materials **4**, 56 (2018).

[46] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).

[47] C. Sammut and G. I. Webb, *Encyclopedia of machine learning* (Springer Science & Business Media, 2011).

[48] N. W. Ashcroft and N. D. Mermin, *Solid State Physics* (Saunders College Publishing, Orlando, 1976).