PHYSICAL REVIEW MATERIALS **4**, 063802 (2020)

Machine learning surrogate models for prediction of point defect vibrational entropy

Clovis Lapointe,[1,*] Thomas D. Swinburne,[2,†] Louis Thiry,[3] Stéphane Mallat,[4,5] Laurent Proville,[1]
Charlotte S. Becquart,[6] and Mihai-Cosmin Marinica[1,‡]
[1]*DEN-Service de Recherches de Métallurgie Physique, CEA Saclay, 91191 Gif-sur-Yvette, France*
[2]*Aix-Marseille Université, CNRS, CINaM UMR 7325, Campus de Luminy, 13288 Marseille, France*
[3]*Département d'informatique, ENS, CNRS, PSL University, Paris, France*
[4]*Collège de France, ENS, PSL University, Paris, France*
[5]*Flatiron Institute, New York, New York 10010, USA*
[6]*Université Lille, CNRS, INRA, ENSCL, UMR 8207 - UMET - Unité Matériaux et Transformations, F-59000 Lille, France*

(Received 3 February 2020; revised manuscript received 5 May 2020; accepted 15 May 2020;
published 15 June 2020)

The temperature variation of the defect densities in a crystal depends on vibrational entropy. This contribution to the system thermodynamics remains computationally challenging as it requires a diagonalization of the system's Hessian which scales as $O(N^3)$ for a crystal made of N atoms. Here, to circumvent such a heavy computational task and make it feasible even for systems containing millions of atoms, the harmonic vibrational entropy of point defects is estimated directly from the relaxed atomic positions through a linear-in-descriptor machine learning approach of order $O(N)$. With a size-independent descriptor dimension and fixed model parameters, an excellent predictive power is demonstrated on a wide range of defect configurations, supercell sizes, and external deformations well outside the training database. In particular, formation entropies in a range of $250k_B$ are predicted with less than $1.6k_B$ error from a training database whose formation entropies span only $25k_B$ (training error less than $1.0k_B$). This exceptional transferability is found to hold even when the training is limited to a low-energy superbasin in the phase space while the tests are performed for a different liquid-like superbasin at higher energies.

DOI: 10.1103/PhysRevMaterials.4.063802

I. INTRODUCTION

The aging of crystalline materials is heavily influenced by the thermodynamic and kinetic properties of point defects. Their evolution gives rise to an extraordinarily diverse range of defect morphologies [1–8] whose distributions in size, character, and density exhibit significant variations with temperature.

The stability of defect populations changes in response to temperature variation [9–11] according to the system entropy in which one distinguishes three distinct contributions associated with (i) various geometry configurations [12], (ii) electronic thermal excitations [13], and (iii) lattice thermal vibrations [14]. For an isolated vacancy close to melting temperature, both electronic and vibrational entropies have the same order of magnitude [13] around $3k_B/2$ while configurational entropy reduces to the mixing entropy and thus is negligible in dilute systems [14]. Below the melting temperature, the electronic entropy decreases linearly in temperature as the width of the Fermi surface sharpens. The vibrational contribution becomes thus dominant up to a few kelvins where quantum effects yields an abrupt decrease, similar to the phonon heat capacity. For more complex defects, the configurational entropy is augmented by a term $k_B \ln(\mathcal{N}_c)$, where $\mathcal{N}_c$ is the number of different geometries corresponding to the same internal energy. Since it does not vary with temperature, this contribution does not affect the stability of defect structures. We shall also leave aside the temperature dependence of the internal energy which is inherent to the thermal expansion of solids [14]. Our study is devoted to the computation of the vibrational entropy as it represents surely [15] an important contribution to the stability of defects in a wide range of temperature.

For a solid containing N atoms, the standard harmonic approximation of entropy [16] requires an $O(N^2)$ calculation of the dynamical matrix and an $O(N^3)$ diagonalization to find the vibrational spectrum. The procedure is schematically represented in Fig. 1(a). For instance, the computational load for such a task in a crystal made of 2×10^5 atoms requires more than 20 TB of memory and 10 hours over thousands of the most recent CPUs. Different methods have been developed [17–25] to compute directly the free energy of defects including the nonharmonic contributions from energy and entropy in an indistinguishable manner. However these methods remain computationally very heavy as they usually rely on sampling the system phase space through the construction of random or optimized trajectories. The essential problem arises from the convergence of such methods, as to achieve a reliable sampling the number of iterations needed scales as $O(N^2)$ or $O(N^3)$, in the more favorable case. Furthermore we notice that according to neutron scattering experiments, the nonharmonic

---

[*] clovis.lapointe@cea.fr
[†] swinburne@cinam.univ-mrs.fr
[‡] mihai-cosmin.marinica@cea.fr

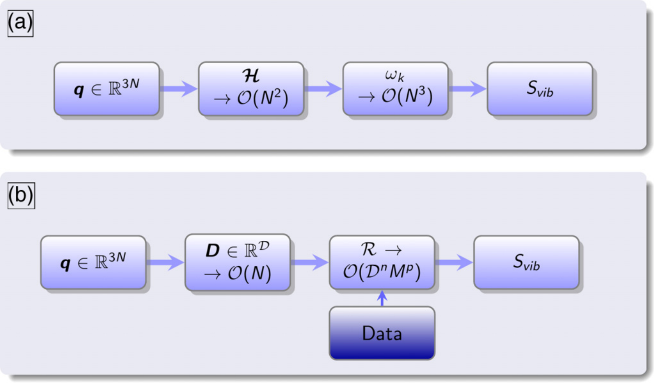

FIG. 1. Two strategies to evaluate the harmonic formation entropy of defects embedded into crystalline structure having N atoms: (a) the traditional approach based on the phonon spectrum of Hessian $\boldsymbol{\mathcal{H}}$ (the second derivatives of the potential energy of the system); (b) the machine learning surrogate model based on the M instances of the database, fitted via regression in the descriptor space $\mathbb{R}^{\mathcal{D}}$. The descriptor or feature space is the representation of the atomic configuration through the descriptor functions. Both scaling coefficients of the regression n and p range between 0 and 2 depending on the method used. For the linear fit, used in the present study, $n = 1$ and $p = 0$.

contributions to the formation of defects are not essential to the computation of vibrational entropy in a broad range of temperature, i.e., up to 700 K for example in α-Fe [26] and Al [27].

In the present study we thus propose a surrogate model to evaluate the harmonic vibrational entropy using a linear-in-descriptor machine learning (LDML) approach with $O(N)$ computational cost [28–37]. The method is applied to a wide class of point defects using only the relaxed atomic positions to determine directly the vibrational entropy. We chose to exemplify our computational technique using empirical potential interactions in α-Fe. The accuracy of interatomic potentials is currently undergoing a renaissance due to ever larger databases and new potential formalisms employing machine learning techniques [29,33–36,38–43], statistical on-the-fly learning [44,45], and mixed elastic-atomic models [3,46], among others. The empirical interatomic potential models employed here were fitted on *ab initio* [47] or experimental data [48]. Their relative simplicity allowed us to rapidly assess a wide range of defect structures and to explore a large data set in large crystals inaccessible via standard *ab initio* methods [49–51].

Our main finding is that the LDML approach we propose exhibits an exceptional degree of transferability, giving the ability to rapidly assess defect vibrational entropy at realistic temperatures in different systems of body-centered cubic (bcc) Fe containing defects. The same machine learning parametrization allows us to predict the formation entropies of different defects over a wide range of $250k_B$ to within a root mean square error (RMSE) inferior to $2k_B$, despite the rather narrow training set having a total formation entropy range of only $25k_B$.

In Sec. II we define the harmonic vibrational entropy and emphasize the equivalence between the local and the eigen descriptions for phonons in the harmonic approximation. In Sec. III we describe the machine learning approach for vibrational entropy, by introducing the LDML model and the descriptor functions. In Sec. IV the data set production and model training are detailed before being applied to predict the formation entropy of various defects in Sec. V. The excellent transferability found in initial applications to point defects is pushed further in Sec. V D, where our model trained on defect structures is applied to predict the formation entropy of highly defected structures generated by random displacements in bcc iron supercells. The success of the present approach opens many perspectives for high-throughput, multiscale materials science calculations which are discussed in Sec. VI.

II. VIBRATIONAL ENTROPY IN THE HARMONIC APPROXIMATION

To evaluate vibrational properties under the harmonic approximation we consider the normal modes of a system with N atoms, obtained from the spectrum of the force constant matrix $\boldsymbol{\phi} \in \mathbb{R}^{3N\times 3N}$ through

$$\left(\boldsymbol{\phi} - \boldsymbol{M}\omega_\nu^2\right)|\nu\rangle = |0\rangle, \tag{1}$$

where $\boldsymbol{M} \in \mathbb{R}^{3N\times 3N}$ is a diagonal mass matrix. If we only consider phonons in the Debye approximation, appropriate for phonons near the center of the Brillouin zone, the force constant matrix becomes the Hessian operator $\boldsymbol{\mathcal{H}}$, the matrix of second derivatives of the potential energy U. As discussed above, to obtain the eigenvalues ω_ν^2 and eigenvectors $|\nu\rangle$ the Hessian must be diagonalized. In the classical approximation, i.e., when the temperature is larger than the crystal Debye temperature such as $\frac{\hbar\omega_\nu}{k_B T} \ll 1 \ \forall|\nu\rangle$, the entropy becomes [52,53]

$$S(T,N) = k_B \sum_\nu \left[\ln\left(\frac{k_B T}{\hbar\omega_\nu}\right) + 1\right], \tag{2}$$

where k_B and $\hbar$ are the Boltzmann and Planck constants, respectively. For finite crystalline systems containing N_b bulk atoms and $\pm N_d$ point defects, the vibrational formation entropy S_f is defined as

$$S_f(T,N_d) = S_d(T,N_b \pm N_d) - \frac{N_b \pm N_d}{N_b} S_b(T,N_b), \tag{3}$$

where the entropies S_b and S_d of the bulk and defective systems are computed at the same volume V. With Hessian eigenvalues $\omega_{\nu_b}^2$ and $\omega_{\nu_d}^2$ for the bulk and defect systems, Eq. (2) yields a harmonic formation entropy:

$$S_f(T,N_d) = k_B \ln\left(\frac{\prod_{\nu_b}\left(\hbar\omega_{\nu_b}\right)^{\frac{N_b \pm N_d}{N_b}}}{\prod_{\nu_d}\hbar\omega_{\nu_d}}\right). \tag{4}$$

Figure 1 schematically summarizes the numerical treatment required to compute entropy through the diagonalization of the Hessian matrix, which has an $O(N^3)$ computational demand that typically prohibits application to large systems. We have to note that an $O(N)$ approach has been developed by different authors [54], who treat the summation in Eq. (2) as an expectation value over the eigenvalue distribution. They approximated this distribution using a set of Chebyshev polynomials with a random basis set. While this stochastic approach is indeed more efficient than $O(N^3)$ treatment for large systems, a converged result requires a proper selection of a large set of polynomials and basis vectors, requiring a computational effort which is still impractically high for the high-throughput evaluation desired in the present work, motivating our use of LDML models.

A. Green's function formalism for vibrational entropy calculations

Within the harmonic approximation, evaluation of the vibrational entropy requires knowledge of the full phonon spectrum, which can be directly computed from the secular equation (1). The Green's function formalism is an alternative and elegant way to iteratively solve this same eigenproblem. Taking eigenmodes $|\nu\rangle$, the Green's function $\boldsymbol{\mathcal{G}} \in \mathbb{C}^{3\times 3} \otimes \mathbb{C}^{N\times N}$ is written [52]

$$\boldsymbol{\mathcal{G}}(\omega) = \sum_{\nu} \frac{|\nu\rangle \otimes \langle\nu|}{\omega_\nu^2 - \omega^2}. \tag{5}$$

The total phonon density of states is then the imaginary part of the trace of the Green's function [52],

$$\Omega(\omega) = \frac{2\omega}{\pi}\mathrm{Im}[\mathrm{Tr}\,\{\boldsymbol{\mathcal{G}}(\omega)\}], \tag{6}$$

where $\mathrm{Im}(\cdot)$ is the imaginary part and $\mathrm{Tr}(\cdot)$ is the trace operator. It is straightforward to verify that the total degrees of freedom are respected through $\int_0^\infty \Omega(\omega)d\omega = 3N$, while the classical vibrational entropy of the system at a given temperature T

$$S = -k_B \int_0^\infty \left[\ln\left(\frac{\hbar\omega}{k_B T}\right) - 1\right]\Omega(\omega)d\omega. \tag{7}$$

While equations (6) and (7) provide a clear connection between the phonon Green's function and the vibrational entropy, the normal modes $|\nu\rangle$ are typically delocalized across many atoms, complicating an analysis based on a mapping to localized atomic descriptors. As a result, we now emphasize this same formalism using a local basis set to give the local density of states [52].

B. Local basis for densities of states of phonons

The local density of states can be deduced directly in the Green's function formalism [52]. Our goal is to replace the delocalized basis $|\nu\rangle$ in the above results with a localized basis $|i\alpha\rangle$, where each basis vector is localized on a coordinate α of an atom i. As both bases $|\nu\rangle$ and $|i\alpha\rangle$ are complete, they are related by a rotation in $\mathbb{R}^{3N}$,

$$|\nu\rangle = \sum_i \sum_\alpha \xi^{i\alpha}(\nu)|i\alpha\rangle; \tag{8}$$

the square of the rotation matrix elements, $|\xi^{i\alpha}(\nu)|^2$, can be seen as the probability of the phonon $|\nu\rangle$ to be localized on the atom i and along the α direction. By direct substitution into Eq. (5) and using generic properties of rotation matrices we obtain

$$\varrho^{i\alpha}(\omega) = \frac{2\omega}{\pi}\mathrm{Im}[\mathcal{G}_{i\alpha}(\omega)], \tag{9}$$

$$\varrho^{i\alpha}(\omega) = \sum_\nu |\xi^{i\alpha}(\nu)|^2\delta(\omega - \omega_\nu), \tag{10}$$

where $\varrho^{i\alpha}$ is the local DOS of the phonon projected on the atom i following the α direction. The classical vibrational entropy of the system from Eq. (7) can then be written as the local entropy contribution of each atom:

$$\begin{aligned} S &= \sum_i \underbrace{\left\{-k_B \sum_\alpha \sum_\nu \left[\ln\left(\frac{\hbar\omega_\nu}{k_B T}\right) - 1\right]|\xi^{i\alpha}(\nu)|^2\right\}}_{S_i,\ \text{local information}} \\ &= \sum_i \left[\sum_\alpha s^{i\alpha}\right], \end{aligned} \tag{11}$$

where $s^{i\alpha}$ accounts for the local entropy from the ith atom in the α direction and $S_i = \sum_\alpha s^{i\alpha}$ represents the total contribution from the same atom. The above equation is a consequence also of the fact that the total density of states of phonons is the sum of the local contribution:

$$\Omega(\omega) = \sum_i \left[\sum_\alpha \varrho^{i\alpha}(\omega)\right]. \tag{12}$$

The redistribution of the total entropy of the system into local contribution, Eq. (11), is exact, as is the total density of states, Eq. (12). As with the local density of states, the local entropy is related to the local environment of the atom. It should be noted that local entropies gather the full spectrum of the dynamical matrix. The Green's formalism allows us to formulate the vibrational entropy problem as a linear problem of sources in term of local geometric environments. This formalism describes long-range interactions in term of source as linear elasticity describes long-range interactions in term of elastic dipoles.

III. MACHINE LEARNING SURROGATE MODEL FOR VIBRATIONAL ENTROPY

Prediction of the vibrational entropy S directly from the relaxed atomic coordinates is impractical due to the high dimension of the input space and the highly nonlinear regression required. In addition, physical constraints such as extensivity in N and V or symmetry under exchange of identical atoms are very hard to enforce.

In common with the majority of machine learning models [28,33,38,40,42], we instead replace a highly complex nonlinear regression task on atomic coordinates $\boldsymbol{q} \in \mathbb{R}^{3N}$ with a much simpler linear regression task on nonlinear *descriptor functions* of the atomic coordinates, which we dub the linear-in-descriptor machine learning (LDML) model. Often, the effective dimension of the descriptor space is larger than the dimension of the original input space [34–36,42,55], though in the present work our mapping will achieve a significant dimensional reduction. The precise choice of descriptor functions is presented in Sec. III A.

To build the LDML model, we first assume that the most general model input is a set of N evaluations of $\mathcal{D}$ descriptor functions, giving a descriptor vector $\underline{D}^i \in \mathbb{R}^{\mathcal{D}}$ for each atom i. To build $\underline{D}^i$, the $\mathcal{D}$ descriptor functions take as input the atomic environment around an atom i. The atomic environment around i can in principle be the entire system, a point we return to in the next section. This procedure thus initially maps an input space of $\mathbb{R}^{3N}$ to a descriptor space of $\mathbb{R}^{\mathcal{D}\times N}$. We then assume that these descriptor functions are sufficiently diverse and well chosen such that the local entropy for an atom i can be written as the linear relation

$$\left[\sum_{\alpha} s^{i\alpha}\right] = \underline{w}^i \cdot \underline{D}^i, \tag{13}$$

where $\underline{w}^i \in \mathbb{R}^{\mathcal{D}}$ is a vector of $\mathcal{D}$ weights. In principle, the total entropy $S = \sum_{i=1}^{N} [\sum_{\alpha} s^{i\alpha}]$ then requires the determination of $\mathcal{D}N$ parameters for the $\{\underline{w}^i\}$; however, while linearity in the $\underline{w}^i$ is sufficient to give a thermodynamically extensive entropy [56], to respect symmetry under identical exchange we further require that the weight vectors be identical among indistinguishable atoms, implying that $\underline{w}^i = \underline{w}$ for the elemental systems considered here; i.e., all weight vectors are identical. This gives a total entropy of

$$S = \sum_{i=1}^{N}\left[\sum_{\alpha} s^{i\alpha}\right] = \underline{w}\cdot\left(\sum_{i=1}^{N}\underline{D}^i\right) = N\underline{w}\cdot\langle\underline{D}\rangle, \tag{14}$$

where $\langle\underline{D}\rangle = N^{-1}\sum_{i=1}^{N}\underline{D}^i \in \mathbb{R}^{\mathcal{D}}$ is the average descriptor vector, meaning that we map the original input $\boldsymbol{q} \in \mathbb{R}^{3N}$ to a descriptor space of dimension $\mathcal{D} \ll 3N$, which is system size independent. The predicted entropy Eq. (14) is invariant under identical exchange. Furthermore if one considers n copies of the system then the average vector $\langle\underline{D}\rangle$ over the n copies is unchanged compared to the original system, thence proving thermodynamic extensivity. The fixed dimensionality of vector $\langle\underline{D}\rangle$ allows the dimension of the input space (i.e., number of atoms) to vary, which is essential to compute the LDML model formation entropy. Using Eq. (3) we find the formation entropy for a defective system containing $\pm N_d$ defects in a bulk lattice of N_b atoms:

$$S_f = (N_b \pm N_d)\,\underline{w}\cdot[\langle\underline{D}\rangle_d - \langle\underline{D}\rangle_b], \tag{15}$$

where $\langle\underline{D}\rangle_d$ and $\langle\underline{D}\rangle_b$ are the average descriptor vectors for the defect-containing and bulk systems, respectively. The formation entropy is therefore the inner product between the model weight vector $\underline{w}$ and the difference in the average descriptor vectors $\langle\underline{D}\rangle_d - \langle\underline{D}\rangle_b$, multiplied by the total number of atoms. Equation (15) is a central result of this paper, defining our LDML model. While many choices for machine-learning-based surrogate models exist, including the popular kernel methods and neural networks [29,55], the conceptually simpler approach followed here offers many advantages in transferability, overfitting control and analytic connection to thermodynamic properties. In the next section we consider candidate descriptor functions.

A. Choice of descriptor functions

We have seen that the input vector to our LDML model is the total descriptor vector $N\langle\underline{D}\rangle$, which is symmetric under identical exchange. However, the choice of descriptor functions must also preserve the symmetries and the invariances of the local atomic environment. The notion of descriptor in material science was introduced by Behler and Parrinello [28–30]. They proposed the $\mathbf{G}_2$ descriptor, defined below, that underlines the radial distribution of neighboring atoms weighted by a Gaussian. Since this pioneering work, many descriptors have been developed by (i) introducing the explicit angular description, as the $\mathbf{G}_3$ [28], (ii) using the spectral decomposition in 3D or 4D spherical functions of the atomic density [31,32], (iii) particular design for a given system [44,57–60], (iv) using even machine/deep learning methods in order to find the appropriate descriptors [61–64], and (v) hybrid descriptors that can mix all others classes mentioned above [33]. We note that there are also particular types of descriptors that take the fingerprint of the whole system, offering significant advantages when the observable targeted by the surrogate model cannot be described as a sum of local quantities. Within this particular formalism, the full atomic density is decomposed through a multiscale convoluted wavelet network, giving a vector of atom-delocalized scattering coefficients [34–37]. This method is particularly relevant for coarse-graining systems where several scales interact. The dimension of descriptors is flexible and is often used to control the level of the accuracy necessary to represent the local atomic environment in the descriptor space. There is therefore always a trade-off between computational efficiency, accuracy, and the sensitivity to overfitting which can arise for large input space dimensions. This work compares three local descriptors: the angular Fourier series (AFS) [32], the bispectrum SO(4) [bSO(4)] [31,32], and a scattering transform descriptor [34,35].

The AFS descriptor $\mathcal{A}_{n,l}$ combines the radial and angular information of the local atomic environment. The n and l components account for the radial and angular information of the neighborhood structure centered on the i atom; defining as $\mathcal{R}_i$ the set of indices for atoms less than r_{cut} from i, we have

$$\mathcal{A}^i_{n,l} = \sum_{k,k'\in\mathcal{R}_i} g_n(r_{ik})g_n(r_{ik'})\cos(l\theta_{ik,ik'})f_i(r_{ik})f_i(r_{ik'}),$$

where r_{ik} is the distance between atom i and atom k, and $\theta_{ik,ik'}$ is the angle formed by the triplet of atoms i, k, k' centered on i. The sum involves the pairs and the triplets of atoms formed by the central ith atom and the neighboring atoms inside the sphere with the radius r_{cut} around atom i. f is a cutoff function, which for the distances $r \geqslant r_{\mathrm{cut}}$ gives $f_i(r) \equiv 0$. The radial functions g_n are decreasing polynomials with the distance r having the degree of $\alpha + 2$ for $0 \leqslant \alpha \leqslant n$. The angular functions are the Tchebyshev polynomials [32] with $0 \leqslant l \leqslant l_{\max}$. As $\mathcal{A}_{n,l}$ is formed from a product of the radial and angular functions, the descriptor has a total of $n_{\max}(l_{\max}+1)$ components. The AFS descriptor enables wide-ranging levels of accuracy on radial and angular information by imposing $n_{\max}$ and $l_{\max}$, respectively. Otherwise stated, in this paper we have used $n_{\max} = 20$, and $l_{\max} = 10$, and the cutoff distance of 5 Å. The total number of components for the AFS descriptor used here therefore is 220.

The bSO(4) descriptor $\mathrm{bSO(4)}_{j_{\max}}$ is a spectral descriptor based on the decomposition of the atomic density in 4D hyperspherical harmonics [31,32]. The three components of the vector $\mathbf{r} \in \mathbb{R}^3$ can be recast into the three angles of the unit sphere $\mathcal{S}^4 \in \mathbb{R}^4$. The local environment of the ith atom is described as a density $\rho_i(\boldsymbol{r})$, and can be decomposed on the 4D hyperspherical harmonics basis:

$$\rho_i(\boldsymbol{r}) = \sum_{k \in \mathcal{R}_i} w_k \delta(\boldsymbol{r} - \boldsymbol{r_k}) \tag{16}$$

$$= \sum_{k \in \mathcal{R}_i} \sum_{j=0}^{\infty} \sum_{m,m'=-j}^{j} \boldsymbol{c}_{i,j}^{m,m'} U_j^{m,m'}, \tag{17}$$

where w_k is the species-dependent weight, and $\boldsymbol{c}_{i,j}^{m,m'}$ are the result of the scalar product between the density centered on atom i and the hyperspherical harmonic $U_j^{m,m'}$. From the above equation and the $\boldsymbol{c}_{i,j}^{m,m'}$ coefficients, the power and the bispectrum of the atomic density can be deduced. The bispectral components of bSO(4) are defined by the following equation, where $j \leqslant j_{\max}$ and $|j_1 - j_2| \leqslant j \leqslant j_1 + j_2$:

$$B^i_{j j_1 j_2} = \left(\boldsymbol{c}_{i,j}^{m,m'}\right)^{\dagger} \boldsymbol{H}^{j_1 j_2} \left(\boldsymbol{c}_{i,j_1}^{m_1,m_1'} \otimes \boldsymbol{c}_{i,j_2}^{m_2,m_2'}\right), \tag{18}$$

where $\boldsymbol{H}^{j_1 j_2}$ is related with the Clebsch-Gordan coefficient of the SO(4) group. A detailed description can be found in [31,32]. Following the analysis of the results of the first trial regressions, presented in Sec. V, in this study we use the $j_{\max} = 3.5$ and select only the diagonal components $j_1 = j_2$ [31,32,65], yielding as the total number of components 26; the cutoff distance is set to 5 Å.

The solid harmonic wavelet scattering transform [34,35] is a multiscale translation-rotation invariant descriptor. First a global density ρ is computed as a sum of Gaussian functions g centered at the atomic positions:

$$\rho(\boldsymbol{r}) = \sum_i g(\boldsymbol{r} - \boldsymbol{r}_i). \tag{19}$$

Scattering coefficients $S^{\mathcal{J},L}\rho[j,\ell]$, $j \in \mathcal{J}$, $0 \leqslant \ell \leqslant L$, are then computed with convolutions of this density ρ with solid harmonic wavelets $\psi^m_{j,\ell}$ of scale $j \in \mathcal{J}$, followed by an integral to have the rotation-translation invariance:

$$S^{\mathcal{J},L}\rho[j,\ell] = \int_{\mathbb{R}^3} \left[\sum_{m=-\ell}^{\ell} \left|\rho * \psi^m_{j,\ell}(\boldsymbol{r})\right|^2\right]^{1/2} d\boldsymbol{r}, \tag{20}$$

$$\psi^m_{j,\ell}(\boldsymbol{r}) = \frac{1}{(\sqrt{2\pi})^3} e^{-\frac{1}{2}|\frac{r}{2^j}|^2} \left|\frac{\boldsymbol{r}}{2^j}\right|^{\ell} Y^m_{\ell}\left(\frac{\boldsymbol{r}}{|\boldsymbol{r}|}\right). \tag{21}$$

In this paper, we have used $L = 9$ and 9 scales $\mathcal{J} = \{0, 0.25, 0.5, 0.75, 1, 2, 3, 4, 5\}$, yielding a descriptor of dimension 90.

IV. TRAINING LDML MODEL

A. Production of the configuration database

Any surrogate model is clearly heavily reliant on the database used for training. In this work, we used the ART method [66–69], following the methodology of previous studies [2], to generate a large number of configurations for small vacancy and interstitial clusters in bcc Fe. All clusters contained between 1 and 4 removed or additional atoms, which we label as V_n and I_n respectively, with $n = 1, 2, 3, 4$. Despite their apparent simplicity, the energy landscape of such defect configurations is known to have many thousands of binding configurations [2,4,11]. To test the sensitivity of our surrogate model to the underlying energy model, all calculations were performed in duplicate using two interatomic potentials for bcc Fe, the embedded atom model (EAM) potential developed by Ackland *et al.* [47] and the modified embedded atom model (MEAM) potential introduced by Alireza and Asadi [48].

TABLE I. Database used for training the present regression model. N is the number of atom in the perfect system, N_{cf} the number of distinct instances for a point defect class. I_{2-4} and V_4 denote the interstitial clusters with 2 up to 4 self-interstitial atoms and the quadrivacancy, respectively. The sizes of these systems with defects are $N + (2\ldots4)$ and $N - 4$ for I_{2-4} and V_4, respectively. ϵ is the isotropic and homogeneous rate of deformation for the system.

	Type of point defects (N_{cf})				
System (N, ϵ)	I_2	I_3	I_4	V_4	Total
1024, $\epsilon = +0\%$	434	1105	1280	1701	4520
1024, $\epsilon = -1\%$	434	1105	1280	1701	4520
1024, $\epsilon = +1\%$	434	1105	1280	1701	4520
1024, $\epsilon = +2\%$	434	1105	1280	1701	4520
1024, $\epsilon = +3\%$	434	1105	1280	1701	4520
2000, $\epsilon = +0\%$	434	1105	1280	1701	4520
3456, $\epsilon = +0\%$	434	1105	1280	1701	4520
Total	3038	7735	8960	11907	31640

After an initial period of structure generation, all configurations were pairwise compared to ensure the final database only contained nonequivalent structures. Two configurations are considered as nonequivalent provided that two conditions are verified: (i) their energies differ by more than 10^{-2} eV; (ii) in the case of interstitial defects the sum of squares of the principal components of inertia tensor are different. Interstitial atoms are localized using the Wigner-Seitz method [70].

The resulting database is one order of magnitude larger than that obtained in our previous work [2], due to a wider exploration of phase space with ART*n*. The resulting database is summarized in Table I.

The local descriptors of the retained configurations were computed using the MiLaDy package [33,71] and the scattering coefficients using the PyScatHarm package [35]. To compute the harmonic entropy for each configuration, the Hessian was computed from $3N$ force evaluations using the standard finite-difference formula with a displacement of 10^{-3} Å. Each configuration was tested to be a minimum by checking that the eigenfrequencies are real. For each configuration, we perform an energy relaxation by using LAMMPS [72]. Then, the phonon spectrum and vibrational entropy are computed using the PHONDY package [73–76].

B. Regression procedure

We wish to choose a parametrization $\underline{w}$ for the LDML model Eq. (14) which is able to approximate the M calculated entropies $\underline{S} \in \mathbb{R}^M$ from the M total descriptor vectors $\underline{\underline{D}} \in \mathbb{R}^{\mathcal{D}\times M}$. The general training procedure takes a random subset of $M_t < M$ entropies $\underline{S}_t \in \mathbb{R}^{M_t}$ and descriptor vectors

$\underline{\underline{D}}_t \in \mathbb{R}^{\mathcal{D}\times M_t}$, performs a multilinear regression to determine $\underline{w}$, then tests the prediction against the remaining $M_r = M - M_t$ entropies $\underline{S}_r \in \mathbb{R}^{M_r}$ and descriptor vectors $\underline{\underline{D}}_r \in \mathbb{R}^{\mathcal{D}\times M_r}$ by taking statistical measures of the vector-valued training error $\underline{S}_t - \underline{w}\cdot\underline{\underline{D}}_t$ and test error $\underline{S}_r - \underline{w}\cdot\underline{\underline{D}}_r$. As is standard in machine learning development, we compare both the root mean square (RMSE) and mean absolute (MAE) errors. By defining the L_p magnitude $\|\underline{v}\|^p$ of a vector $\underline{v} \in \mathbb{R}^M$ as $\|\underline{v}\|^p \equiv \sum_{l=1}^{M}|v_l|^p$, the RMSE and MAE errors for the vector-valued error $\underline{S}_s - \underline{w}\cdot\underline{\underline{D}}_s \in \mathbb{R}^{M_s}$ read

$$\sqrt{M_s^{-1}\|\underline{S}_s - \underline{w}\cdot\underline{\underline{D}}_s\|^2} \quad \text{(RMSE)}, \tag{22}$$

$$M_s^{-1}\|\underline{S}_s - \underline{w}\cdot\underline{\underline{D}}_s\|^1 \quad \text{(MAE)}, \tag{23}$$

where $s = t$ gives the training error and $s = r$ the test error.

While multilinear regression is conceptually simple, in practice the optimal parametrization can be difficult to obtain when the number of parameters (here the descriptor vector dimension $\mathcal{D}$) is large. The purpose of standard regression is to minimize the L_2 error: $\|\underline{S}_t - \underline{w}\cdot\underline{\underline{D}}_t\|^2$ with respect to $\underline{w}$. However this can lead to overfitting or highly heterogeneous parametrization. In order to avoid such difficulties we used a ridge regression where a penalty term is added in the minimization: $\|\underline{S}_t - \underline{w}\cdot\underline{\underline{D}}_t\|^2 + \lambda\|\underline{w}\|^2$. To optimize the parameter λ we use Bayesian ridge regression, a probabilistic generalization of multilinear ridge regression that was applied commonly in machine learning [55].

Briefly, in the Bayesian approach one models the error $\underline{S}_t - \underline{w}\cdot\underline{\underline{D}}_t \in \mathbb{R}^{M_t}$ as a multidimensional Gaussian random variable with a diagonal covariance matrix $\sigma^2\mathbb{I}_{M_t}$. In the language of Bayesian estimation, σ is a *hyperparameter* of our estimation procedure, to be distinguished from the model parameters $\underline{w}$ which we want to estimate. This gives a Gaussian likelihood of observing output data $\underline{S}_t$ given model parameters $\underline{w}$, input data $\underline{\underline{D}}_t$, and error variance σ^2 of

$$L(\underline{S}_t|\underline{w}, \underline{\underline{D}}, \sigma) \propto \exp(-\|\underline{S}_t - \underline{w}\cdot\underline{\underline{D}}_t\|^2/2\sigma^2), \tag{24}$$

which is clearly a Gaussian of the L_2 loss function $\|\underline{S}_t - \underline{w}\cdot\underline{\underline{D}}_t\|$ with variance σ^2. A prior distribution of the model parameters $\underline{w}$ is required and as is standard in a Bayesian approach, we chose another multidimensional Gaussian $p_0(\underline{w}|\sigma_w) = \exp(-\|\underline{w}\|^2/2\sigma_w^2)$ with σ_w as the second and final hyperparameter. The product of the prior distribution with the likelihood $L(\underline{S}_t|\underline{w}, \underline{\underline{D}}, \sigma)p_0(\underline{w}|\sigma_w)$ gives a Gaussian of the ridge regularized L_2 loss function with $\lambda = \sigma^2/\sigma_w^2$. The prior distributions $p_0(\sigma)$, $p_0(\sigma_w)$ for the hyperparameters are chosen as Gamma distributions, which can be shown to facilitate the analytical derivations when using Gaussian likelihoods [55]. In practice, the hyperparameters reflect the confidence in the final parametrization.

Defining integrals over the joint hyperparameter prior $p_0(\sigma)p_0(\sigma_w)$ as $\int_{\sigma,\sigma_w}\dots$, the posterior distribution for $\underline{w}$ given training data $\underline{S}_t, \underline{\underline{D}}_t$ reads

$$p(\underline{w}|\underline{S}_t, \underline{\underline{D}}_t) = \mathcal{N}\int_{\sigma,\sigma_w} L(\underline{S}_t|\underline{w}, \underline{\underline{D}}_t, \sigma)p_0(\underline{w}|\sigma_w), \tag{25}$$

where $\mathcal{N}^{-1} = \int d^{\mathcal{D}}\underline{w}\int_{\sigma,\sigma_w}\langle L(\underline{S}_t|\underline{w}, \underline{\underline{D}}_t, \sigma)p_0(\underline{w}|\sigma_w)\rangle_{\sigma,\sigma_w}$ ensures normalization. We aim to find the mode of the posterior distribution to determine the optimal parametrization $\underline{w}$. This is equivalent to maximizing any monotonic function of the posterior with respect to $\underline{w}$, in particular the logarithm, which avoids calculation of the normalization constant $\mathcal{N}$. Our final variational problem for Bayesian ridge regression is thus

$$\underline{w} = \arg\max_{\underline{w}'\in\mathbb{R}^{\mathcal{D}}} \log\int_{\sigma,\sigma_w} L(\underline{S}_t|\underline{w}', \underline{\underline{D}}_t, \sigma)p_0(\underline{w}'|\sigma_w), \tag{26}$$

which is typically the most stable method to determine the optimal parametrization. Bayesian linear regressions have been performed by using the "scikit-learn" package [77]. The initial value of σ_w for the prior is set by default in the code.

V. TESTING OF THE LDML MODEL

A. Influence of interatomic potential and descriptor set

The LDML model formalism was tested on bcc defect systems as described above (Table I), initially in a supercell of size $8a_0 \times 8a_0 \times 8a_0$. The bulk lattice contained 1024 atoms before the introduction of 2–4 interstitial atoms to produce I_{2-4} defects or the removal of 4 atoms for the V_4 quadrivacancies. The volume of the defected supercell is fixed to the equilibrium bulk volume.

We first tested the influence of the underlying interatomic potentials by training and testing the LDML model on either EAM [47] or MEAM [48] data sets. The MEAM potential augments the EAM potential form with angular three-body terms and typically employs analytic expressions for the pair and embedding functions [78–80] as opposed to the tabulated cubic splines of EAM potentials. We also compared three sets of descriptors for the LDML model: (i) $\mathcal{A}_{20,10}$ descriptors, (ii) the bSO(4)$_{3.5}$ bispectrum, both with $r_{\text{cut}} = 5.0$ Å, and (iii) the global scattering descriptor $S^{\mathcal{J},L}$. In our tests the $\mathcal{A}_{20,10}$ was around 50% faster to evaluate than bSO(4)$_{3.5}$ and $S^{\mathcal{J},9}$.

For the three descriptors the MEAM results have a lower RMSE and MAE (see Fig. 2), a feature we found replicated across other training sets. The present surrogate model estimates the multidimensional curvature of the potential energy surface solely from the geometric structure of the minimum basin. Consequently, a smooth energy landscape is assumed; i.e., mathematically speaking, the underlying potential energy surface is a smooth function with regular derivatives. The EAM potential using spline functions does not satisfy the assumed regularity, inducing error in the fitting procedure. The MEAM force field is coded on smoother functions resulting in a smaller intrinsic error in the regression model. This inconvenience is not related to the capacity of the force field, EAM or MEAM, to describe the physics of phonons in bcc iron. In order to reduce the intrinsic error due to the regularity of the force field in this paper we use the MEAM potential exclusively [48].

The results of regressions to the MEAM data with different descriptor functions are shown in Fig. 2. On formation entropies ranging between $8k_B$ and $28k_B$, the performance in descriptor sets has limited variation but we find that $\mathcal{A}_{20,10}$ consistently outperforms bSO(4)$_{3.5}$, and $S^{\mathcal{J},9}$ despite the

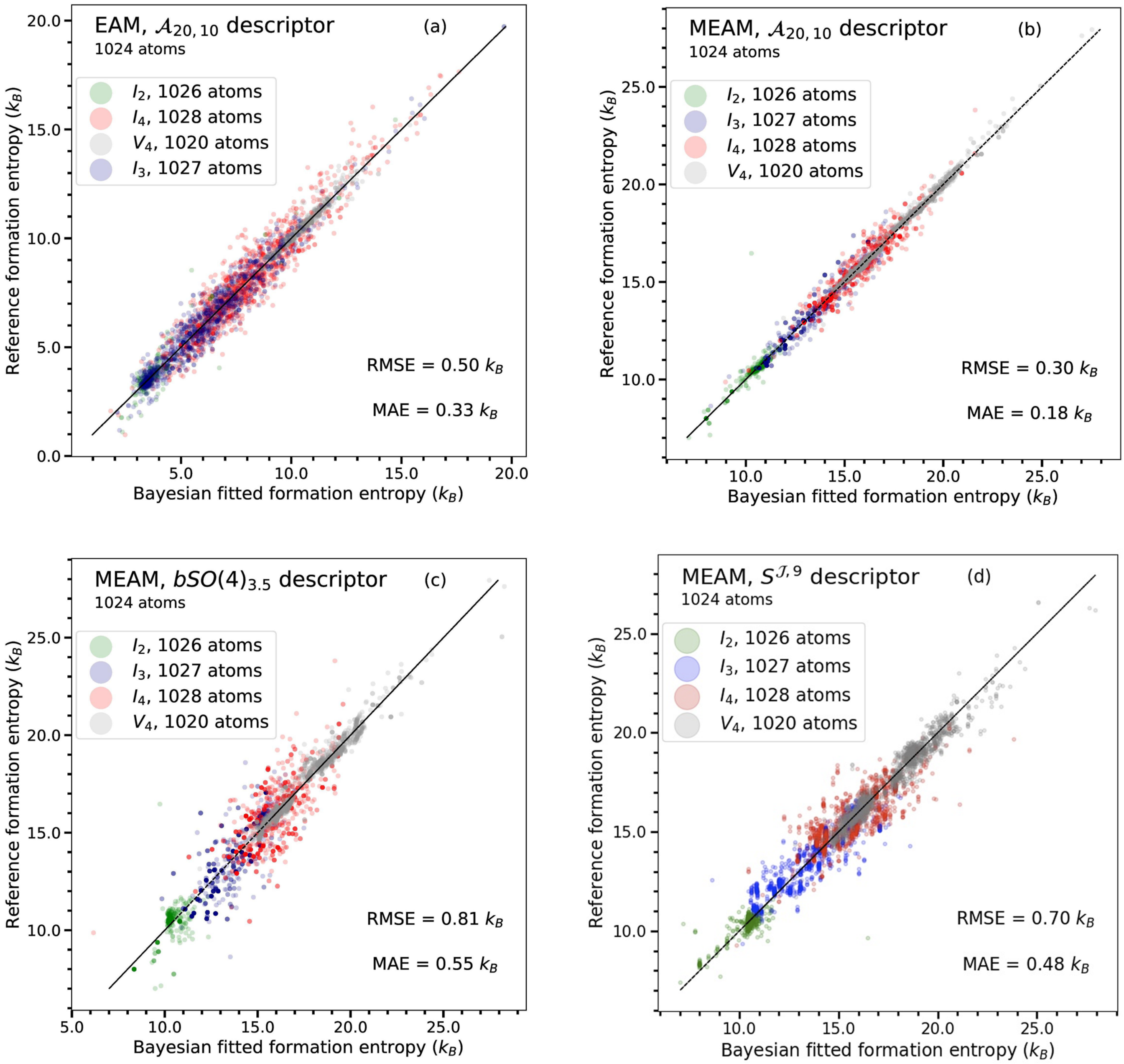

FIG. 2. Formation entropy computed from the numerical diagonalization of Hessian against the formation entropy computed from LDML model using (a) EAM and (b)–(d) MEAM data set for 2–4 interstitial clusters I_{2-4} and for quadrivacancies V_4 in $(8a_0)^3$ supercells. The number of configurations is given at first line of Table I. The descriptors in the present study are (a), (b) $\mathcal{A}_{20,10}$, (c) bSO(4)$_{3.5}$, and (d) $S^{\mathcal{J},L}$ with scales $\mathcal{J} = \{0, 0.25, 0.5, 0.75, 1, 2, 3, 4, 5\}$.

greater computational efficiency, with an RMSE of $0.8k_B$ and $0.7k_B$ to $0.3k_B$, respectively.

B. Modeling data sets with multiple defect species and variable supercell volume

It is highly desirable to have predictability on the changes in formation entropy under deformations of the simulation supercell, as this can be used as a proxy for changes in the formation entropy under varying microstructural environments.

In addition, as LDML model formation entropy Eq. (14) receives an input vector of fixed dimension, independent of system size, it is possible to simultaneously train the model on data sets with a variable number of atoms.

As a first application, we trained the LDML model on a large data set of I_{2-4} and V_4 configurations, found through the ARTn searches. The simulation cell is the same $8a_0$ cubic supercell as above, where each configuration was additionally copied, subjected to an isotropic dilation of -1% to 3% before a new calculation of descriptor vector and harmonic entropy. The number of configurations in the data set has been increased by a factor 5. Figure 3 illustrates the accuracy of the LDML model using a single weight vector $\underline{w}$ for the entire data set. We notice that the RSME error is only $0.4k_B$.

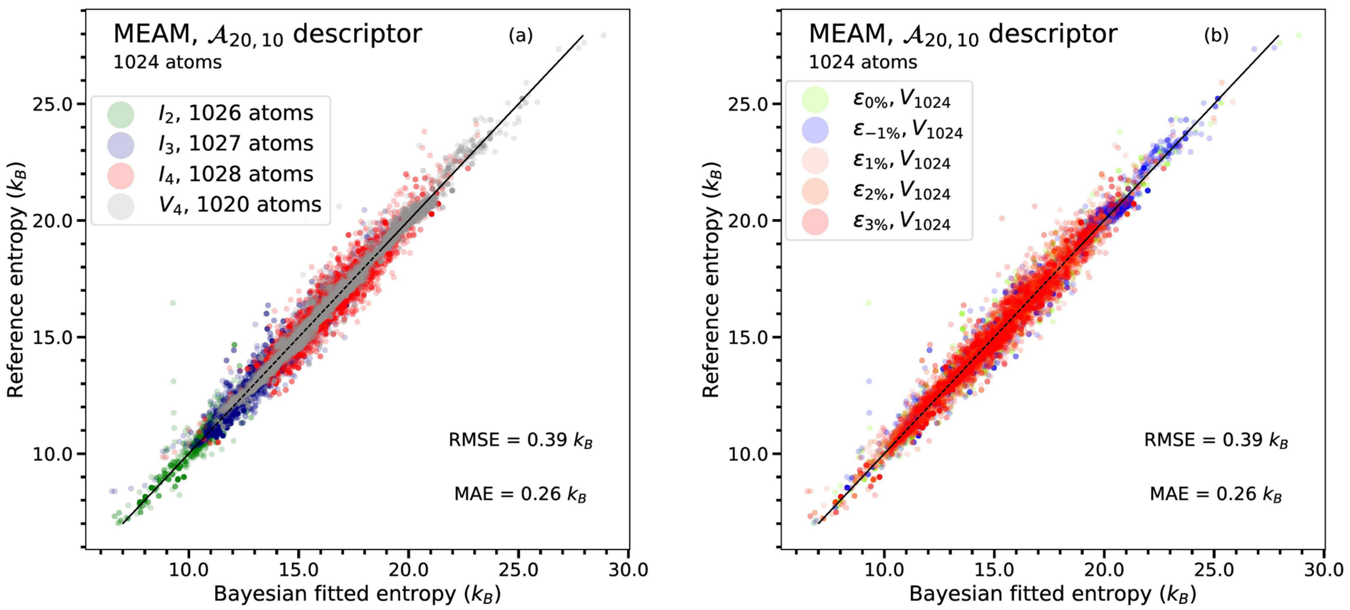

FIG. 3. Illustration of the performance of the training of the surrogate model using deformed supercells of I_{2-4}/V_4 clusters (MEAM database) with the $\mathcal{A}_{20,10}$ descriptor. The initial configurations have a $(8a_0)^3$ volume and have been deformed by applying a homogeneous and isotropic dilatation of the supercell. The deformation ranges from −1% to 3%. (a) Illustration of the results of the regression model depending on the type of defect in the supercell; (b) same as in (a) but for the quadrivacancy V_4 and various deformation rates.

C. Training on combined disordered and crystalline data sets

To test the ability of the LDML model and descriptor functions to predict formation entropies beyond crystalline structures, we created an additional database of highly disordered structures from an ART*n* database of I_{2-4} and V_4 configurations in cubic supercells of dimension $8a_0$, $10a_0$, and $12a_0$, as described in Table I. For each configuration, a large number of individual atoms were subjected to random displacements, which creates many Frenkel pairs even after relaxation. Once the relaxation procedure has been realized we obtained a highly defective structure containing up to 22 vacancies and 26 interstitials. The set of such structures will be referred to as the *random* database. The distribution of defects in the random database is presented in Fig. 4. The difference between the number of interstitials and vacancies is conserved before and after the disordering procedure, giving a strong correlation between the effective vacancy and interstitial count. We also present the distribution of formation entropies, Fig. 5(b), and the distribution of distances between point defects, Fig. 5(a), associated with Fig. 4. The distribution of formation entropies Fig. 5(b) emphasizes that the selected configurations are diverse and carefully selected. Concerning the distribution of distances between defects Fig. 5(a), we can notice that about 1/3 of the distances are less than the “interaction distance” defined by $2r_{\text{cut}} = 10$ Å for the descriptors. Moreover, the interaction between defects is not limited to the cutoff distance of the force field. The point defects used in the present database have a strong elastic dipole tensor [3,81–83] that induces a strong elastic field far beyond the defects and makes almost all the defects interact with each other. Figure 6(a) presents the results of the LDML model trained on this highly diverse data set. We find that the RMSE error is only $0.8k_B$, which is to be compared to a formation entropy range of approximately $250k_B$. This high value of formation entropy in comparison to the ART*n* database, Fig. 2, is ascribed to the much higher effective number of defects in the system.

In order to prove the robustness and the transferability of the model illustrated in Fig. 6(a), a training/testing procedure

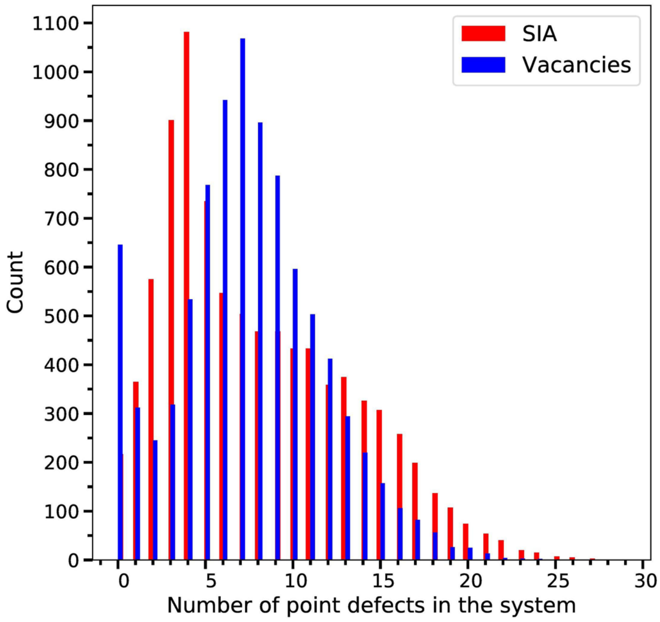

FIG. 4. The analysis of the distribution of randomly generated point defects in the random database. This database is derived from the ART*n* database only using the supercells of volume $(10a_0)^3$ and $(12a_0)^3$ by random creation of Frenkel pairs. The plot emphasizes the occurrences in the entire random database of number of self-interstitial atoms (SIAs) and vacancies in the same supercell. The random database contains 9016 configurations.

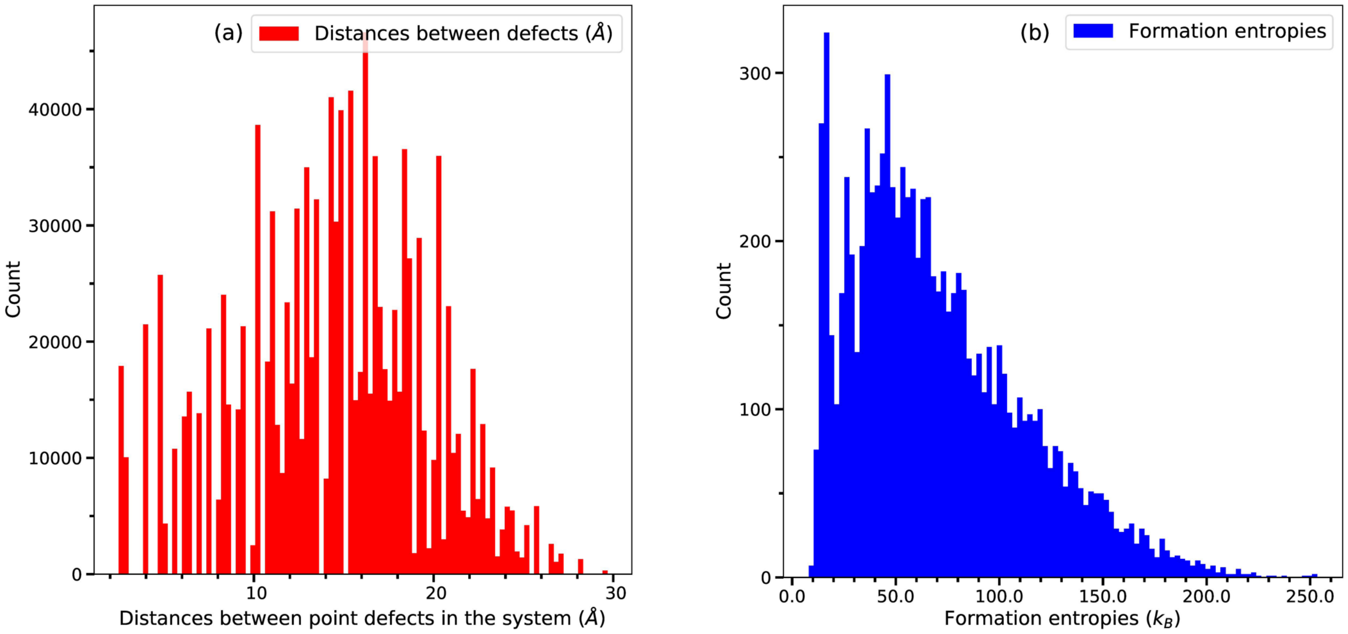

FIG. 5. The analysis of the distribution of formation entropies (b) and of distances between point defects in the random database. Formation entropies follow the same distribution as defect numbers in simulation boxes drawn in Fig. 4. Concerning distance distribution, about 1/3 of the point defects are separated by less than 10 Å = $2r_{\text{cut}}$.

is performed. The database is split randomly into two sets following the proportion p. One set corresponds to the training set with the proportion $(1 - p)$; the second one is the test set with p proportion. The LDML model is adjusted on the training set and a prediction is realized for the test set. In order to reduce bias on the random procedure we iterate this training/testing set sampling a hundred times for a given proportion p and we average the values of RMSE and MAE for the training and test sets. RMSE and MAE calculated for both sets are presented in the inset of Fig. 6(a). The weak variability of RMSE against the proportion p indicates the extremely good quality of predictions, up to a splitting of 90%. We notice for a 90% training/10% testing ratio that the testing error is less than the training error. This behavior reflects the natural tendency of any regression: there are more data in training and therefore the RMSE training is higher. Moreover,

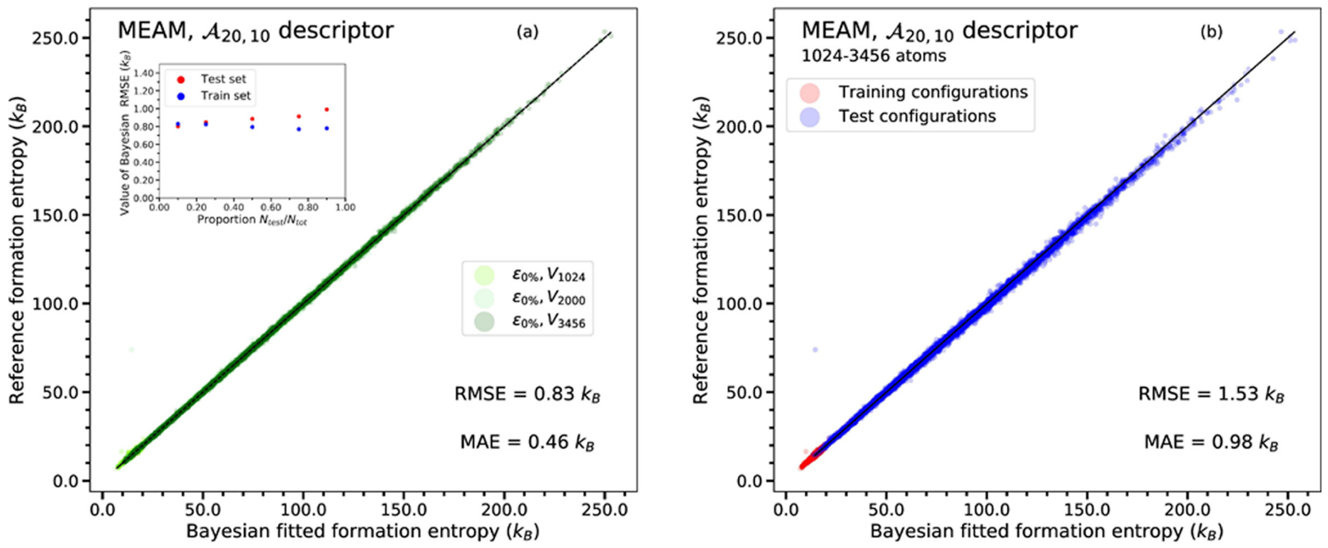

FIG. 6. The robustness and the transferability of the surrogate model are tested by (a) crossing validation using several splitting proportions between training and testing configurations of the joined ARTn database and the random database (the entropies are computed using MEAM potential [48] and the LDML model employs $\mathcal{A}_{20,10}$ descriptor). The statistical indicators are given by the RMSE and MAE. The inset shows RMSE for training and testing data sets against testing proportion (defined in the text). (b) The predictive power of the LDML model trained on the ARTn database and validated on the random database. The statistical indicators, RMSE and MAE, are computed for the random database. The order of magnitude of the statistical indicator is the same as in (a) while the model is in the extrapolation regime.

at that unbalanced ratio it is possible to have some stochastic fluctuations. So far, the probability to have predicted data with a large systematic error (symptomatic data) is higher than the probability to find symptomatic data in the test set.

Defects from the random database are representative structures of bcc iron under irradiation. These structural properties of defects influence drastically the phonon properties. Let us take the case of $\langle 111 \rangle$ interstitial clusters [74,84,85]. These interstitial clusters exhibit a soft mode due to an almost free translation of the dumbbell along the $\langle 111 \rangle$ direction. This phonon mode is highly active in the α-γ martensitic transition of Fe as well as in the pair kinks nucleation in the $\frac{1}{2}\langle 111 \rangle$ dislocation [86] and is delocalized over distances larger than 10 Å.

The ability of the present LDML model to mimic the physics of those soft modes is nontrivial, as the characteristic wavelength is far beyond the cutoff radius of the descriptors used to sample the local atomic environment. Despite this, the linear regression in the descriptor space is able to reconstruct the correlation between high formation entropies and the large phonon wavelengths.

D. Transferability of crystalline model to disordered structures

In this final example, we artificially tested the transferability of the LDML model by training *only* on the ART*n* database of defect structures, before attempting to predict the formation entropies of the random database. As illustrated in Fig. 6(b), the LDML model achieves a remarkable predictive accuracy with an RSME error of only $1.53k_B$. Such a performance is obtained while the prediction is made for a basin of the energy landscape which is disjointed from the training basin where formation entropy is bounded by $25k_B$.

In order to prove the transferability of the LDML in noncrystalline structures we investigate LJ_{38}, the Lennard-Jones cluster containing 38 atoms. It is an archetypal system with thousands of minima organized in many attraction basins [87]. This system is often the benchmark for advanced numerical methods in the exploration of the complex energetic landscapes [88–90]; we used the LJ_{38} database from Cambridge University [90]. For such a system the cluster entropies could be easily calculated by direct diagonalization of the Hessian of the system. We have randomly chosen up to 10 000 different LJ_{38} configurations. The present surrogate model used the AFS(20,10) descriptor having $r_{\text{cut}} = 5$ Å. Figure 7 illustrates the results of the regression model for LJ_{38}, and the inset shows the results of the training/testing procedure. The surrogate model presents the same score and transferability behavior as in the case of the bcc iron system.

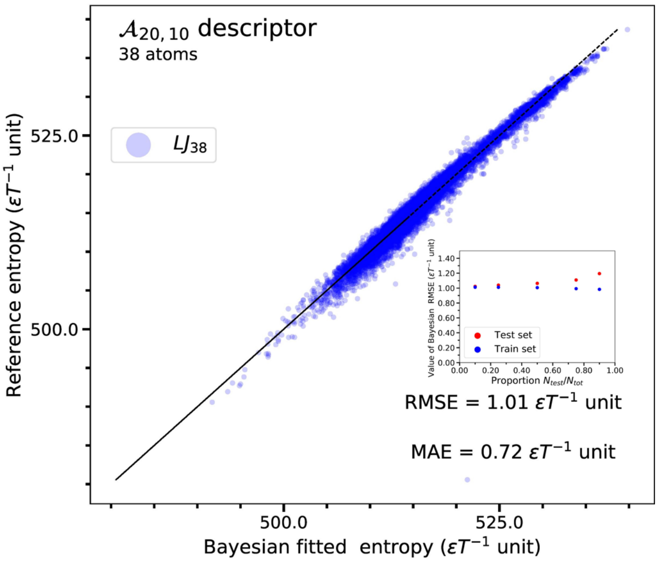

FIG. 7. LDML applied to the Lennard-Jones clusters of 38 atoms LJ_{38} to adjust the vibrational entropy of clusters in ϵT^{-1} units. Training/testing procedure is described in Sec. V C; results are presented in the inset. The statistical indicators RMSE and MAE remain stable even for large proportions of testing set.

VI. CONCLUSIONS AND PERSPECTIVES

This work proposes a strategy to predict the vibrational entropy of structural defects in crystalline solids from the Cartesian coordinates of atoms. After a training phase, the procedure is based solely on geometrical information and does not require explicit knowledge of the Hessian and its spectrum. The $\mathcal{D}$ chosen descriptor functions are calculated for each atom in a relaxed configuration, then summed across all N atoms, giving a model input space of dimension $\mathcal{D}$ independent of the system size N. This reduction is based on the physics of the harmonic approximation that enables us to compute the total vibrational entropy of the system as the sum of the local atomic entropies. This reduction is exact within the harmonic approximation and justifies the summation over the local descriptors in order to build the descriptor of the simulation box. The regression entropy descriptor is then parametrized via methods developed in the machine learning community, specifically Bayesian ridge regression. The extensivity for entropy, in number of particles, volume, and deformations, was carefully checked.

The physics background of the present surrogate model ensures robustness and outstanding transferability. By using two disjoint parts of an extensive database we demonstrate the transferability from supercells containing only one defect cluster to complex configurations having more defects and clusters. The low error in predictions, around $1k_B$ for the absolute values ranging from $20k_B$ to $250k_B$, opens many perspectives; e.g., the defects can be trained separately in small cells, while complicated structures such as those in radiation damage can be accurately predicted [91].

Moreover, the routine calculations of formation vibrational entropies of defects is hindered by the computational cost. The harmonic approximation scales as cubic in number of atoms. In the present approach, the evaluation is very rapid. By far, most of the numerical evaluation is reserved by the calculation of the fingerprint of the local atomic environment. The present algorithm scales linearly with the number of atoms and can be easily parallelized for massive systems.

The present surrogate model opens the way for the prediction of the vibrational entropy of defects up to nanometric size with a host medium up to millions of atoms. Moreover, the present strategy can be integrated in on-the-fly skims such as relaxed Monte Carlo for fast evaluation of the kinetic pathways of the system on the free energy landscape.

ACKNOWLEDGMENTS

This work has been carried out within the framework of the EUROfusion Consortium and has received funding from Euratom Research and Training Programme 2019-2020 under Grant Agreement No. 633053. C.L. and M.-C.M. acknowledge support from GENCI (CINES/CCRT) computer center under Grant No. A0070906973. L.T. and S.M. acknowledge support from the ANR-19-P3IA-0001 program of the PRAIRIE 3IA Institute. We are grateful to Prof. Wales for providing the access to the Cambridge Cluster Database [90].

---

[1] V. V. Bulatov, L. L. Hsiung, M. Tang, A. Arsenlis, M. C. Bartelt, W. Cai, J. N. Florando, M. Hiratani, M. Rhee, and G. Hommes, Nature (London) **440**, 1174 (2006).
[2] M.-C. Marinica, F. Willaime, and N. Mousseau, Phys. Rev. B **83**, 094119 (2011).
[3] R. Alexander, M.-C. Marinica, L. Proville, F. Willaime, K. Arakawa, M. R. Gilbert, and S. L. Dudarev, Phys. Rev. B **94**, 024103 (2016).
[4] M.-C. Marinica, F. Willaime, and J.-P. Crocombette, Phys. Rev. Lett. **108**, 025501 (2012).
[5] C. Woodward and S. I. Rao, Phys. Rev. Lett. **88**, 216402 (2002).
[6] D. A. Terentyev, T. P. C. Klaver, P. Olsson, M.-C. Marinica, F. Willaime, C. Domain, and L. Malerba, Phys. Rev. Lett. **100**, 145503 (2008).
[7] D. Mordehai, E. Clouet, M. Fivel, and M. Verdier, Philos. Mag. **88**, 899 (2008).
[8] Yu. N. Osetsky, D. J. Bacon, Z. Rong, and B. N. Singh, Philos. Mag. Lett. **84**, 745 (2004).
[9] D. J. Wales, *Energy Landscapes* (Cambridge University Press, 2003).
[10] D. J. Wales, J. Chem. Phys. **130**, 204111 (2009).
[11] T. D. Swinburne and D. Perez, Phys. Rev. Mater. **2**, 053802 (2018).
[12] S. S. Kapur, M. Prasad, J. C. Crocker, and T. Sinno, Phys. Rev. B **72**, 014119 (2005).
[13] A. Satta, F. Willaime, and S. de Gironcoli, Phys. Rev. B **57**, 11184 (1998).
[14] G. H. Vineyard and G. J. Dienes, Phys. Rev. **93**, 265 (1954).
[15] H. B. Huntington, G. A. Shirn, and E. S. Wajda, Phys. Rev. **99**, 1085 (1955).
[16] N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Holt-Saunders International Editions: Science: Physics (Holt, Rinehart, and Winston, 1976).
[17] T. Lelièvre, G. Stoltz, and M. Rousset, *Free Energy Computations: A Mathematical Perspective* (Imperial College Press, London, 2010).
[18] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).
[19] A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U.S.A. **99**, 12562 (2002).
[20] L. Maragliano and E. Vanden-Eijnden, Chem. Phys. Lett. **446**, 182 (2007).
[21] Weinan E, Weiqing Ren, and Eric Vanden-Eijnden, Phys. Rev. B **66**, 052301 (2002).
[22] T. Lelièvre, M. Rousset, and G. Stoltz, J. Chem. Phys. **126**, 134111 (2007).
[23] L. Bonati and M. Parrinello, Phys. Rev. Lett. **121**, 265701 (2018).
[24] E. Darve, D. Rodríguez-Gómez, and A. Pohorille, J. Chem. Phys. **128**, 144120 (2008).
[25] T. D. Swinburne and M.-C. Marinica, Phys. Rev. Lett. **120**, 135503 (2018).
[26] L. Mauger, M. S. Lucas, J. A. Muñoz, S. J. Tracy, M. Kresch, Y. Xiao, P. Chow, and B. Fultz, Phys. Rev. B **90**, 064303 (2014).
[27] D. Bansal, A. Aref, G. Dargush, and O. Delaire, J. Phys.: Condens. Matter **28**, 385201 (2016).
[28] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[29] J. Behler, J. Chem. Phys. **134**, 074106 (2011).
[30] G. Imbalzano, A. Anelli, D. Giofrè, S. Klees, J. Behler, and M. Ceriotti, J. Chem. Phys. **148**, 241730 (2018).
[31] A. P. Bartók, Gaussian approximation potential: An interatomic potential derived from first-principles quantum mechanics, Ph.D. thesis, Cambridge University, 2009 .
[32] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).
[33] A. M. Goryaeva, J.-B. Maillet, and M.-C. Marinica, Comput. Mater. Sci. **166**, 200 (2019).
[34] M. Eickenberg, G. Exarchakis, M. Hirn, and S. Mallat, in *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., New York, USA, 2017), pp. 6540–6549.
[35] M. Eickenberg, G. Exarchakis, M. Hirn, S. Mallat, and L. Thiry, J. Chem. Phys. **148**, 241732 (2018).
[36] M. Hirn, S. Georges Mallat, and N. Poilvert, Multiscale Model. Simul. **15**, 827 (2017).
[37] E. Homer, D. M. Hensley, C. Rosenbrock, A. Nguyen, and G. Hart, Front. Mater. **6**, 168 (2019).
[38] D. Dragoni, T. D. Daff, G. Csányi, and N. Marzari, Phys. Rev. Mater. **2**, 013808 (2018).
[39] W. J. Szlachta, A. P. Bartók, and G. Csányi, Phys. Rev. B **90**, 104108 (2014).
[40] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[41] C. Chen, Z. Deng, R. Tran, H. Tang, I.-H. Chu, and S. P. Ong, Phys. Rev. Mater. **1**, 043603 (2017).
[42] A. Thompson, L. Swiler, C. Trott, S. Foiles, and G. Tucker, J. Comput. Phys. **285**, 316 (2015).
[43] V. Botu and R. Ramprasad, arXiv:1410.3353.
[44] J. R. Kermode, A. Gleizer, G. Kovel, L. Pastewka, G. Csányi, D. Sherman, and A. De Vita, Phys. Rev. Lett. **115**, 135501 (2015).
[45] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Phys. Rev. Lett. **114**, 105503 (2015).
[46] L. Dezerald, L. Proville, L. Ventelon, F. Willaime, and D. Rodney, Phys. Rev. B **91**, 094105 (2015).
[47] G. J. Ackland, M. I. Mendelev, D. J. Srolovitz, S. Han, and A. V. Barashev, J. Phys.: Condens. Matter **16**, S2629 (2004).
[48] S. A. Etesami and E. Asadi, J. Phys. Chem. Solids **112**, 61 (2018).
[49] B. Grabowski, L. Ismer, T. Hickel, and J. Neugebauer, Phys. Rev. B **79**, 134106 (2009).

[50] B. Grabowski, T. Hickel, and J. Neugebauer, Phys. Status Solidi B **248**, 1295 (2011).

[51] A. Glensk, B. Grabowski, T. Hickel, and J. Neugebauer, Phys. Rev. Lett. **114**, 195901 (2015).

[52] P. H. Dederichs, R. Zeller, and K. Schroeder, *Point Defects in Metals II: Dynamical Properties and Diffusion Controlled Reactions* (Springer Tracts in Modern Physics, Berlin, 1980).

[53] Prog. Mater. Sci. **55**, 247 (2010).

[54] C. Huang, A. F. Voter, and D. Perez, Phys. Rev. B **87**, 214106 (2013).

[55] C. E. Rasmussen, *Gaussian Processes in Machine Learning* (Springer, Berlin, 2004).

[56] D. Frenkel and S. Berend, *Understanding Molecular Simulation: From Algorithms to Applications* (Academic Press, Cambridge, UK, 2002).

[57] H. Zong, G. Pilania, X. Ding, G. J. Ackland, and T. Lookman, npj Comput. Mater. **4**, 48 (2018).

[58] G. Ferré, J.-B. Maillet, and G. Stoltz, J. Chem. Phys. **143**, 104114 (2015).

[59] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, J. Phys. Chem. C **121**, 511 (2017).

[60] E. D. Cubuk, S. S. Schoenholz, J. M. Rieser, B. D. Malone, J. Rottler, D. J. Durian, E. Kaxiras, and A. J. Liu, Phys. Rev. Lett. **114**, 108001 (2015).

[61] F. Noe and C. Clementi, J. Chem. Theory Comput. **11**, 5002 (2015).

[62] K. T. Schutt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, J. Chem. Phys. **148**, 241722 (2018).

[63] A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis, and P. G. Debenedetti, Chem. Phys. Lett. **509**, 1 (2011).

[64] W. F. Reinhart, A. W. Long, M. P. Howard, A. L. Ferguson, and A. Z. Panagiotopoulos, Soft Matter **13**, 4733 (2017).

[65] R. Kakarala, The bispectrum as a source of phase-sensitive invariants for Fourier descriptors: A group-theoretic approach, Ph.D. thesis, University of California, Irvine, 1992.

[66] G. T. Barkema and N. Mousseau, Phys. Rev. Lett. **77**, 4358 (1996).

[67] R. Malek and N. Mousseau, Phys. Rev. E **62**, 7723 (2000).

[68] E. Cancès, F. Legoll, M.-C. Marinica, K. Minoukadeh, and F. Willaime, J. Chem. Phys. **130**, 114711 (2009).

[69] E. Machado-Charry, L. K. Bèland, D. Caliste, L. Genovese, T. Deutsch, N. Mousseau, and P. Pochet, J. Chem. Phys. **135**, 034102 (2011).

[70] E. Wigner and F. Seitz, Phys. Rev. **43**, 804 (1933).

[71] M. C. Marinica *et al.*, *MiLaDy: Machine Learning Dynamics* (CEA, Saclay, 2015–2019).

[72] S. Plimpton, J. Comput. Phys. **117**, 1 (1995).

[73] M. C. Marinica *et al.*, *Phondy: Phonon Dynamics* (CEA, Saclay, 2007–2019).

[74] M.-C. Marinica and F. Willaime, Solid State Phenomena **129**, 67 (2007).

[75] A. Soulié, F. Bruneval, M.-C. Marinica, S. Murphy, and J.-P. Crocombette, Phys. Rev. Mater. **2**, 083607 (2018).

[76] F. Berthier, J. Creuze, T. Gabard, B. Legrand, M.-C. Marinica, and C. Mottet, Phys. Rev. B **99**, 014108 (2019).

[77] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011).

[78] M. S. Daw and M. I. Baskes, Phys. Rev. B **29**, 6443 (1984).

[79] M. S. Daw, S. M. Foiles, and M. I. Baskes, Mater. Sci. Rep. **9**, 251 (1993).

[80] M. I. Baskes, Phys. Rev. B **46**, 2727 (1992).

[81] C. Varvenne, F. Bruneval, M. C. Marinica, and E. Clouet, Phys. Rev. B **88**, 134102 (2013).

[82] M. Boleininger and S. L. Dudarev, Phys. Rev. Mater. **3**, 093801 (2019).

[83] D. R. Mason, D. Nguyen-Manh, M.-C. Marinica, R. Alexander, A. E. Sand, and S. L. Dudarev, J. Appl. Phys. **126**, 075112 (2019).

[84] G. Lucas and R. Schäublin, Nucl. Instrum. Methods Phys. Res., Sect. B **267**, 3009 (2009).

[85] S. Chiesa, P. M. Derlet, and S. L. Dudarev, Phys. Rev. B **79**, 214109 (2009).

[86] L. Proville, D. Rodney, and M.-C. Marinica, Nat. Mater. **11**, 845 (2012).

[87] D. Wales, *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*, Cambridge Molecular Science (Cambridge University Press, Cambridge, UK, 2004).

[88] D. J. Wales, Mol. Phys. **100**, 3285 (2002).

[89] D. J. Wales, Mol. Phys. **102**, 891 (2004).

[90] D. J. Wales, Cambridge Cluster Database, http://www-wales.ch.cam.ac.uk/CCD.html.

[91] S. J. Zinkle and B. N. Singh, J. Nucl. Mater. **199**, 173 (1993).