Machine learning surrogate models for strain-dependent vibrational properties and migration rates of point defects

Clovis Lapointe, Thomas D. Swinburne, Laurent Proville, Charlotte Becquart, Normand Mousseau, Mihai-Cosmin Marinica

Machine learning surrogate models for strain-dependent vibrational properties and migration rates of point defects

Clovis Lapointe,[1,*] Thomas D Swinburne,[2,†] Laurent Proville,[1] Charlotte S. Becquart,[3] Normand Mousseau,[4] and Mihai-Cosmin Marinica[1,‡]

[1] *Université Paris-Saclay, CEA, DES-Service de Recherches de Métallurgie Physique, F-91191, Gif-sur-Yvette, France*
[2] *Aix-Marseille Université, CNRS, CINaM UMR 7325, Campus de Luminy, 13288 Marseille, France*
[3] *Université Lille, CNRS, INRA, ENSCL, UMR 8207 - UMET - Unité Matériaux et Transformations, F-59000 Lille, France*
[4] *Département de Physique and Regroupement québécois sur les matériaux de pointe, Université de Montréal, C.P. 6128, Succursale Centre-Ville, Montréal, H3C 3J7 Québec, Canada*

(Dated: November 29, 2022)

Machine learning surrogate models employing atomic environment descriptors have found wide applicability in materials science. In our previous work, this approach yielded accurate and transferable predictions of the vibrational formation entropy of point defects for $\mathcal{O}(N)$ computational cost. The present study investigates the limits of data driven surrogate models in accuracy and applicability for vibrational properties. We propose an improvement of the accuracy by extending the fitting capacity of the model by increasing the dimension of the descriptor space. This is achieved by using a non-linear relation between descriptors - target observables and when it is possible by including physical relevant information of the underlying energy landscape. The non-linear extension is used to learn the formation entropy of defects with or without applied strain whilst including physical information such as the minimum-saddle point sequences are employed for the migration of point defects, a key ingredient of transition state theory rate approximations. We find excellent predictive power after augmenting the dimensionality of the descriptor space, as demonstrated on large defect databases in α-iron and amorphous silicon based on semi-empirical force fields. The current linear surrogate models are used to investigate the correlation between migration entropy and energy. Our approaches reproduce the Meyer-Neldel compensation law observed from direct calculations in amorphous Si systems. Moreover, the same abstract descriptor space representation for entropy and energy is then used for the statistical correlation analysis. For linear surrogate models, we show that the energy-entropy statistical correlations can be reinterpreted in descriptor space. This provides a simple statistical criterion for marginal interpretation of the compensation law. More generally, the present work shows how linear surrogate models can accelerate high-throughput workflows, aid the construction of mesoscale material models and provide new avenues for correlation analysis.

Keywords: Harmonic approximation, defects, Machine Learning, vibrational entropy, attack frequency, migration rates, energy-entropy correlations, compensation law

I. INTRODUCTION

The thermodynamic and kinetic properties of defects drive the microstructural evolution of materials [1, 2]. The appropriate thermodynamic potential of defects, such as free energy and enthalpy, gives the equilibrium density of defects, while the defects' kinetic properties steer the kinetic pathways of the microstructure towards equilibrium. Therefore, the study of solids at finite temperature requires the correct description of the free energy thermodynamic potential of defects. The defects' free energy accounts for the contribution of the system's microstates around a particular atomic configuration. In materials, the morphology of defects shows extraordinary variety and is related to the complexity of the underlying energy landscape.

Defects have a wide range of shapes and sizes: from localized 2D dislocation loops [3, 4] to 3D clusters such as voids [5, 6], stacking fault tetrahedra [7, 8], small interstitial cluster [9], insertions with particular crystallographic structure [10] etc. The complexity of the energetic landscape is amplified by the fact that those defects are embedded in a host matrix. The complexity does not come from just the arrangement of atoms within the defects, but also from the interface between the defects and the surrounding matrix. Very often, in order to characterise these complex energetic landscapes, the configurations of defects are indexed by the local minima of the energy landscape and the fluctuations associated with thermal vibrations are neglected [11–18]. Therefore, the entire domain of the free energy's phase space $\mathbb{R}^{3N}$, of N atoms, has particular topology: it is sparse and sliced into basins of attraction, which can be indexed with discrete labels. This approach is similar to that used to describe atomic clusters in vacuum [19]. Any configuration $\boldsymbol{q} \in \mathbb{R}^{3N}$ is then a member of some discrete states, which belong to the same basin of attraction [19]. Hence, the entire basin can be represented by the corresponding local minimum to which $\boldsymbol{q}$ converges to under local minimization of the internal energy $U(\boldsymbol{q})$. Within this framework, probably the most common visual representation

* clovis.lapointe@univ-lorraine.fr
† thomas.swinburne@cnrs.fr
‡ mihai-cosmin.marinica@cea.fr

of energetic landscape is provided by the disconnectivity graph techniques [19–21]. For defects, it is very practical to account for the formation free energy, i.e. the difference between the free energy of the perfect and the defective solid containing the same number of N atoms. The formation free energy converges to a well-defined value in the thermodynamic limit $N \to \infty$ [22]. However, the kinetic evolution of microstructure is driven by the pathways and the connections of the sequences minimum - saddle point - minimum [11, 23–25]. For these kinds of sequences, the quantification of the energy contribution has been widely studied and the community has well-established tools [12, 24, 26–28]. The entropic contribution is still challenging to quantify because of the methodological and numerical complexity [1, 29, 30].

The present study focuses on the evaluation of the vibrational entropy S^{vib} for minima and saddle points configurations of the energy landscape in the framework of the harmonic approximation. The vibrational entropy $S^{\text{vib}}(\boldsymbol{q})$ of some defect is directly related to the curvature of the phase space $\boldsymbol{q} \in \mathbb{R}^{3N}$ in some particular point of the potential energy surface on which the defect is located. Many alternative methods to harmonic approximation have been designed and tested in the community [31–39] to compute the free energy of defects, including even the non-harmonic contributions from energy and entropy in an indistinguishable manner. However, these methods remain computationally very demanding as they usually rely on sampling the phase space of the system through the construction of random or optimized trajectories. Furthermore, in most cases the main limitation is not the numerical efficiency and the poor scalability of computational methods. There is a conceptual problem: the community does not have the appropriate theoretical tools in order to handle in a systematic manner the anharmonic *finite temperature* vibrational entropy of a complex energy landscape. Currently, there are no existing general sampling methods to estimate the free energy barrier profile between two states of a complex energy landscape. The community proposes some promising methods for particular cases [31, 33–40] where is possible to build *intuitively* or *automatically* an *ad hoc* reaction coordinate.

In the harmonic approximation, the vibrational entropy can be computed from the knowledge of the frequency of normal modes, which itself requires the evaluation and the diagonalization of local Hessian matrix operations that have $\mathcal{O}(N^2)$ and $\mathcal{O}(N^3)$ complexity, respectively. This traditional procedure requires an interatomic interaction that can be treated in the framework of electronic structure calculations, such as *ab initio* calculations, or of empirical interatomic potentials. This workflow can be bypassed by the recent surrogate model proposed by Lapointe et al. [41], which proposes a linear correlation between the atomic descriptors of the local atomic environment and vibrational entropy. That model was applied for point defects in crystals and nanoparticles [41] and opens many perspectives for the fast evaluation of vibrational properties around energy-minimum configurations. Moreover, the comparison of the surrogate model with direct calculations of formation vibrational entropies of defects reveals an excellent accuracy and predictive power. The direct evaluation of the Hessian ($\mathcal{O}(N^2)$) and its diagonalization ($\mathcal{O}(N^3)$) is replaced with $\mathcal{O}(N)$ computational effort. The utility of such a model is huge: the numerical efficiency increases drastically. Consider the case of evaluating the formation entropy of a dislocation loop containing 200 self-interested atomsin α-Fe. To avoid finite size effects, traditional direct evaluation requires simulation cells with over 120,000 atoms. For this size, solely the diagonalization of the Hessian matrix requires 6 hours on 3000 modern CPUs, whilst the surrogate model provides the same observable, within 5 % error, in less than 10 minutes on one CPU i.e. more than 10^5 faster.

The aim of this study is to explore the performance of surrogate model approaches for other physical observables in the field of materials science. Here, we apply the surrogate model for amorphous systems as well as defects in crystals under deformation. The deformed systems can be defects in minima configurations but equally first-order saddle points configurations. With this motivation, the previous surrogate model [41] is revisited by integrating specific physics, such as the metastable character of saddle point configurations. Moreover, in order to have a better representation in the descriptor space of the non-deformed minimum configurations, we introduce a non-linear extension of the machine learning surrogate model. The surrogate model formalism is then used to learn and predict the kinetical transition rates during defect migration.

Figure 1(A) and Figure 1(B) present the graphical summary of the traditional and current workflow, respectively, for the computation of vibrational harmonic entropies. The only inputs required for the present surrogate model are the optimized atomic coordinates of various defect configurations. Minimum and saddle point configurations are generated employing a 0 K method for systematic search in complex energetic landscapes, Activation-Relaxation Technique nouveau (`ARTn`). The collection of found defect configurations, in α-Fe and amorphous Si, defines the database of our machine learning surrogate model. The efficient `ARTn` 0 K method is used employing an inter-atomic force field based on a semi-empirical potential. Then, the system's Hessian matrix is computed and diagonalized to obtain the vibrational harmonic entropy S^{vib}. Figure 1(B) emphasizes the machine learning (ML) sequence of the workflow. Descriptors of each configuration of the database are computed. Then, using a supervised regression models the machine learning model is trained in order to estimate S^{vib}.

The paper is structured as follows. In Section II, we summarize the link between the vibrational entropy and the Green function formalism. The latter enables the total entropy of the system to be decomposed into local

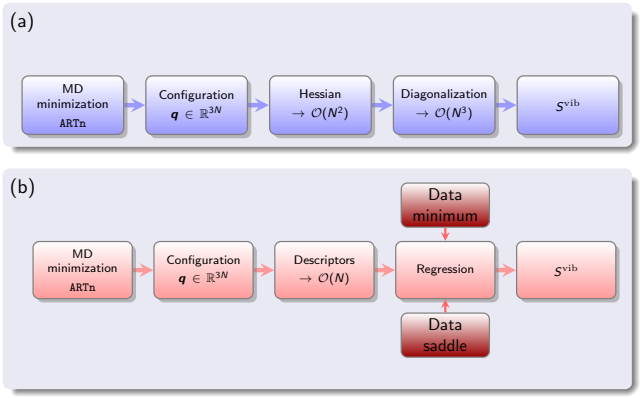

FIG. 1: Graphical abstract of the vibrational entropy calculations using the traditional (A) and the present workflows of vibrational harmonic entropies. Atomic configurations are generated and optimized using the ARTn method and semi-empirical potentials for α-Fe and amorphous Si. Then, in the traditional approach (A), the system's Hessian matrix is computed and diagonalized in order to obtain vibrational harmonic entropy S^{vib}. In the present machine learning workflow (B), to estimate S^{vib} we calculate the atomic descriptors of the same atomic configurations. Then the surrogate regression model is trained using the corresponding database of minimum and saddle point configurations.

atomic contributions. In the same section, we describe the descriptor space used in order to obtain an appropriate representation of the atomic environments. The construction of this space is based on the encoding, using atomic descriptors, of the local atomic neighborhood of each atom by preserving all the geometrical invariances. In Section III, we explore the limits of the present non-linear extension of the surrogate model in order to increase the accuracy of vibrational entropy predictions when the system is constrained to a small deformation $\boldsymbol{\epsilon}$. The efficacy of our approach is demonstrated on strained defects in α-Fe. In Section IV, we use our approach to build a surrogate model for attack frequencies, in particular for the logarithm of the attack frequency, which is proportional to a vibrational entropy difference between the initial minimum atomic configuration and the saddle point configuration. The model is applied for the case of amorphous silicon. In Section V, the same surrogate model architecture is employed to predict the transition energy barrier, giving a complete surrogate model for harmonic transition state theory rate calculations. We use linear surrogates models to investigate the correlation between migration energy and entropy. Within the present framework we give a statistical insight of the compensation Meyer-Neldel law from the perspective of the energetic and entropic surrogate models and descriptor space.

II. THE HARMONIC VIBRATIONAL ENTROPY AND GREEN FUNCTIONS

The normal modes of a system with N atoms are obtained from the spectrum of the dynamical matrix $\tilde{\mathfrak{D}}$:

$$(\tilde{\mathfrak{D}} - \omega_\nu^2 \mathbb{I}) \cdot \hat{\mathbf{e}}_\nu = \mathbf{0}, \tag{1}$$

where $\mathbb{I}$ is the identity matrix. When the dynamical matrix includes only the phonons from the center of Brillouin zone, it reduces to the mass normalized Hessian operator of the system: $\boldsymbol{M}^{-\frac{1}{2}}.\boldsymbol{\mathcal{H}}.\boldsymbol{M}^{-\frac{1}{2}}$, where $\boldsymbol{M}$ is a diagonal matrix which return the mass of each atom. For unary systems of mass m, $\boldsymbol{M} = m\mathbb{I}$. The Hessian is built from the second derivatives of the potential energy $U(\mathbf{q})$ of the system: $\mathcal{H}_{i\alpha j\beta} = \partial^2 U/\partial q_{i\alpha}\partial q_{j\beta}$. Vibrational mode (or Γ point phonons) frequencies ω_ν^2 and displacements $\hat{\mathbf{e}}_\nu$ are the eigenvalues and eigenvectors of the above secular equation (1). At high temperature T (above the Debye temperature in cristalline solids) in the so-called harmonic approximation, the classical vibrational entropy of the system becomes:

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

where k_B and $\hbar$ are the Boltzmann and Planck constants, respectively.

A. Green function formalism for vibrational entropy calculations

Within the harmonic approximation, vibrational entropy estimation requires the full spectrum of the secular equation (1). The Green function formalism provides an iterative solution to the mentioned eigen-problem. The total density of states of vibrational modes $\Omega(\omega)$ is extracted from the imaginary part of the trace of Green function $\boldsymbol{\mathcal{G}} \in \mathbb{C}^{3N\times 3N}$ [23]:

$$\Omega(\omega) = \frac{2\omega}{\pi}\Im\left(\mathrm{Tr}\left\{\boldsymbol{\mathcal{G}}(\omega)\right\}\right), \tag{3}$$

$$\boldsymbol{\mathcal{G}}(\omega) = \sum_\nu \frac{\hat{\mathbf{e}}_\nu \otimes \hat{\mathbf{e}}_\nu}{\omega_\nu^2 - \omega^2} = \left[\tilde{\mathfrak{D}} - \omega^2\mathbb{I}\right]^{-1}, \tag{4}$$

and the above total density of state (DOS) of Γ-point phonons, $\Omega(\omega)$, verifies the following constraint with respect to the degrees of freedom of the system under 3D periodic boundary conditions:

$$\int_0^\infty \Omega(\omega)d\omega = 3N - 3. \tag{5}$$

We can express the classical vibrational entropy of the system at a given temperature T, from the total density of vibrational modes states:

$$S^{\text{vib}} = -k_B \int_0^\infty \left[\ln\left(\frac{\hbar\omega}{k_B T}\right) - 1\right]\Omega(\omega)d\omega. \tag{6}$$

Very often, the eigenmode ($\hat{\mathbf{e}}_\nu$, ω_ν) is delocalized over several atoms. For example, low energy and large wavelength Rayleigh phonons induce concerted atomic motions across the entire system. However, the local contribution to S_i^{vib} can be deduced [23] by transforming from the delocalized basis $\hat{\mathbf{e}}_\nu$ of the Green function in equation (4) to a localized basis $\hat{\mathbf{e}}_{i\alpha}$ of atomic sites, which will yield the local density of states $\Omega_i(\omega)$. This transformation is not unique, and, in the next section, we will choose the standard local projection.

B. Local basis for densities of states of vibrational modes

The orthonormal bases of normal modes $\hat{\mathbf{e}}_\nu$ and atomic sites $\hat{\mathbf{e}}_{i\alpha}$ (where $\hat{\mathbf{e}}_{i\alpha}$ are unit vectors for displacement of atom i along direction α) are both complete over the configurational phase space, and are therefore related by a unitary transformation:

$$\hat{\mathbf{e}}_\nu = \sum_i \sum_\alpha \xi^{i\alpha}(\nu)\, \hat{\mathbf{e}}_{i\alpha}, \tag{7}$$

where the square of the rotation matrix elements, $|\xi^{i\alpha}(\nu)|^2$, can be seen as the probability of the mode $\hat{\mathbf{e}}_\nu$ to be localized on atom i and along the α direction.

By analogy with equation (4), the local density of state of Γ-point phonons reads:

$$\mathrm{Tr}(\boldsymbol{\mathcal{G}}(\omega)) = \sum_{i,\alpha} \hat{\mathbf{e}}_{i\alpha} \cdot \boldsymbol{\mathcal{G}}(\omega) \cdot \hat{\mathbf{e}}_{i\alpha}\,, \tag{8}$$

$$\Omega_i(\omega) = \sum_\alpha \frac{2\omega}{\pi} \Im\left(\hat{\mathbf{e}}_{i\alpha} \cdot \boldsymbol{\mathcal{G}}(\omega) \cdot \hat{\mathbf{e}}_{i\alpha}\right)\,, \tag{9}$$

where $\Im(\cdot)$ selects the imaginary component, giving a local vibrational entropy

$$S_i^{\text{vib}} = -k_B \int_0^\infty \left[\ln\left(\frac{\hbar\omega}{k_B T}\right) - 1\right] \Omega_i(\omega) d\omega. \tag{10}$$

We note that the vibrational entropy of the system is completely defined by projected normal modes on each atom. Furthermore, in this local basis the total entropy can be exactly decomposed into local entropies. Regression of vibrational entropy can thus be achieved by a local approach depending on the atomic neighborhood. This local problem needs an accurate and systematic representation of the local atomic environment as well as the fact that this projection mixes the eigenvectors.

C. Local atomic environment encoded into local atomic descriptors

At the heart of the current approach is the relationship between the local density of states and the local atomic environment. Atomic descriptors are numerical tools developed to describe quantitatively the local atomic environments [42–46]. A large number of the methods and studies were presented in the literature to build regression models between local physical observable and local atomic descriptors [46–52]. In the next section, we will introduce simple models which link the local density of states and the local atomic environment. Even if the choice of local basis is not unique we will numerically show that the above local basis is a good choice to state the proportionality of descriptors and the local environment. The relation between the two quantities is established by the use of local atomic descriptors that provide the non-linear encoding of the geometrical neighborhood of each atom. These atomic descriptors project the atomic configurations into descriptor space. This non-linear function maps the entire neighborhood of a central atom into a space with fixed dimension, hence, we assume the existence of a set of functions $\{\underline{D}^i\}_{0\leq i\leq N}$ such that

$$\begin{aligned} \underline{D}^i : \mathbb{R}^{3N} &\to \mathbb{R}^{\mathcal{D}} \\ \boldsymbol{q} &\to \underline{D}^i(\boldsymbol{q}), \end{aligned}$$

where $\underline{D}^i$ is the local atomic descriptor for the i^{th} atom and $\boldsymbol{q}$ is the vector of coordinates of the entire system. Moreover, $\mathcal{D}$ is independent of the number of atoms in the system. Descriptor functions should respect the symmetries of the system, i.e. applying symmetry operations (e.g. permutations, translations and rotations) on the input coordinate vector $\boldsymbol{q}$ should not change the value of $\underline{D}^i$.

The notion of atomic descriptors in material science was introduced by Behler and Parrinello [42–44]. They proposed the $\mathbf{G}_2$ descriptor, which is sensitive to the radial distribution of neighboring atoms, weighted by a Gaussian. Since then, many descriptors have been developed by (i) introducing the explicit angular description, as the $\mathbf{G}_3$ descriptor [42], (ii) using the spectral decomposition in $3D$ or $4D$ spherical functions of the atomic density [45, 46] (iii) decomposition of the total energy in manybody contributions that are expanded in tensorial decomposition in particular basis [53–57] (iv) particular design for a given system [58–63], (v) or using machine / deep learning methods in order to find the appropriate descriptors [49, 50, 64, 65], (vi) hybrid descriptors that can mix all other classes mentioned above [51]. The dimension of the descriptor space, $\mathcal{D}$, ranges from few tens to few thousands. The dimension is flexible and is often used to control the level of accuracy necessary to represent the local atomic environment in the descriptor space. The numerical cost is also proportional to $\mathcal{D}$.

In this paper we choose to work with bi-spectrum SO(4) ($bSO(4)$) descriptor [45, 46]. This descriptor is based on the $4D$ hyper-spherical harmonics decomposition onto the unit sphere of $\mathbb{R}^4$. With this projection any function returns values in $\mathbb{R}^3$ as described in [45]. The local environment of the i^{th} atom is described as a density $\rho_i(\boldsymbol{q})$, and can be decomposed on the basis of $4D$

spherical harmonics

$$\rho_i(\boldsymbol{q}) = \sum_{k\in\mathcal{R}_i} w_k \delta(\boldsymbol{q}-\boldsymbol{q}_k+\boldsymbol{q}_i), \tag{11}$$

$$= \sum_{k\in\mathcal{R}_i}\sum_{j=0}^{\infty}\sum_{m,m'=-j}^{j} \boldsymbol{c}_{i,j}^{m,m'} U_j^{m,m'}, \tag{12}$$

w_k is species-dependent weight, the $\boldsymbol{c}_{i,j}^{m,m'}$ are the result of the scalar product between the density centered on atom i and the hyper-spherical harmonic $U_j^{m,m'}$. $\mathcal{R}_i$ is the cut-off radius for atom i, $\boldsymbol{q}_k$ and $\boldsymbol{q}_i$ are the coordinates of atom k and i, respectively. The components of $bSO(4)$ are defined by the following equation:

$$B^i_{jj_1j_2} = (\boldsymbol{c}_{i,j}^{m,m'})^\dagger \boldsymbol{H^{j_1j_2}} (\boldsymbol{c}_{i,j_1}^{m_1,m_1'} \otimes \boldsymbol{c}_{i,j_2}^{m_2,m_2'}), \tag{13}$$

where $j \le j_{max}$, $|j_1-j_2| \le j \le j_1+j_2$ and $\boldsymbol{H^{j_1j_2}}$ is related with the Clebsch-Gordan coefficient of SO(4) group. In this study we use $j_{max}=4$ and select only the diagonal components $j_1=j_2$ [45, 46, 66] yielding the total number of components to 35. Otherwise stated, the cut-off distance is set to 5Å . The $bSO(4)$ descriptor is an over-complete basis of representation of the SO(3) group. In the present study, we extend our previous formalism [41] to a quadratic machine learning model. The main improvements of the regression model for formation entropy or attack frequency are achieved by increasing the dimension of the descriptor space: from $2\mathcal{D}$ (for attack frequencies) to $\mathcal{D}^2$ (for formation entropy of systems under deformations), where $\mathcal{D}$ is the initial dimension of the descriptor space. The higher the dimension of fit, the more parameters are required. This high-dimensional regression increases the risk of overfitting and decreases the predictive power of the model. An accurate description of the local atomic environment can be obtained using the bSO(4) descriptor with a relatively low number (35) of components [45, 46, 51] Therefore, descriptor bSO(4) provides the right balance between the low dimensional descriptor space (critical for the ML quadratic model) and the precision of the representation. This low dimensional descriptor space helps to prevent the risk of overfitting.

III. EXTENSION OF SURROGATE MODEL FOR VIBRATIONAL ENTROPY AT HIGHER ORDERS

In previous work [41] we have shown that the local entropy $S_i^{\rm vib}$ is proportional to the local atomic descriptors:

$$S_i^{\rm vib} = \underline{w}\cdot\underline{D}^i\,, \tag{14}$$

where $\underline{w}$ is the weight vector that parametrizes the surrogate model. $\underline{D}^i = \{B^i_{jj_1j_2}\}_{0\le j\le j_{max}}$ such as $|j_1-j_2| \le j \le j_1+j_2$ is the local atomic descriptor of the atom i given by Eq. (13). For $j_{max}=4.0$, the number of components is equal to 35. Moreover, this proportionality expressed in [41] is reinforced by the definition of harmonic vibrational entropy given by Dederichs *et al.* [23], $S^{\rm vib}$ as a sum of local terms on each atom :

$$S^{\rm vib} = \sum_i S_i^{\rm vib}. \tag{15}$$

The above equation (15) is exact when summing over all atoms in the system. In this study, we explore how the mere proportionality of Eq. 14 may be enhanced beyond linearity, e.g., by considering a quadratic model. Before introducing the quadratic model, we introduce the following vector and matrix notations for the rest of the study:

$$\underline{D} = \sum_i \underline{D}^i \in \mathbb{R}^{\mathcal{D}}, \quad \boldsymbol{D} = \sum_i \underline{D}^i \cdot [\underline{D}^i]^\top \in \mathbb{R}^{\mathcal{D}\times\mathcal{D}}\,. \tag{16}$$

By using the above notations our quadratic model for the vibrational entropy reads

$$S^{\rm vib} = \underline{w}\cdot\underline{D} + {\rm Tr}\left\{\boldsymbol{W}\cdot\boldsymbol{D}\right\}, \tag{17}$$

where $\underline{w}$ is the weight vector of the linear model Eq. (14) and $\boldsymbol{W} \in \mathbb{R}^{\mathcal{D}\times\mathcal{D}}$ is a learnable matrix. We note that $\underline{D}/N$ and $\boldsymbol{D}/N - \underline{D}\underline{D}^\top/N^2$ give the descriptor means and sample covariances, respectively. An equivalent way of writing this quadratic model is given by:

$$S^{vib} = \sum_i [\underline{w}\cdot\underline{D}^i + \boldsymbol{W} : \underline{D}^i\otimes\underline{D}^i + o\left(\|\underline{D}^i\otimes\underline{D}^i\otimes\underline{D}^i\|\right)], \tag{18}$$

where : denotes the double contraction operator. The second term reads $\boldsymbol{W} : \underline{D}^i\otimes\underline{D}^i = \sum_{k=1}^{\mathcal{D}}\sum_{l=1}^{\mathcal{D}} W_{k,l} D_k^i D_l^i$. Finally, the third term accounts to $\underline{D}^i\otimes\underline{D}^i\otimes\underline{D}^i = \sum_{k=1}^{\mathcal{D}}\sum_{l=1}^{\mathcal{D}}\sum_{m=1}^{\mathcal{D}} D_k^i D_l^i D_m^i\, \underline{e}_k\otimes\underline{e}_l\otimes\underline{e}_m$ (where $\{\underline{e}_i\}_{1\le i\le\mathcal{D}}$ are the canonical basis of the descriptor space) and corresponds to 3^{rd} order (i.e. cubic) term. In the present study, we consider only the quadratic expansion. Increasing the order of expansion increases the number of model parameters and thus the fitting capacity of the model, but also increases the risk of overfitting. We show below that a well chosen quadratic model form gives better accuracy than the linear model while maintaning excellent transferability.

We will call as "Extended Quadratic" Machine Learning (EQML) this second order formulation. The current EQML model preserves the entropy extensiveness, is local and enables direct comparison with the linear model Eq. (14). The performances of EQML model (17) are tested on the database of small defects under strain in bcc iron built in our previous work [41].

A. The database: small defects in BCC iron

A point defect can be created in a perfect crystalline system of N_b atoms by adding or removing N_d atoms. We

are going to denote with $\mathcal{C}$ the resultant defective structure. The formation vibrational entropy $S_{f,\mathcal{C}}$ at temperature T reads:

$$S_{f,\mathcal{C}}(T, N_d) = S_{d,\mathcal{C}}(T, N_b \pm N_d) - \frac{N_b \pm N_d}{N_b} S_b(T, N_b), \tag{19}$$

By definition, the vibrational entropies of the defected structure $S_{d,\mathcal{C}}(T, N_b \pm N_d)$ and of the perfect bulk structure $S_b(T, N_b)$ are computed at the same volume V. Under the harmonic approximation, the formation entropy will thus depend only on the energies $\hbar\omega_\nu$ of normal modes. For high temperatures, larger than the crystal Debye temperature, such that $\max\left(\frac{\hbar\omega_\nu}{k_B T}\right) \ll 1$, the expression becomes simply

$$S_{f,\mathcal{C}}(N_d) = k_B \ln\left(\frac{\prod_{\nu_b} (\hbar\omega_{\nu_b})^{\frac{N_b \pm N_d}{N_b}}}{\prod_{\nu_d} \hbar\omega_{\nu_d}}\right), \tag{20}$$

where ω_{ν_b} and ω_{ν_d} are the frequencies of the perfect bulk and defect configurations, respectively. This classical formulation of the formation vibrational entropy is useful in the present study because it is possible to assign to each defect configuration a single value for vibrational entropy, instead of a function that depends on temperature (as would be the case in the limit of quantified phonons). This formulation is not a limitation of the present study, but a choice to demonstrate the feasibility of current surrogate models.

To train our machine learning model we used a database of 4506 configurations of bcc iron, including various defect morphologies such as 2-4 self interstitials and 4 vacancy clusters in supercells of $(8a_0)^3$ ($N_b = 1024$). All these instances of the defects are minima of the energy landscape and have been generated using the `ARTn` method. [24, 67]. The Activation-Relaxation Technique nouveau [24, 67–69] is a powerful method for searching saddle points and transition pathways of a given potential energy surface. The `ARTn` method is designed to explore the energy landscape of the system using only the lowest eigenvalue and the associated eigenvector of the Hessian. The `ARTn` method is composed of two main steps, the activation step and the relaxation step. The activation step consists in moving the system from a local minimum to a saddle point. The relaxation step consists in relaxing the system, from the computed saddle point, to another local minimum. At the end the `ARTn` method provides 0 K minimum - saddle point - minimum sequences. All the 4506 minima configurations are different. In order to obtain a more complete dataset, we perform data augmentation. On the `ARTn` minimum configurations of the dataset, we have applied a homogeneous and isotropic strain of -1% to 3%, resulting in a total of 22530 configurations. Reference harmonic vibrational entropies are computed by using molecular dynamics simulations with `LAMMPS` [70] and the `PHONDY` package [71–73] from the direct evaluation of vibrational spectrum. `PHONDY` package enables the evaluation of the phonon spectrum of crystalline systems by direct diagonalization of the dynamical matrix of the system [1, 71–73] and allows the computation of vibrational properties in the framework of the harmonic approximation. We use the modified embedded atom potential (MEAM) developed by Alireza and Asadi [74]. Descriptors of each configuration have been computed by using the `MILADY` package [51, 52, 75].

B. Train/test procedure

The linear model is parametrized with the same procedure described in [41] using Eq. (14). The parametrization of the EQML model is sequential, in two steps procedure, being preconditioned by the linear model and then adjusting the quadratic part. Firstly, we set the values of the linear model, $\underline{w}$ in Eq. (17) by a linear fit using Bayesian ridge regression. Secondly, the target property becomes the differences $\Delta S_k \equiv S_k - \hat{S}_k^{\text{lin}}$ for each configuration k, where $\hat{S}_k^{\text{lin}}$ is the estimation of S_k with the linear model. The values of the elements of the tensor $\boldsymbol{W}$, in Eq. (17), are parametrized using ΔS_k values using the same Bayesian ridge regression. All the fitting procedure is performed with ridge Bayesian Regression by using `scikit-learn` package [76] (the initial value of σ_w Bayesian prior has been set by default).

The robustness and the transferability of the surrogate EQML model is checked by performing a train/test procedure. Here, we define two statistical quantities in order to evaluate the quality of the surrogate model:

$$\sqrt{\frac{1}{M}\sum_k^M \left(\hat{S}_k^{\text{vib}} - S_k^{\text{vib}}\right)^2} \quad \text{(RMSE)}, \tag{21}$$

$$\frac{1}{M}\sum_k^M |\hat{S}_k^{\text{vib}} - S_k^{\text{vib}}| \quad \text{(MAE)}, \tag{22}$$

where S_k^{vib} and $\hat{S}_k^{\text{vib}}$ are the formation entropy and the predicted formation entropy of the k^{th} configuration, respectively. RMSE and MAE are the Root Mean Square Error and the Mean Absolute Error, respectively. The database configurations were randomly divided into two sets at a certain ratio of p. The percentage of the training configurations is set at $(1-p)$ (from the total database) and the rest of the database, with a ratio of p, is reserved for tests. The surrogate model is fitted on the training set and the predictions are evaluated for the test set. RMSE and MAE are calculated for both sets. In order to reduce bias of the random procedure selection, we iterate the procedure a hundred times for a given ratio of p and we average the value of RMSE and MAE for the training and test set.

C. The performance of the EQML surrogate model

Here, we compare the two regression models presented in the previous sections: (i) the linear model given by

equation (14) and (ii) the EQML model given by equation (17). The comparison is made using Figure 2: linear machine learning approach from [41] is presented in the Figure 2(a) while EQML model is presented in Figure 2(b). Both models, linear and quadratic, use the same parametrization for the $bSO(4)$ descriptor i.e. $j_{\max} = 4.0$ and the cut-off radius is set to 5 Å . The number of dimensions for linear and quadratic fit are $\mathcal{D} = 35$ and $\mathcal{D} + \mathcal{D}(\mathcal{D}+1)/2 = 665$, respectively. EQML model provides better RMSE/MAE fit errors than the simple linear model. The non-linearity of EQML model introduces coupling between descriptor components giving larger fit capacity. For each class of deformations, EQML model has lower RMSE ($0.34 k_B$) compared to linear model, $0.82\, k_B$. From the learning curves presented in the insets of Figure 2 it is possible to deduce that the linear model is transferable even for a very small fraction $1 - p$. EQML remains stable even for large test / small train database's partition. EQML learning curves emphasize small over-fitting behavior for the ratio $p > 0.7$. Even under these conditions, the RMSE of EQML model is inferior to the linear model. It is important to note that the additional quadratic term preserves the good transferability capacity of the linear model. The strong linear preconditions of the quadratic parametrization keep the characteristics of the linear fit. Moreover, as we pointed out in [41], the information provided by the local vibrational modes is sufficient to quantitatively recover vibrational modes properties, even if they are delocalized over the cutoff distance of descriptors.

IV. APPLICATION OF LINEAR MODELS FOR HARMONIC TRANSITION STATE THEORY (HTST)

The kinetic pathways of the microstructural evolution of the system are driven by the migration energies / free energies landscape. Sequences of minima and saddle points drive complex phenomena such as the agglomeration of defects in a larger structure, e.g. a dislocation loop [77] or the mobility of the dislocations [1, 78]. In the following, we will call a sequence of connected minimum - saddle point - minimum as an event $\mathcal{E}$. The transition rate is the probability of realisation of that event in a specific order. The migration coefficients can be computed from the transition rates between the relevant minima of the energy landscape [13, 15, 79–82]. As such, the transition rates are observable of paramount importance in the implementation of multiscale simulations such as Kinetic Monte Carlo (KMC), whatever the variant: objects [83–89], events [90–93], or off-lattice framework [11, 94]. Each variation of the KMC method mentioned above has a special recipe for defining the physical reality corresponding to the transition rate. In object KMC, for example, each rate is associated to an atomistic transition rate of a particular defect which is the main concept of the present study. The event and off-lattice KMC deals with rates, which correspond to some physical phenomenons such as collective migration and reaction of defects.

Let's consider an event $\mathcal{E}$, which is defined by two instances: the initial minimum state $\mathcal{E}, m$ and an associated saddle point state $\mathcal{E}, s$. We define the rate from state $\mathcal{E}, m$ to state $\mathcal{E}, s$ as $R_{\mathcal{E},m\to s}$. The overall sequence of the event contributes to the transition rate [95, 96]. However, within the framework of the most used approximation in computational materials science for the transition rate, namely the Transition State Theory (TST) [95, 97], only the initial minimum and the saddle point of the events account for the transition probability. In the harmonic approximation, the transition rates gives [95, 97]:

$$R_{\mathcal{E},m\to s} = \nu^*_{\mathcal{E},ms} e^{-\beta \Delta E_{\mathcal{E},m\to s}}, \tag{23}$$

where $\beta = (k_B T)^{-1}$, $\Delta E_{\mathcal{E},m\to s}$ is the energy difference between the saddle point and the minimum and $\nu^*_{\mathcal{E},ms}$ is the attack frequency defined in the framework of harmonic TST as:

$$\nu^*_{\mathcal{E},ms} = \frac{\prod_{\nu_{m'}\in\mathcal{S}(\mathcal{E},m)} \nu_{m'}}{\prod_{\nu_{s'}\in\mathcal{S}(\mathcal{E},s)} \nu_{s'}}, \tag{24}$$

where the numerator is the product of frequencies at the minimum $\mathcal{S}(\mathcal{E}, m)$ and the denominator is the product of frequencies at the saddle point $\mathcal{S}(\mathcal{E}, s)$. By $\mathcal{S}(\mathcal{E}, m)$ and $\mathcal{S}(\mathcal{E}, s)$, we denote the ensemble of the Hessian spectrum at the minimum $(\mathcal{E}, m)$ and saddle point $(\mathcal{E}, s)$ configuration of event $\mathcal{E}$, respectively. Obviously, the unstable modes such as the three zero modes due to the periodic boundary conditions as well as the imaginary mode of the saddle point are not included in ensemble $\mathcal{S}$. In practice, under harmonic approximation, $\nu^*_{\mathcal{E},ms}$ is obtained by diagonalizing the dynamical matrix of the system's minimum and saddle point. The numerical complexity of this procedure is $\mathcal{O}(N^3)$ and limits the size of the routinely studied systems to a few tens of thousands of atoms. Due to this complexity, the majority of multiscale studies use phenomenological laws of unique value for the attack frequencies (such as the Debye frequency for crystalline materials). Here, we propose to extend the linear machine learning approach in the descriptor space to compute and predict the attack frequencies for a collection of events $\{\mathcal{E}^i\}$. In this way, as in the case of formation vibrational entropy, we avoid a direct diagonalization of the dynamical matrix which is replaced by the $\mathcal{O}(N)$ operation of computing descriptors.

A. Reformulation of the attack frequency

For the attack frequency defined in Eq. (24), it is better to handle the logarithm of frequencies for ensemble $\mathcal{S}$, which are positively definite.

$$\ln\left(\nu^*_{\mathcal{E},ms}\right) = \sum_{\nu_{m'}\in\mathcal{S}(\mathcal{E},m)} \ln\left(\nu_{m'}\right) - \sum_{\nu_{s'}\in\mathcal{S}(\mathcal{E},s)} \ln\left(\nu_{s'}\right), \tag{25}$$

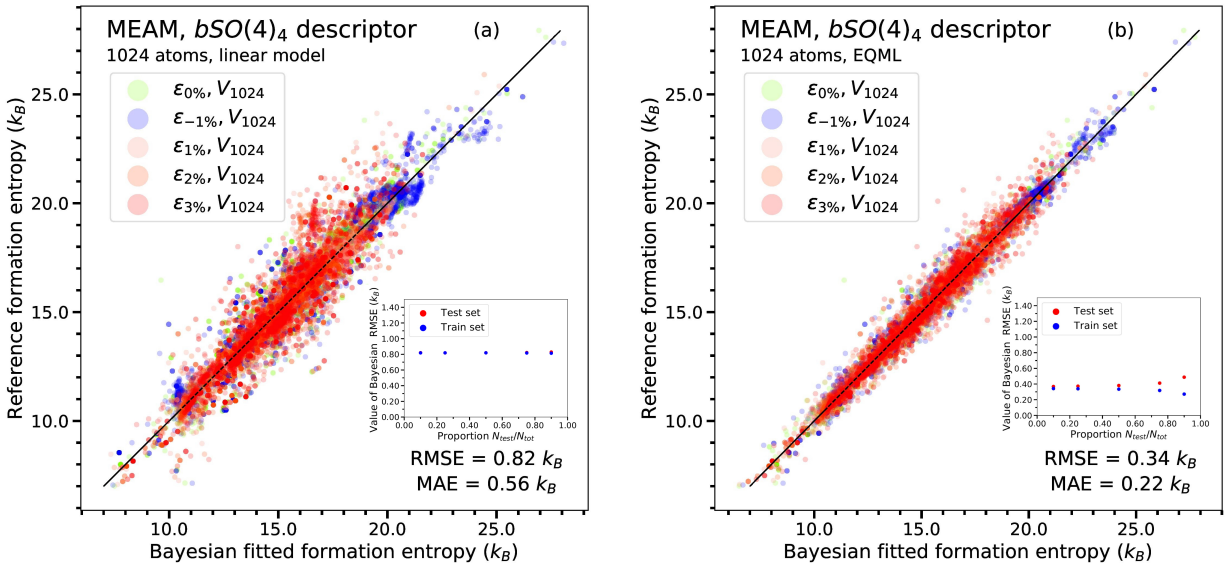

FIG. 2: Illustration of the performance of linear and EQML surrogate models using deformed supercells of I_{2-4}/V_4 clusters and by using $bSO(4)_4$ descriptor. The initial configurations have a $(8a_0)^3$ volume and have been deformed by applying an homogeneous and isotropic dilatation of the supercell. The deformation rates range from -1% to 3%. Figures illustrate the performances of linear (a) and EQML (b) of surrogate models.

which can be related to the density of states $\Omega_j(\omega)$ for the j^{th} state of the system:

$$\sum_{\nu_{j'} \in \mathcal{S}(\mathcal{E},j)} \ln\left(\nu_{j'}\right) = \int_0^{+\infty} \ln\left(\frac{\omega}{2\pi}\right) \Omega_j(\omega) d\omega. \qquad (26)$$

Similarly, the logarithm of the attack frequency can be decomposed into local contributions. The associated surrogate model can be set up using a linear model in descriptor space. The regression model attack frequency - descriptors now has the following formulation:

$$\ln\left(\nu^*_{\mathcal{E},ms}\right) = \underline{w}_1 \cdot \left(\underline{D}_{\mathcal{E},m} \oplus \underline{D}_{\mathcal{E},s}\right), \qquad (27)$$

where $\underline{D}_{\mathcal{E},m/s} = \sum_{d \in \mathcal{E},m/s} \underline{D}^d \in \mathbb{R}^D$ is the total descriptor vector for $\mathcal{E}, m$ or $\mathcal{E}, s$. We denote by $\cdot \oplus \cdot$ the direct concatenation operator for two descriptor vectors. This extended descriptor space is sufficiently general to capture the fact that the evaluation of reaction rate $R_{\mathcal{E},m\to s}$ of the event $\mathcal{E}$ requires information about the minimum configuration m and the saddle point configuration s. The direct sum of descriptor vectors for states $\mathcal{E}, m$ and $\mathcal{E}, s$ gives an extended descriptor vector of dimension $2\mathcal{D}$.

B. The database for transition rates: amorphous Si

The performance of the surrogate model for attack frequencies was tested on amorphous Si. This system is a challenging test case where each minimum of the energy landscape is connected to a large number of saddle points. This system was widely studied in the past using the `ARTn` method [15, 94, 98, 99]. The amorphous system is contained within a cubic cell of 4096 atoms at constant volume. The spectrum of partial Hessian is estimated by using the Lanczos method described by Marinica *et al.* [9], and we fix the admissible error on Lanczos eigenvalues at $1\times10^{-6}(1.018049\times10^{-2}\text{ps})^{-2}$. The amorphous Si system is explored through inherent states. The inherent states are representative states of disordered materials corresponding to attraction basins (local minimum surrounded by many other minima and saddle points). Inherent states give access to a potentially astronomical number of different configurations and activation barriers. The database of events $\mathcal{E}$ is generated from 20 independent inherent states of amorphous silicon, simulated using a modified version of the Stillinger-Weber potential [100]. To obtain inherent states, we run canonical molecular dynamics simulations [70], with 4096 Si atoms, a fixed density of 2.192 g.cm^{-3}, and using a timestep of 1 fs. Random configurations are first equilibrated at 2300 K for 20 ns, and then directly relaxed (i.e. without any intermediate quench) at 700 K during 100 ns. Finally, the system energy is minimized using the `FIRE` algorithm [101], until all components of the force vector are lower than 10^{-9} eV.Å^{-1}. Once initial minima are prepared, we sample saddle points with `ARTn` (converging towards them until all force vector components are lower than 10^{-7} eV.Å^{-1}). The connectivity of saddle points to initial minima is systematically checked using the steepest descent method, and duplicates are removed by comparing saddle points energies and displacements of the most displaced atom [30]. At the end, the amorphous Si database contains an average of 420 saddle points per minimum (a total of 10502 distinct activated events).

C. Surrogate model for attack frequencies

The surrogate model was trained using the $bSO(4)$ with $j_{\max} = 4.0$. For consistency reasons, the descriptor was computed using a cut-off equal to that of the empirical potential used to perform the MD simulations, $r_{\text{cut}} = 3.77\mathring{A}$ [100]. The direct sum descriptor spans a descriptor space with $35 + 35$ components. The performances of the linear model are presented in figure 3 for logarithm (a) and plain values (b) of attack frequencies. The spectrum of attack frequencies is wide over many orders of magnitude.The linear model is capable of predicting magnitude changes in a fairly accurate manner. The value of the $\log(\nu/\nu_0)$ (ν_0 is fixed at 1 THz) has the RMSE of 0.65 for a range of 10 $\log(\nu/\nu_0)$. The results of the train/test procedure, presented in the inset of the figure 3.(a), indicate that the model is stable even for a small train fraction of the database. This behavior of the present linear model is similar to that of the previous surrogate model for vibrational formation entropies [41].

The results of plain attack frequencies are shown in Figure 3.(b), the RMSE is about 1400 THz for a range of values from 1×10^{-1} to 1×10^{5} THz. The logarithm of the stochastic noise for the direct attack frequency model, $\log(\nu/\hat{\nu})$ where $\hat{\nu}$ is the predicted frequency, follows a normal distribution presented in the inset of Figure 3.(b) with a mean $\mu = 0$ and a standard deviation $\sigma = 0.653$. We can estimate that 98 % of the predictions $\hat{\nu}$ verify the following ratio $0.27 \approx e^{-2\sigma} \leq \nu/\hat{\nu} \leq e^{2\sigma} \approx 3.7$. Further analysis of the distribution of errors will be carried out in the context of the statistical formulation of the compensation law, in the next section.

In conclusion, under the linear formalism in descriptor space, the attack frequencies can be inferred and predicted. This approach is purely geometric and based on the local decomposition of the density of states of vibrational modes, bypassing the direct calculations and diagonalization of the dynamical matrix. This type of approach scales like $\mathcal{O}(N)$ and has the same order of magnitude as the calculation of descriptors and could be used in any multi-scale atomistic methodology that uses transition rates.

V. THE COMPENSATION MEYER-NELDEL LAW IN DESCRIPTORS SPACE

It is frequently found in thermally activated processes - i.e. following an Arhenius-like law given in Eq. (23) - that when the activation energy increases within a family of processes, the prefactor also increases. Thus, observed first in chemistry [102] and then in physics by Meyer and Neldel [103], the increase of a rate prefactor somewhat "compensates" for the decrease in the Arrhenius exponential term governing the dependence on temperature. This experimental observation can be expressed as a simple correlation between the observed prefactor (ν_{exp}) and the slope (ΔE_{exp}) of the measured Arrhenius law: $\ln \nu_{\text{exp}} = a + b\Delta E^{\alpha}_{\text{exp}}$ (where α is an exponent, which usually is taken as 1). Evidence of a direct link between the experimental measurements and numerical simulations is difficult to obtain [30, 104, 105]. The main complexity is that in experimental measured Arrhenius law there is a contribution of many kinetic pathways for which it is difficult to account theoretically / numerically the huge number of thermally activated sequences of events. Evidence of a direct link between experimental measurements and numerical simulations is difficult to obtain [30, 104, 105]. The main complexity is that in experimental measurements of the Arrhenius law, there is a contribution from many kinetic pathways which are difficult to theoretically and/or numerically account for the huge number of thermally activated sequences of events. However, in the case of simple energetic landscapes, a direct association can be established between the migration mechanism and experimental observation. This is the case for self-diffusion in crystalline solids or migration of particular defects, e.g. diffusion of ad-atoms on metallic surfaces [106, 107] or thermal activated dislocation movements [108]. This compensation was even demonstrated analytically in the case of well identified thermal overcoming of energetic barriers, such as in the correlation between the magnitude of the gap in semiconductors and the thermally activated conductivity [109]. This compensation is explained by some authors with the concept of multi-excitation entropy and its consequences. When a fluctuation involving a large number of excitations occurs, for example when a large activation barrier is overcome, there must be a large entropy associated with this fluctuation [105]. Otherwise, when the kinetic pathways are difficult to identify, such as the migration in disordered solids, some studies [30, 104, 105] describe a compensation effect for large set of events between the average value of the pre-exponential factor logarithm $\ln\left(\nu^{*}_{\mathcal{E},m\to s}\right)$, in a given bin / window of activation energy, and the energy barrier. This correlation is called enthalpy-entropy compensation law and results from the averaging over a large number of kinetic pathways [30]. Moreover, in the general framework of thermally activated diffusion of kinetic processes, beyond an array of interpretations all along the last 100 years, Gelin *et. al* [30] recently proposed (based on numerical simulations) a general interpretation for which the compensation is a statistical law associated with the deformation of the vibrational spectrum caused by the local deformations of the atomic lattice and that can be calculated within the harmonic TST.

The goal of this section is two fold: (i) we first examine if the present surrogate model for attack frequencies is able to recover enthalpy-entropy compensation law that was previously identified by the direct calculations. (ii) Then, in the framework of the linear model in descriptors space, we try to give a statistical insight of the compensation law by the simple fact that the present surrogate

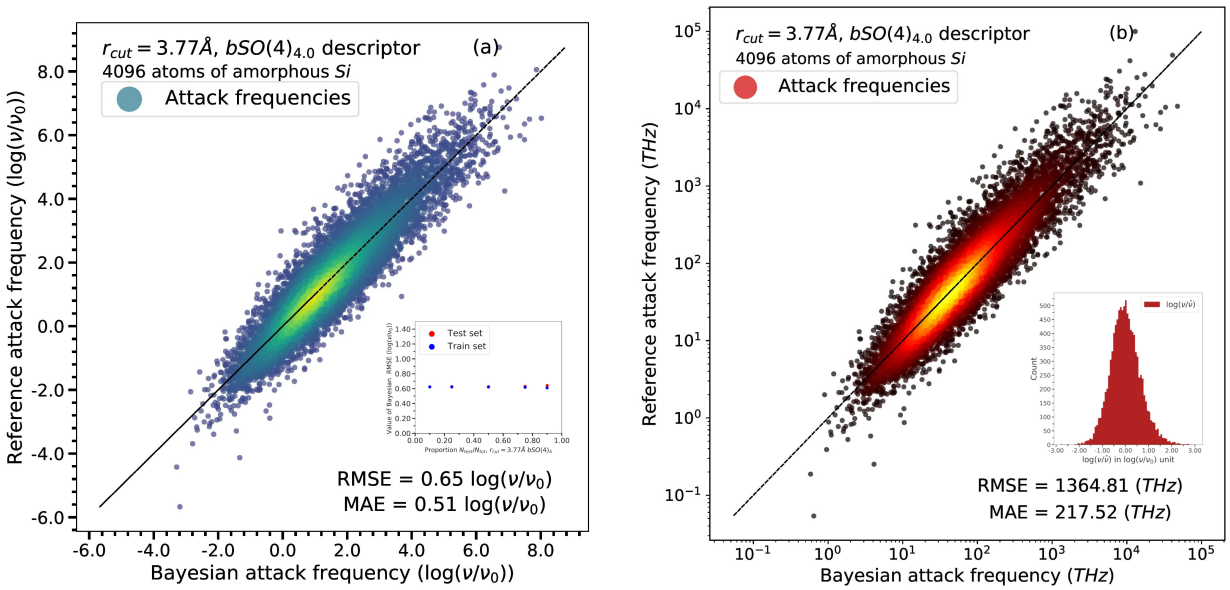

FIG. 3: Illustration of the linear model to adjust the (log) attack frequencies of the amorphous Si database, where $\nu_0 = 1$ THz. The color gradient represents the data distribution, with yellow corresponding to dense data zones. Figure 3.(a) shows the regression of the logarithm of attack frequencies; the results of the train/test validation procedure are presented in the inset. The values of statistical indicators remain stable, even for a large proportion of validation sets. Figure 3.(b) emphasizes the results of attack frequencies. The logarithm of noise $\log(\nu/\hat{\nu})$ ($\hat{\nu}$ is the frequency estimation using the surrogate model of the direct computed frequency ν from ML model) is presented as inset of Figure 3.(b) and follows a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 0.653$.

model for the activation entropy and the activation energy share the same descriptor space.

First, let's define the proportionality between the value of energy barrier $\Delta E_{\mathcal{E},m\to s}$ and the logarithm of attempt frequency $\log\left(\nu^*_{\mathcal{E},ms}\right)$. We define two possible statistical correlations: (i) the direct correlation of activation energy-entropy (DCAEE) for a particular event, and (ii) the enthalpy-entropy compensation (EEC) law based on averaging many activated events. In the case of DCAEE the correlation can be expressed in simple terms for event i:

$$\log(\nu_i/\nu_0) = \gamma \Delta E_i + \log(\nu_\gamma/\nu_0), \tag{28}$$

where γ in eV^{-1} and $\log(\nu_\gamma/\nu_0)$ are parameters that define DCAEE correlation. As we have mentioned before, this law is observed in some simple thermally activated events such as metal conductivity [110], diffusion of adatoms on metallic surfaces [106, 107], dislocation glide in Zr [108] etc. The other type of correlation is a marginal proportionality. This average EEC relation can be expressed as:

$$\mathbb{E}\left[\log(\nu(\Delta E)/\nu_0)|\Delta E\right] = \gamma^* \Delta E + \log(\nu^*_\gamma/\nu_0), \tag{29}$$

where γ^* in eV^{-1} and $\log(\nu^*_\gamma/\nu_0)$ are parameters. We denote by $\mathbb{E}\left[\log(\nu(\Delta E)/\nu_0)|\Delta E\right]$ the average value of $\log(\nu(\Delta E)/\nu_0)$ for each configuration whose associated energy barrier is between ΔE and $\Delta E + \delta\epsilon$, where $\delta\epsilon$ is the width of the energy bin. This expression has been used by Gelin *et al.* [30] for EEC in thermally activated systems in materials science. We note that the DCAEE law implies its marginal EEC definition. These correlations have practical consequences being a "physical" surrogate models in order to estimate the prefactor of some transitions without doing any computations or experiments. In the following, we give a new statistical insight of the DCAEE compensation law by representing the energy barriers and attack frequencies in the same descriptor space and using the same surrogate model with just different parameterisations.

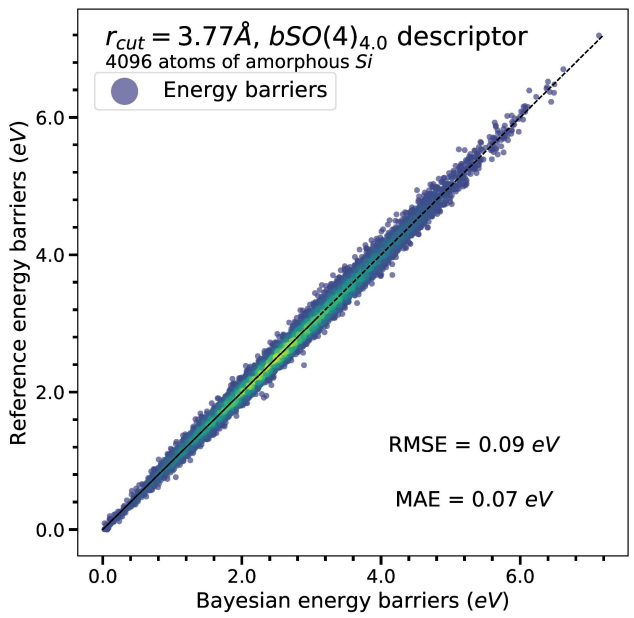

FIG. 4: The correlation between the predicted and computed energy barrier ΔE using the linear regression between descriptors and energy barriers on the amorphous Si database. The relative accuracy of the linear regression model is better for barriers than for the log of prefactors (a RMSE of 0.09 eV over a range of 6 eV).

Using the formalism of the current surrogate model, we built a linear regression model for energy barriers, using the relation:

$$\Delta E_{\mathcal{E},m\to s} = \underline{w}_2 \cdot \left(\underline{D}_{\mathcal{E},m} \oplus \underline{D}_{\mathcal{E},s}\right), \tag{30}$$

where $\underline{D}_{\mathcal{E},m/s} = \sum_{d\in\mathcal{E},m/s} \underline{D}^d \in \mathbb{R}^D$ is the total descriptor vector for $\mathcal{E}, m$ or $\mathcal{E}, s$. The weight vector $\underline{w}_2 \in \mathbb{R}^{1\times 2\mathcal{D}}$ is parametrized using the same database of amorphous Si with the particularity that the target property is now the barrier energy of the event. The corresponding energy barriers have been calculated using the same `ARTn` exploration. For the mapping of the atomic configuration into the descriptor space we have used the same descriptor $bSO(4)$ with the same parameters i.e. $j_{\max} = 4.0$ and the cut-off radius $r_{\text{cut}} = 3.77$ Å . Regression results are presented in Figure 4. The surrogate energy model given by Eq. (30) provides a RMSE of 0.09 eV. Interestingly, the regression is more accurate for the energy barrier than for the log attack frequency. In [41], it is shown that the accuracy performances can be explained by the nature of the force fields. This source of errors for surrogate models for the migration energy or the attack frequency has exactly the same origins. With these two surrogate models for the energy and the attack frequency we have the appropriate tools to investigate the DCAEE and EEC "compensation laws".

Firstly, we investigate the ability of the present surrogate model to recover the direct EEC law from [30]. We emphasize in Figure 5 the correlation between the value of energy barrier ΔE and the corresponding attack frequency, in terms of $\log(\nu/\nu_0)$, using direct and machine learning surrogate approaches. We observe a clear linear relation between ΔE and $\log(\nu/\nu_0)$. In Figure 5.(a) the blue spots emphasize the direct atomistic calculations and in Fig 5.(b) the red spots are the predictions of the current surrogate model. Regression models for both datasets are given in the inset in figure 5. The correlations between ΔE and $\log(\nu/\nu_0)$ for the direct method and the surrogate model are quantitatively very close. The current correlation coefficient is defined by the following equation:

$$r(\log(\nu/\nu_0), \Delta E) = \frac{\mathbb{C}[\log(\nu/\nu_0), \Delta E]}{\sqrt{\mathbb{V}[\log(\nu/\nu_0)]\mathbb{V}[\Delta E]}} \tag{31}$$

where $\mathbb{C}[\log(\nu/\nu_0), \Delta E]$ is the covariance between $\log(\nu/\nu_0)$ and ΔE. $\mathbb{V}$ is the variance of the corresponding observables. The correlation coefficient is a quantitative measurement of whether there is a linear relation between two quantities. If the DCAEE law is exact, it implies this $r(\log(\nu/\nu_0), \Delta E) = 1$. The correlation coefficient is 0.61 for the database and 0.65 for the data predicted by the surrogate model. The current surrogate machine learning approach is accurate enough in order to reconstruct the correlation between the two observables.

Here, we go further. Are there particular conditions of realisation of DCAEE law that can be characterized in the descriptor space? Let's consider two distinct events $\mathcal{E}^1$ and $\mathcal{E}^2$ and the associated attempt frequencies $\nu^*_{\mathcal{E}^1,m\to s}$ and $\nu^*_{\mathcal{E}^2,m\to s}$ such as $\ln\left(\nu^*_{\mathcal{E}^1,m\to s}\right) = \alpha \ln\left(\nu^*_{\mathcal{E}^2,m\to s}\right)$, Eq. (27) implies:

$$\underline{w}_1 \cdot (\underline{D}_{\mathcal{E}^1,m} \oplus \underline{D}_{\mathcal{E}^1,s} - \alpha \underline{D}_{\mathcal{E}^2,m} \oplus \underline{D}_{\mathcal{E}^2,s}) = 0, \tag{32}$$

where $\underline{D}_{\mathcal{E}^1,m}, \underline{D}_{\mathcal{E}^1,s}$ and $\underline{D}_{\mathcal{E}^2,m}, \underline{D}_{\mathcal{E}^2,s}$ are the descriptors of minimum and saddle point for the two events $\mathcal{E}^1$ and $\mathcal{E}^2$, respectively. $\underline{w}_1$ is the weight vector introduced by Eq. (27). If we set $\Delta E_i - \Delta E_0 = \gamma \log(\nu_i/\nu_0)$ and $\Delta E_j - \Delta E_0 = \gamma \log(\nu_j/\nu_0)$, so we can deduce from equation (28) $\log(\nu_i/\nu_0) = \frac{\Delta E_i - \Delta E_0}{\Delta E_j - \Delta E_0} \log(\nu_j/\nu_0)$. If there exists a DCAEE relation then we can build the following relation based on Eq. (32) such as:

$$\underline{w}_1 \cdot \left(\underline{D}_{\mathcal{E}^i,m} \oplus \underline{D}_{\mathcal{E}^i,s} - \frac{\Delta E_i - \Delta E_0}{\Delta E_j - \Delta E_0} \underline{D}_{\mathcal{E}^j,m} \oplus \underline{D}_{\mathcal{E}^j,s}\right) = 0, \tag{33}$$

where $\mathcal{E}^i$, $\underline{D}_{\mathcal{E}^i,m} \oplus \underline{D}_{\mathcal{E}^i,s}$ are the events and the associated descriptors for each event $\mathcal{E}^i$ of the Si amorphous database. $\underline{w}_1$ is the weight vector defined by Eq. (27). This orthogonality relation is valid if the DCAEE law defined by equation (28) is exact. As shown in Figure 5 the DCAEE relation (28) is qualitatively true but implies

an imperfect correlation i.e. 0.61 instead of 1.0 for the perfect theoretical correlation. In order to quantify the notion of orthogonality in the descriptor space, given by Eq. (33), for the realistic case of imperfect correlations of DCAEE law, we introduce the following vectorial quantity: $\underline{\mathfrak{D}}^{ij} = \underline{D}_{\mathcal{E}^i,m} \oplus \underline{D}_{\mathcal{E}^i,s} - \frac{\Delta E_i - \Delta E_0}{\Delta E_j - \Delta E_0} \underline{D}_{\mathcal{E}^j,m} \oplus \underline{D}_{\mathcal{E}^j,s}$. In the end, we focus on the following ratio:

$$\kappa_{ij} = \frac{\left|\underline{w}_1 \cdot \underline{\mathfrak{D}}^{ij}\right|}{\left|\underline{w}_1 \cdot \left(\underline{D}_{\mathcal{E}^i,m} \oplus \underline{D}_{\mathcal{E}^i,s} + \underline{D}_{\mathcal{E}^j,m} \oplus \underline{D}_{\mathcal{E}^j,s}\right)\right|}. \quad (34)$$

If DCAEE law is valid, the following ratio κ_{ij} should be as small as much as possible. For the amorphous Si dataset we can compute the average ratio:

$$\langle \kappa \rangle \equiv \frac{1}{M^2} \sum_{i,j}^{M,M} \kappa_{ij} \ll 1, \quad (35)$$

where M is the number of events. The average value of κ over the entire database is equal to $\langle \kappa \rangle = 0.11$, which is low value. This suggests a weak DCAEE law for this particular database.

Furthermore, for the same amorphous Si database we investigate the variance of the marginal correlation of the EEC law. In order to compute the numerical value of bin average of the $\mathbb{E}\left[\log(\nu/\nu_0)|\Delta E\right]$, given by Eq. (29), the barrier energies domain ΔE is split into 50 uniform bins, each of them having the width $\delta\epsilon = 0.144$ eV. Consequently, $\mathbb{E}\left[\log(\nu/\nu_0)|\Delta E\right]$ is computed with the average of $\log(\nu/\nu_0)$ in each bin of the energy domain. The analysis is performed, over the Si database, for brute and predicted data from our surrogate model. For both we compute the marginal variance $\mathbb{V}\left[\log(\nu/\nu_0)|\Delta E\right]$. The marginal variance represents the width of the intrinsic stochastic noise over $\log(\nu/\nu_0)$ for a bin in $[\Delta E, \Delta E + \delta\epsilon]$. A constant marginal variance implies constant intrinsic noise for all energy bins. Consequently, constant variance over the entire database indicates that correlation between $\log(\nu/\nu_0)$ and ΔE is blurred by a noise with the same origin e.g. induced by the same physical phenomena over the entire range of the energy barriers domain ΔE. Moreover, marginal variance $\mathbb{V}\left[\log(\nu/\nu_0)|\Delta E\right]$ is a quantitative indicator of the DCAEE law's validity. From this perspective, the DCAEE is a particular case of marginal law with $\mathbb{V}\left[\log(\nu/\nu_0)|\Delta E\right] = 0$. The less is the variance, the better is the accuracy of DCAEE to predict $\log(\nu/\nu_0)$ from ΔE.

Results of this analysis are given in Figures 5.(a) and (b) emphasize the correlation between $\log(\nu/\nu_0)$ and ΔE for the Si amorphous database employing direct and surrogate model, respectively. Grey points indicate the marginal observable $\mathbb{E}\left[\log(\nu/\nu_0)|\Delta E\right]$ and his estimated standard deviation. The ajusted marginal EEC from equation (29) is given as an inset for each figures. Average EEC models are very similar for direct and surrogate models. Figure 5.(c) and Figure 5.(d) draw the marginal variance $\mathbb{V}\left[\log(\nu/\nu_0)|\Delta E\right]$ for direct and predicted data, respectively. For these figures we also give the estimated value of standard deviation variance $2\sigma\left[\mathbb{V}\left[\log(\nu/\nu_0)|\Delta E\right]\right]$. This indicator allows to quantify the uncertainty of the marginal variance $\mathbb{V}\left[\log(\nu/\nu_0)|\Delta E\right]$. Low deviation of marginal variance implies that DCAEE law could be extended depending only on $\mathbb{V}\left[\log(\nu/\nu_0)|\Delta E\right]$. From Figure 5.(c) and Figure 5.(d) it can be noted that the marginal variance of $\log(\nu/\nu_0)$ remains almost constant for every bin of barrier domain ΔE. Moreover, the marginal variance $\sigma[\log(\nu/\nu_0)|\Delta E]^2 = \mathbb{V}\left[\log(\nu/\nu_0)|\Delta E\right]$ is quite similar between direct and predicted data. $2\sigma\left[\mathbb{V}\left[\log(\nu/\nu_0)|\Delta E\right]\right]$ slightly depends on the value of ΔE. Consequently, the DCAEE relation can be extended by adding only a dependency in $\mathbb{V}\left[\log(\nu/\nu_0)|\Delta E\right]$. If we suppose that all events present in the Si database are independent, we can give a simple stochastic reformulation of the DCAEE law, Eq. (28), by assuming that the stochastic noise follows a normal distribution:

$$\log(\nu/\nu_0) = \gamma \Delta E + \log(\nu_\gamma/\nu_0) + \mathcal{N}(\sigma^2[\log(\nu/\nu_0)|\Delta E]), \quad (36)$$

where $\mathcal{N}(\sigma^2[\log(\nu/\nu_0)|\Delta E])$ is a centered normal distribution of standard deviation equal to $\sigma[\log(\nu/\nu_0)|\Delta E]$. In the limit $\sigma[\log(\nu/\nu_0)|\Delta E] \to 0$, we find the DCAEE relation. Combining the two perspectives of the compensation law, direct correlation and marginal, allows us to give a more general formulation of the enthalpy-entropy compensation. This formulation is valid for simple (e.g. small point defects in crystalline lattice) and disordered systems (such as the present amorphous system).

The present statistical analysis within the descriptors formulation of the enthalphy-entropy correlations gives two important information. Firstly, by using two surrogate models - one for barrier energy and one for attack frequency - which underlay the same descriptor space, we are able to recover the statistical correlations of the EEC law given by direct calculation of Gelin *et al.* [30]. The results, provided by our linear surrogate models over an amorphous Si database, emphasize the same correlations with the direct HA TST calculations and are solely based on geometrical considerations. Our statistical analysis of the compensation law underlines the ability of the present linear models to capture the complex information about the potential surface of the energy landscape. Secondly, the geometrical information that feed the surrogate models is local, i.e. the descriptors on which is based the linear models encode the geometric structure of atoms within a cut-off distance around a central atom. It is interesting to note that our surrogate model is able to reconstruct well the enthalpy-entropy compensation law solely from this local geometric information. This means that the current surrogate models enable the reconstruction of harmonic vibration quantities, which are diffuse by definition. The ability of the local surrogate model to reconstruct non-local components of vibrational quantities implies a particular structuration of the data in the

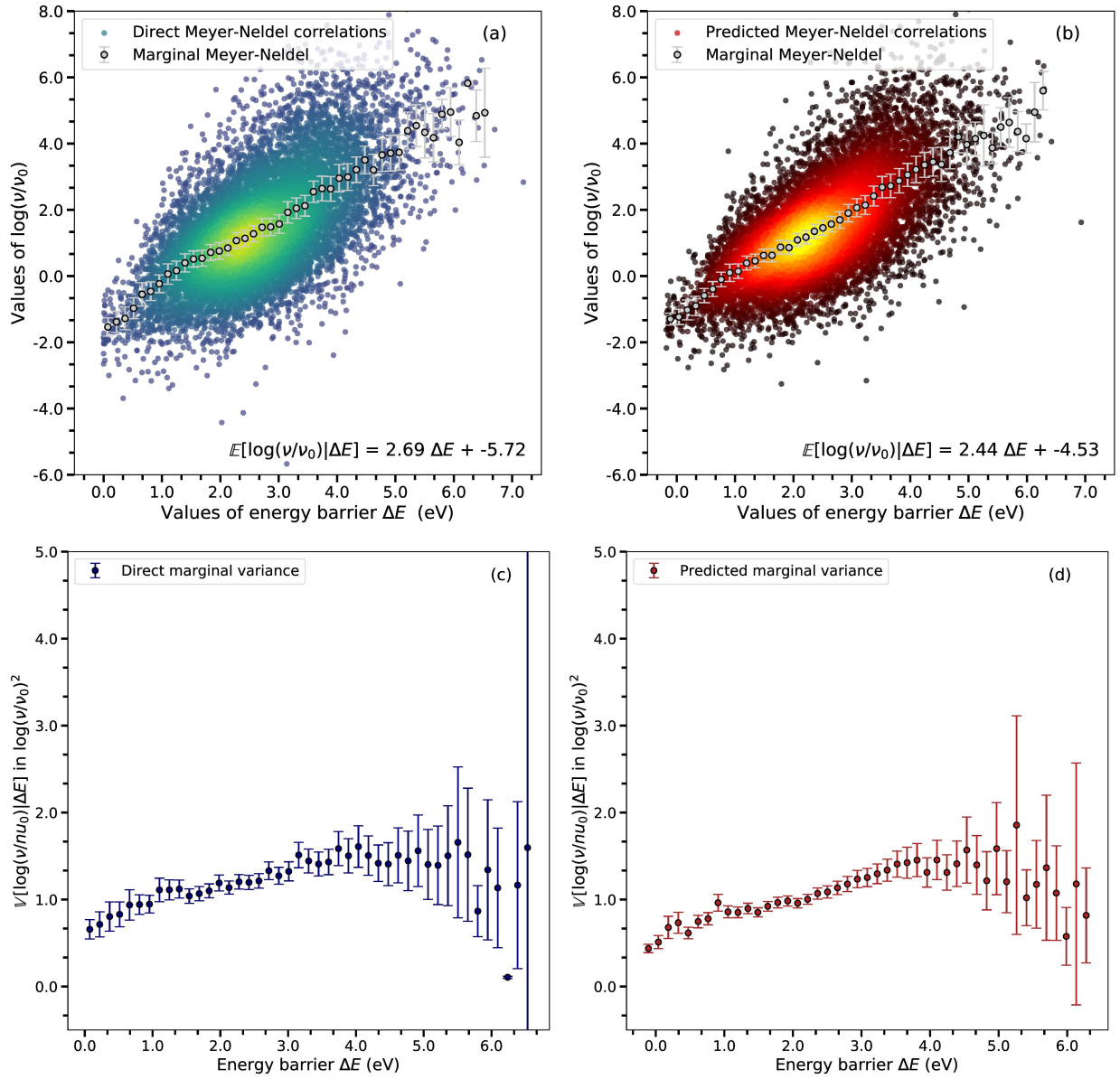

FIG. 5: Drawing the enthalpy-entropy compensation relation for the Si amorphous database. Figure 5.(a) shows values of ΔE and $\log(\nu/\nu_0)$ computed with `ARTn` method. Figure 5.(b) shows predicted values of ΔE and $\log(\nu/\nu_0)$ with linear model. The color gradient represents data distribution; yellow corresponds to dense data zones for both types of points. Adjusted EEC relations, following Eq. (29), for both direct and surrogate models are emphasized in (a) and (b) by white points with black contour. Both models, the direct and the surrogate, have distributions with very close correlation indicators. Marginal variance distribution for both dataset is presented in Figure 5.(c) and Figure 5.(d) for direct and surrogate data respectively. The marginal variance is quantitavely almost the same for both dataset.

descriptor space. In fact, data in the descriptor space seems to present a simple topology that allows for new formulations, in terms of elementary geometry in high dimensional space, of non-trivial physical laws in the configuration space, such as the present Meyer-Neldel correlation law.

VI. CONCLUSIONS AND PERSPECTIVES

The current study explores the limits of machine learning surrogate models for the vibrational properties of solids that make a direct link between atomic geometry and vibrational observables of interest. This current approach is developed using the traditional harmonic approximation, which allows vibrational properties to be

exactly decomposed into local contributions around each atom. This feature allowed the model to be easily integrated within machine learning frameworks. Traditional methods based on harmonic approximation use for the evaluation of the vibrational entropy the derivatives of the interatomic potential energy surface in particular points of the potential energy surface i.e. minima or saddle points. Consequently, here we treat only the information from particular instances of the phase space such as minima or saddle points of the first-order.

Within the framework of harmonic approximation and transition state theory we proposed surrogate models for the vibrational entropy that account for (i) variations under applied hydrostatic strains (ii) attack frequencies for thermally activated events. The traditional evaluation of the system's Hessian and its diagonalization is bypassed; the only need is to provide reliable atomic positions for the minimum / saddle point configurations. In the present study we have used the `ARTn` method in order to provide those configurations.

For the present surrogate model for formation vibrational entropy, with or without strain, compared to the previous work, we have extended the dimension of atomic descriptors through quadratic coupling between various components of the original atomic descriptor. We have deduced the correlation between the new atomic descriptors and the vibrational observables within the framework of the quadratic surrogate model (EQML). This EQML model is more accurate than previous linear models and it is not too much less transferable. We keep the balance between accuracy and transferability of the fit by the parametrization procedure of the quadratic fit, which is preconditioned by the linear fit. Furthermore, the EQML model has very good transferability in the sense of train/test procedure. Over a database containing various point defects in α-Fe we have numerically demonstrated that this approach has better accuracy for homogeneous, isotropic deformations from -1% to 3% strain than the previous linear model introduced in [41].

We have replaced the local atomic descriptors with the appropriate descriptor of the activated event, i.e., the sequence minimum - saddle point, for estimating the attack frequency of thermally activated processes. For this purpose, the surrogate model uses structural information, through descriptor projection, of the geometrical configurations of the minimum and of the saddle configuration. The relation descriptors - target local observables is based on the linear model Eq. (14). However, this choice is not restrictive and the model framework can be extended to higher orders, such as quadratic EQML. This new surrogate model has been tested on a database of activated events in amorphous Si [30] and provides good accuracy over many orders of magnitude (6 orders of magnitude) of the attack frequency. This model has a good transferability with very stable train / test learning curves.

Our study shows that it is possible to adjust the formation entropy of defects and logarithm of the attack frequency of their activated events only with $\mathcal{O}(N)$ numerical estimation. The present workflow avoids the time-consuming evaluation of system's dynamical matrix ($\mathcal{O}(N^2)$) and its spectrum ($\mathcal{O}(N^3)$). The current efficient solution opens many avenues in the field of on-the-fly exploration of complex energetic landscapes, for example using semi-automatic atomistic frameworks such as lattice or off-lattice and relaxed Kinetic Monte Carlo [11, 83–89, 94].

Finally, employing the framework of machine learning surrogate models, we have proposed an insight into the statistical analysis of enthalpy-entropy compensation law - a non-trivial conjecture observed in several materials [104, 111]. This law states that, for a given transition event, the link between the magnitude of the potential energy function (the value of ΔE) and its curvature (the value of $\log(\nu/\nu_0)$). Here we give a statistical formulation of the compensation energy - frequency law. Sharp statistical analysis emphasizes that we are able to reproduce the same correlation as the direct calculations by using two linear surrogate models for the barrier energy and the attack frequency of activated events. Moreover, by tackling the DCAEE law in the descriptor space, we have provided a geometrical insight into the conditions of realisation of this conjecture, which are quantified by an orthogonality relation. The present formulation requires more investigation in other systems and opens many perspectives for further investigation.

VII. CODES AND DATA

The `ARTn` package and databases of amorphous silicon events are available upon reasonable request to Normand Mousseau (normand.mousseau@umontreal.ca). The `Milady` package is open source software under ASL license and can be downloaded at `https://ai-atoms.github.io/milady/`.

ACKNOWLEDGMENTS

The authors sincerely thank S. Gelin for providing the amorphous Si database and for his valuable suggestions concerning the manuscript. 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 and Grant Agreement No. 900018 from the ENTENTE project. TDS gratefully recognizes support from the Agence Nationale de Recherche, via the MEMOPAS project ANR-19-CE46-0006-1, and access to the HPC resources of IDRIS under the allocation A0090910965 attributed by GENCI. CL and MCM acknowledge the support from GENCI - (CINES/CCRT) computer center under Grant No. A0130906973.

[1] L. Proville, D. Rodney, and M. C. Marinica, Nat. Mater. **11**, 845-849 (2012).
[2] D. Caillard and J. L. Martin, *Thermally Activated Mechanisms in Crystal Plasticity*, Pergamon Materials Series (Elsevier Science, 2003).
[3] 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 **440**, 1174 (2006).
[4] D. Mordehai, E. Clouet, M. Fivel, and M. Verdier, Phil. Mag. **88**, 899 (2008).
[5] M. Loretto and R. Smallman, *Defect Analysis in Electron Microscopy*, Science paperbacks (Chapman and Hall, 1975).
[6] D. Hull and D. J. Bacon, *Introduction to Dislocations* (Butterworth-Heinemann, Amsterdam, 2011).
[7] V. Vítek, Phil. Mag. **18**, 773 (1968).
[8] M. Kiritani, Mater. Chem. Phys. **50**, 133 (1997).
[9] M.-C. Marinica, F. Willaime, and N. Mousseau, Phys. Rev. B **83**, 094119 (2011).
[10] M.-C. Marinica, F. Willaime, and J.-P. Crocombette, Phys. Rev. Lett. **108**, 025501 (2012).
[11] F. El-Mellouhi, N. Mousseau, and L. J. Lewis, Phys. Rev. B **78**, 153202 (2008).
[12] L. K. Béland, P. Brommer, F. El-Mellouhi, J.-F. Joly, and N. Mousseau, Phys. Rev. E **84**, 046704 (2011).
[13] O. A. Restrepo, C. S. Becquart, F. El-Mellouhi, O. Bouhali, and N. Mousseau, Acta Mater. **136**, 303 (2017).
[14] A. Samanta and W. E, J. Chem. Phys. **136**, 124104 (2012).
[15] M. Trochet, L. K. Béland, J.-F. Joly, P. Brommer, and N. Mousseau, Phys. Rev. B **91**, 224106 (2015).
[16] G. Q. Xu and M. J. Demkowicz, Phys. Rev. Lett. **111**, 145501 (2013).
[17] T. D. Swinburne and D. Perez, Phys. Rev. Mat. **2**, 053802 (2018).
[18] T. D. Swinburne and D. Perez, Npj Comput. Mater. **6**, 190 (2020).
[19] D. J. Wales and J. P. Doye, J. Phys. Chem. A **101**, 5111 (1997).
[20] O. M. Becker and M. Karplus, J. Chem. Phys. **106**, 1495 (1997).
[21] S. V. Krivov and M. Karplus, J. Chem. Phys. **117**, 10894 (2002).
[22] J. Braun, M. H. Duong, and C. Ortner, Archive for Rational Mechanics and Analysis **238**, 1413 (2020).
[23] K. S. P. H. Dederichs, R. Zeller, *Point Defects in Metals II, Dynamical Properties and Diffusion Controlled Reactions* (Springer Tracts in Modern Physics, 1980).
[24] R. Malek and N. Mousseau, Phys. Rev. E **62**, 7723–7728 (2000).
[25] D. J. Wales, *Energy Landscapes*, edited by C. U. Press (Cambridge, 2003).
[26] G. Henkelman, B. P. Uberuaga, and H. Jonsson, J. Chem. Phys. **113**, 9901 (2000).
[27] G. Henkelman, G. Johannesson, and H. Jonsson, *Theoretical Methods in Condensed Phase Chemistry*, , 269.
[28] G. Henkelman, Annu. Rev. Mater. Res. **47**, 199 (2017).
[29] 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).
[30] S. Gelin, A. Champagne-Ruel, and N. Mousseau, Nat. Commun. **11**, 3977 (2020).
[31] T. Lelièvre, G. Stoltz, and M. Rousset, *Free energy computations: A mathematical perspective* (Imperial College Press, 2010).
[32] G. M. Torrie and J. P. Valleau, Journal of Computational Physics **23**, 187 (1977).
[33] A. Laio and M. Parrinello, Proceedings of the National Academy of Sciences **99**, 12562 (2002).
[34] L. Maragliano and E. Vanden-Eijnden, Chemical Physics letters **446**, 182 (2007).
[35] E. Weinan, W. Ren, and E. Vanden-Eijnden, Phys. Rev. B **66**, 052301 (2002).
[36] T. Lelièvre, M. Rousset, and G. Stoltz, J. Chem. Phys. **126**, 134111 (2007).
[37] L. Bonati and M. Parrinello, Phys. Rev. Lett. **121**, 265701 (2018).
[38] E. Darve, D. Rodríguez-Gómez, and A. Pohorille, J. Chem. Phys. **128**, 144120 (2008).
[39] T. D. Swinburne and M.-C. Marinica, Phys. Rev. Lett. **120**, 135503 (2018).
[40] J. Baima, A. M. Goryaeva, T. D. Swinburne, J.-B. Maillet, M. Nastar, and M.-C. Marinica, Phys. Chem. Chem. Phys. **24**, 23152 (2022).
[41] C. Lapointe, T. D. Swinburne, L. Thiry, S. Mallat, L. Proville, C. S. Becquart, and M.-C. Marinica, Phys. Rev. Mat. **4**, 063802 (2020).
[42] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[43] J. Behler, J. Chem. Phys. **134**, 074106 (2011).
[44] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, J. Chem. Phys. **148**, 241730 (2018).
[45] A. P. Bartók, *Gaussian Approximation Potential : an interatomic potential derived from first principles Quantum Mechanics*, Ph.D. thesis, University of Cambridge (2009).
[46] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).
[47] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[48] A. P. Bartók and G. Csányi, Int. J. Quantum Chem. **115**, 1051 (2015).
[49] K. T. Schutt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, J. Chem. Phys. **148**, 241722 (2018).
[50] A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis, and P. G. Debenedetti, Chem. Phys. Lett. **509**, 1 (2011).
[51] A. M. Goryaeva, J.-B. Maillet, and M.-C. Marinica, Comput. Mater. Sci. **166**, 200 (2019).
[52] A. M. Goryaeva, J. Dérès, C. Lapointe, P. Grigorev, T. D. Swinburne, J. R. Kermode, L. Ventelon, J. Baima, and M.-C. Marinica, Phys. Rev. Mater. **5**, 103803 (2021).
[53] A. Shapeev, Multiscale Model. Sim. **14**, 1153 (2016).
[54] R. Drautz, Phys. Rev. B **99**, 014104 (2019).
[55] R. Drautz, Phys. Rev. B **102**, 024104 (2020).
[56] A. E. A. Allen, G. Dusson, C. Ortner, and G. Csányi, Machine Learning: Science and Technology **2**, 025017 (2021).
[57] C. v. d. Oord, G. Dusson, G. Csányi, and C. Ortner, Machine Learning: Science and Technology **1**, 015004

(2020).
[58] H. Zong, G. Pilania, X. Ding, G. J. Ackland, and T. Lookman, Npj Comput. Mater. **4**, 1 (2018).
[59] J. R. Kermode, A. Gleizer, G. Kovel, L. Pastewka, G. Csányi, D. Sherman, and A. De Vita, Phys. Rev. Lett. **115**, 135501 (2015).
[60] G. Ferré, J.-B. Maillet, and G. Stoltz, J. Chem. Phys. **143**, 104114 (2015).
[61] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, J. Phys. Chem. C **121**, 511 (2017).
[62] E. Cubuk, S. Schoenholz, J. Rieser, B. Malone, J. Rottler, D. Durian, E. Kaxiras, and A. Liu, Phys. Rev. Lett. **114**, 108001 (2015).
[63] F. Bruneval, I. Maliyov, C. Lapointe, and M.-C. Marinica, J. Chem. Theory Comput. **16**, 4399 (2020).
[64] F. Noe and C. Clementi, J. Chem. Theory Comput. **11**, 5002 (2015).
[65] W. F. Reinhart, A. W. Long, M. P. Howard, A. L. Ferguson, and A. Z. Panagiotopoulos, Soft Matter **13**, 4733 (2017).
[66] R. Kakarala, *Phd Thesis: The bispectrum as a source of phase-sensitive invariants for Fourier descriptors: a group-theoretic approach* (Irvine University, 1992).
[67] G. T. Barkema and N. Mousseau, Phys. Rev. Lett. **77**, 4358 (1996).
[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. Béland, D. Caliste, L. Genovese, T. Deutsch, N. Mousseau, and P. Pochet, J. Chem. Phys. **135**, 034102 (2011).
[70] S. Plimpton, Journal Computational Physics **117**, 1 (1995).
[71] M.-C. Marinica and F. Willaime, Solid State Phenom. **129**, 67 (2007).
[72] A. Soulié, F. Bruneval, M.-C. Marinica, S. Murphy, and J.-P. Crocombette, Phys. Rev. Mat. **2**, 083607 (2018).
[73] F. Berthier, J. Creuze, T. Gabard, B. Legrand, M.-C. Marinica, and C. Mottet, Phys. Rev. B **99**, 014108 (2019).
[74] S. A. Etesami and E. Asadi, Journal of Physics and Chemistry of Solids **112**, 61–72 (2018).
[75] A. M. Goryaeva, C. Lapointe, C. Dai, J. Dérès, J.-B. Maillet, and M.-C. Marinica, Nature Commun. **11**, 1 (2020).
[76] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, Journal of Machine Learning Research **12**, 2825 (2011).
[77] A. Chartier and M.-C. Marinica, Acta Mat. **180**, 141 (2019).
[78] K. Arakawa, M.-C. Marinica, S. Fitzgerald, L. Proville, D. Nguyen-Manh, S. L. Dudarev, P.-W. Ma, T. D. Swinburne, A. M. Goryaeva, T. Yamada, *et al.*, Nat. Mater. **19**, 508 (2020).
[79] T. D. Swinburne, D. Kannan, D. J. Sharpe, and D. J. Wales, J. Chem. Phys. **153**, 134115 (2020).
[80] T. Schuler, L. Messina, and M. Nastar, Comput. Mater. Sci. **172**, 109191 (2020).
[81] T. Schuler and M. Nastar, Phys. Rev. B **93**, 224101 (2016).
[82] L. Huang, T. Schuler, and M. Nastar, Phys. Rev. B **100**, 224103 (2019).
[83] T. Jourdan, F. Soisson, E. Clouet, and A. Barbu, Acta Mater. **58**, 3400 (2010).
[84] D. Carpentier, T. Jourdan, P. Terrier, M. Athènes, and Y. Le Bouar, J. Nucl. Mater. **533**, 152068 (2020).
[85] B. Gámez, L. Gámez, C. Ortiz, M. Caturla, and J. Perlado, J. Nucl. Mater. **386-388**, 90 (2009), fusion Reactor Materials.
[86] N. Castin, G. Bonny, A. Bakaev, C. Ortiz, A. Sand, and D. Terentyev, J. Nucl. Mater. **500**, 15 (2018).
[87] C. Domain and C. Becquart, in *Handbook of Materials Modeling*, edited by Y. Chen, E. Homer, and C. A. Schuh (Springer Nature, Switzerland, 2018).
[88] V. Jansson, L. Malerba, A. De Backer, C. Becquart, and C. Domain, J. Nucl. Mater. **442**, 218 (2013).
[89] M. Chiapetto, C. S. Becquart, C. Domain, and L. Malerba, Phys. Status Solidi C **12**, 20.
[90] C.-C. Fu, J. D. Torre, F. Willaime, J.-L. Bocquet, and A. Barbu, Nat. Mater. **4**, 68 (2005).
[91] M. de Koning, W. Cai, B. Sadigh, T. Oppelstrup, M. H. Kalos, and V. V. Bulatov, J. Chem. Phys. **122**, 074103 (2005).
[92] T. Oppelstrup, V. V. Bulatov, A. Donev, M. H. Kalos, G. H. Gilmer, and B. Sadigh, Phys. Rev. E **80**, 066701 (2009).
[93] W. Cai, M. H. Kalos, M. de Koning, and V. V. Bulatov, Phys. Rev. E **66**, 046703 (2002).
[94] N. Mousseau, L. K. Béland, P. Brommer, F. El-Mellouhi, J.-F. Joly, G. K. N'Tsouaglo, O. Restrepo, and M. Trochet, Comp. Mater. Sci. **100**, 111 (2015), special Issue on Advanced Simulation Methods.
[95] P. Hänggi, P. Talkner, and M. Borkovec, Reviews of Modern Physics **62**, 251 (1990).
[96] C. Dellago, P. Bolhuis, and P. L. Geissler, Adv. Chem. Phys. **1 123**, 1-78 (2002).
[97] G. H. Vineyard, J. Phys. Chem. Solids **3**, 121 (1957).
[98] A. Jay, C. Huet, N. Salles, M. Gunde, L. Martin-Samos, N. Richard, G. Landa, V. Goiffon, S. De Gironcoli, A. Hémeryck, *et al.*, Journal of Chemical Theory and Computation **16**, 6726 (2020).
[99] L. K. Béland, Y. Anahory, D. Smeets, M. Guihard, P. Brommer, J.-F. Joly, J.-C. Pothier, L. J. Lewis, N. Mousseau, and F. Schiettekatte, Phys. Rev. Lett. **111**, 105502 (2013).
[100] R. Vink, G. Barkema, W. van der Weg, and N. Mousseau, Journal of Non-Crystalline Solids **282**, 248 (2001).
[101] E. Bitzek, P. Koskinen, F. Gähler, M. Moseler, and P. Gumbsch, Phys. Rev. Lett. **97**, 170201 (2006).
[102] F. H. Constable and W. J. Pope, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character **108**, 355 (1925).
[103] W. Meyer and H. Nedel, Z. Tech. Phys. (Leipzig) **12**, 588 (1937).
[104] J. Philibert, in *Diffusion in Solids - Past, Present and Future*, Defect and Diffusion Forum, Vol. 249 (Trans Tech Publications Ltd, 2006) pp. 61–72.
[105] A. Yelon, B. Movaghar, and R. S. Crandall, Rep. Prog. Phys. **69**, 1145 (2006).
[106] G. Boisvert, L. J. Lewis, and A. Yelon, Phys. Rev. Lett. **75**, 469 (1995).
[107] M.-C. Marinica, C. Barreteau, D. Spanjaard, and M.-C. Desjonqueres, Phys. Rev. B **72**, 115402 (2005).
[108] E. Maras and E. Clouet, Acta Materialia **223**, 117398 (2022), arXiv: 2112.04284.
[109] D. Emin, Phys. Rev. Lett. **32**, 303 (1974).

[110] Y. Lubianiker and I. Balberg, Phys. Status Solidi B **205**, 119 (1998).
[111] L. Shcherbak, O. Kopach, P. Fochuk, A. E. Bolotnikov, and R. B. James, J. Phase Equilibria Diffus. **36**, 99–109 (2015).