Capabilities and limits of autoencoders for extracting collective variables in atomistic materials science

Jacopo Baima, Alexandra Goryaeva, Thomas Swinburne, Jean-Bernard Maillet, Maylise Nastar, Mihai-Cosmin Marinica

PCCP

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Capabilities and limits of autoencoders for extracting collective variables in atomistic materials science

Jacopo Baima,*[a] Alexandra M. Gorayeva,[a], Thomas D. Swinburne,[b] J.-B. Maillet,[c] M. Nastar[a] and M.-C. Marinica*[a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Free energy calculations in materials science are routinely hindered by the need to provide reaction coordinates that can meaningfully partition atomic configuration space, a prerequisite for most enhanced sampling approaches. Recent studies on molecular systems have highlighted the possibility of constructing appropriate collective variables directly from atomic motions through deep learning techniques. Here we extend this class of approaches to condensed matter problems, for which we encode the finite temperature collective variable by an iterative procedure starting from 0 K features of the energy landscape i.e. activation events or migration mechanisms given by a minimum - saddle point - minimum sequence. We employ the autoencoder neural networks in order to build a scalar collective variable for use with the adaptive biasing force method. Particular attention is given to design choices required for application to crystalline systems with defects, including the filtering of thermal motions which otherwise dominate the autoencoder input. The machine-learning workflow is tested on body-centered cubic iron and its common defects, such as small vacancy or self-interstitial clusters and screw dislocations. For localized defects, excellent collective variables as well as derivatives, necessary for free energy sampling, are systematically obtained. However, the approach has a limited accuracy when dealing with reaction coordinates that include atomic displacements of a magnitude comparable to thermal motions, e.g. the ones produced by the long-range elastic field of dislocations. We then combine the extraction of collective variables by autoencoders with an adaptive biasing force free energy method based on Bayesian inference. Using a vacancy migration as an example, we demonstrate the performance of coupling these two approaches for simultaneous discovery of reaction coordinates and free energy sampling in systems with localized defects.

1 Introduction

Thermally activated processes are ubiquitous in material science. They control a wide range of phenomena, such as defect migration or recombination [1], plasticity [2], and phase transitions [3]. Simulating these processes directly using molecular dynamics (MD) is highly inefficient, owing to the orders of magnitudes of difference between the time-scale of the diffusion processes and the one of atomic vibrations. As alternative, the transformation processes can be modeled through other frameworks, like transition state theory [4–6], which require knowledge of the energy barriers or profiles. Indeed, a wealth of information on such processes can be obtained from the free energy profile sampled along some suitable collective variable (CV) $\xi(\mathbf{q})$, i.e. a smooth function defined between the initial and final states of the system that projects the $3N$ dimensional phase space (N being the number of atoms in the system) into a lower d-dimensional space. For this purpose, the collective variable function must closely approximate the true reaction coordinate. In practice, this requires separating the initial and final states of the process as well as the transition state and any other intermediaries, and partitioning the configuration space as evenly as possible in order to allow proper sampling of any features of the free energy profile. [5,7,8]

If the reaction coordinate is known, a number of methods are available to integrate the free energy profile, including Monte Carlo approaches and enhanced MD [5,6,9]. The latter methods add a bias to the system dynamics in order to flatten the free energy landscape along the chosen collective variables and, thus, to accelerate its exploration. The bias can take different forms, depending on the framework: in the case of umbrella sampling [10,11], metadynamics [12,13], and other adaptive biasing poten-

[a] Université Paris-Saclay, CEA, Service de Recherches de Métallurgie Physique, Gif-sur-Yvette 91191, France
[b] Aix-Marseille Université, CNRS, CINaM UMR 7325, Campus de Luminy, 13288 Marseille, France
[c] Université Paris-Saclay, CEA, LMCE, 91680 Bruyères-le-Châtel, France
* E-mails: jacopo.baima@cea.fr, mihai-cosmin.marinica@cea.fr

tial approaches[14,15], the bias is added to the potential energy, while Adaptive Biasing Force (ABF) methods[16–18] add a force to the system dynamics, which counteracts the average force acting along the CVs. Besides, while some methods only require the knowledge of the $\xi(\mathbf{q})$ function and its derivatives $\nabla_{\mathbf{q}}\xi(\mathbf{q})$[12,17,19], other approaches such as the original formulation of ABF also require a subset of second derivatives of the CV[16].

The most of well-known ABF approaches that require the first CV derivatives only, are the extended (eABF) version and its numerous variants[17,18,20,21]. These methods sidestep the need to compute the derivative of the CV Jacobian by applying the biasing force to a fictitious variable coupled to the CV function by a spring potential. This approach has been very successful despite its requirement of an integration of the additional dynamics for the extended variable, which is often sensitive to the choice of parameters[20,22]. More recently, a probabilistic reformulation of eABF has been proposed, the so-called Bayesian ABF (bABF)[23,24]. This approach substitutes the dynamics for the extended variable with a statistical distribution of its values given the atomic positions. As a consequence, most of the specific difficulties of eABF are avoided, and the resulting approach can be shown to minimize the variance of the estimated thermodynamic quantities in the limit of long simulations[25].

Despite the ever-growing progress in the development of CV methods for free energy integration, the preliminary knowledge of an appropriate reaction coordinate remains their main shortcoming. Although suitable CVs are relatively easy to obtain for some simple systems[26,27], the same does not hold for complex material science problems, such as dislocation glide[19,28], or complex transformations of defect clusters[29,30], or phase transitions including crystallization and amorphization[31].

In this context, machine learning techniques for dimensionality reduction offer a promising way to extract CVs directly from atomistic simulations, ideally even in absence of any previous knowledge of the kinetic path of the system.[31–33]. In this study, we rely on AutoEncoders (AEs)[34,35]. These are neural networks (NNs) trained to generate a low-dimensional representation of the input data and vice versa, i.e., to reconstruct the original input from its low dimensional representation. In order to encode the input data into the latent space while minimizing the reconstruction error, the NN has to learn a representation containing the most relevant information describing the training set. Additional terms can be included to the loss function in order to impose the desired properties on the latent space, e.g. continuity[36,37] or orthogonality[38]. In the case of Variational autoencoders (VAEs)[36,39], a continuous statistical distribution is imposed on the latent space, which makes them very successful as generative models[39]. The utility of AEs was originally demonstrated for image processing, including feature extraction[40], denoising[41], and anomaly detection[42].

More recently, a new application of AEs was found in learning the CVs of molecular systems. An iterative approach alternating between umbrella sampling and CV extraction with AEs[43,44] was applied to the modeling of dihedral rotations in small peptides, without using prior knowledge of the system's minima and saddle points. Further extensions based on VAEs [45–47] or time-lagged AEs[47,48] have also been explored. However, the application of the AEs to materials science, is still not straightforward, due to multiple reasons. First, condensed matter problems can involve a very large number of degrees of freedom. The number of weights in an AE network grows at least linearly with the input size, which may lead to overfitting and large computational costs. Second, even in simple cases, free energy surfaces commonly have multiple minima due to translation and permutation symmetries. Finally, some processes, such as the elastic field of dislocation or soft phonons excitations are characterized by delocalized collective variables, which can be difficult to discriminate from thermal noise.

In this paper, we first explore the performance and applicability range of AE neural networks in condensed matter problems, using localised and extended defects in α-Fe as a test case. Second, we discuss the coupling between AE and bABF free energy exploration, as well as the computational choices that allow a practical coupling between the two algorithms. In Sec. 2, we describe the methods employed to learn the CVs, the utility of Principal Component Analysis to reduce the dimension of the AE input, and the coupling with a most recent bABF enhanced sampling approach. The employed computational settings are summarized in Sec. 3. Further, in Sec. 4.1, we analyze the accuracy of AEs for extraction of reaction coordinates in different systems at 0 K from noisy synthetic data. Then, Sec. 4.2 presents the free energy profiles obtained using our bABF coupled with AE for a vacancy migration at different temperatures. Finally, Sec. 5 contains the conclusions of the present study.

2 Learning collective variables

2.1 Autoencoders

Autoencoders are neural networks trained to compress and then reconstruct input data[34,35]. A typical architecture is schematically represented in the Fig. 1. One of the hidden layers has a (much) lower dimension with respect to the input data. This inner space is called latent representation. The first part of the AE, generating the low-dimensional representation, is the encoder, and the second part, reconstructing the output data, is the decoder. The input, $\mathbf{x} \in \mathbb{R}^D$, is passed to the output trough the encoder and decoder function: $\hat{\mathbf{x}}(\mathbf{x}) = f_d \circ f_e(\mathbf{x}) = f_d(f_e(\mathbf{x}))$. with $f_e : \mathbb{R}^D \to \mathbb{R}^d$ with $d < D$, and $f_d : \mathbb{R}^d \to \mathbb{R}^D$. In this work, we use f_e and f_d functions with symmetric architecture with respect to the bottleneck, and each of them contains L layers of dimension $d_1, \dots d_L$. However, it should be noted that this architecture is not a compulsory requirement. The similarity between the output (or reconstruction) $\hat{\mathbf{x}}$ and the input $\mathbf{x}$ is enforced by minimizing the objective function measuring the difference between the two arrays of data. Hereafter, we take the mean square error (MSE) as objective function. As a consequence of this minimization and of the small size of the latent space, the neural network will map the input into a few variables which account for most of the variance of the input data. When the training data are structural configurations extracted from a MD simulation, the latent space will contain the CVs which explain the largest variance of atomic positions. When the path includes at least one transition event for an activated process, these CVs are often perfectly suited for de-

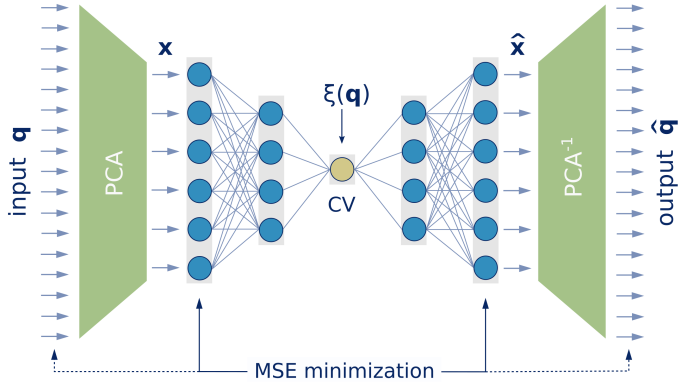

Fig. 1 Schematic illustration of the procedure for learning CV using AE neural network with the number of input channels reduced by PCA preprocessing. The inner part between $\mathbf{x}$ and $\hat{\mathbf{x}}$ is the traditional AE architecture. Linear dimensionality reduction of the AE input $\mathbf{q} \in \mathbb{R}^{3N} \rightarrow \mathbf{x} \in \mathbb{R}^{D}$ and reconstruction $\hat{\mathbf{x}} \rightarrow \hat{\mathbf{q}}$ is ensured using linear PCA encoding / decoding, respectively, as it is detailed in Sec. 2.2.

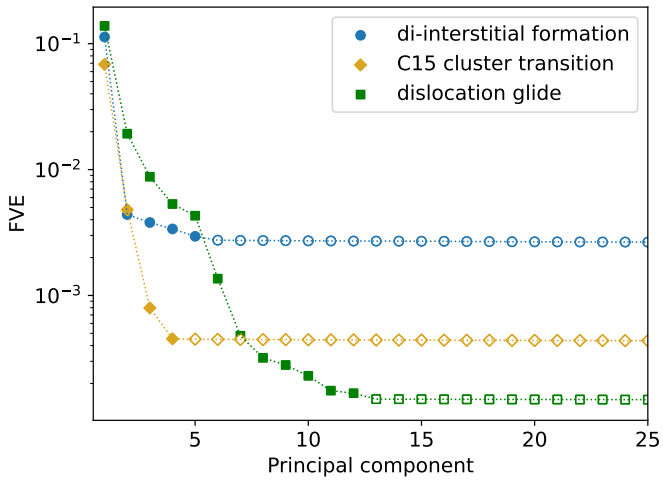

Fig. 2 Fraction of variance explained by the first 30 principal components for three systems discussed in Sec. 4.1. Minimum number of principal components is 5, 4, and 12 for di-interstitials, C15 cluster and dislocations, respectively (filled symbols in the figure).

scribing the process and sampling the relative free energies.[43–45]

In this study we use two AE architectures: (i) the AE with direct encoding of the atomic coordinates $\mathbf{q}$, i.e. $\mathbf{x} = \mathbf{q}$ and $D = 3N$; and (ii) the AE that uses a filtered input $\mathbf{x} \in \mathbb{R}^{D}$ resulting from a linear encoding of the full set of atomic positions, $\mathbf{q} \in \mathbb{R}^{3N}$, in order to reduce the dimensionality of the AE input (i.e. $D < 3N$, as displayed in Fig. 1). The first architecture was used in previous studies for isolated molecules[43,44], whilst the second, more adapted for condensed matter systems, is used in the present study (see next Section 2.2).

2.2 Preparing the input of AE

Using NNs implies a risk of overfitting when the number of weights is comparable with the size of the training dataset. Indeed, an AE network including too many parameters may rapidly learn to trivially reproduce the identity function. Then, the CVs would not reflect the dynamics of the system. In the context of materials science, this issue hinders the studies of large systems. In order to capture nonlinearities in the encoding, the number of neurons in the first hidden layer d_1 is typically chosen to have the same order of magnitude as the number of input channels D, i.e. the number of atomic coordinates. The number of weights to be fitted therefore grows at least as $D \times d_1$, that is approximatively quadratically in D. Therefore, for applications in material science and in general for applications involving a large number of atoms, decreasing the number of input channels is a crucial step in preparing the input data of AE.

To apply AEs to large atomic arrays, we apply linear dimensionality reduction operations to the original coordinate space of the atoms.[46] Linear dimensionality reduction is significantly faster and less prone to overfitting than nonlinear methods such as AEs, but cannot be used to identify a CV which is a nonlinear function of the coordinates. We can, however, use it to project the original coordinate space into a linear subspace which contains the desired CVs. Specifically, we use Principal Component Analysis (PCA), as this algorithm projects the input in the same space as a linear AE[35], i.e. one that minimizes the MSE of the projected input (Fig. 1). This allows the two methods to be combined effectively. In the most common formulation, the principal components are obtained by diagonalizing the sample covariance matrix of the atomic cartesian coordinates, $\mathbf{K} \in \mathbb{R}^{3N \times 3N}$:

$$\mathbf{K} = \sum_{m=1}^{M} \frac{1}{M-1} (\mathbf{q}_m - \overline{\mathbf{q}}) (\mathbf{q}_m - \overline{\mathbf{q}})^{\top} \quad (1)$$

$$\mathbf{K} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^{\top}, \quad (2)$$

where $\mathbf{q}_m \in \mathbb{R}^{3N}$ is a column vector containing the Cartesian coordinates of the m-th atomic configuration in the training dataset; $\overline{\mathbf{q}}$ is the barycenter of the dataset equal to $\frac{1}{M}\sum_{m=1}^{M} \mathbf{q}_m$; $\mathbf{\Lambda}$ is a diagonal matrix, $\mathrm{diag}(\lambda_1, \ldots, \lambda_{3N})$, with eigenvalue sorted in descending order; and the orthogonal matrix $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_{3N})$ contains the associated eigenvectors.

Here, we represent the database in the space span by the most pertinent M vectors, i.e. associated to the first D λ-eigenvalues. $\mathbf{U}_D = (\mathbf{u}_1, \ldots, \mathbf{u}_D) \in \mathbb{R}^{3N \times D}$ is the truncated version of matrix $\mathbf{U}$, that contains only the eigenvectors associated with the largest D eigenvalues. An input atomic configuration, $\mathbf{q}$, is projected onto the low D-representation space as the following: $\mathbf{x} = \mathbf{U}_D^{\top}(\mathbf{q} - \overline{\mathbf{q}}) \in \mathbb{R}^{D}$. The number of principal components, D, is chosen by means of the elbow method[49]. We plot the logarithm of the fraction of variance explained (FVE), $log(\lambda_d / \mathrm{Tr}(\mathbf{\Lambda}))$, as a function of the principal component. We look for the point where the curve plateaus or changes slope. This point indicates the separation between the components accounting for the dominant part of the variance and those mostly containing noise (Fig. 2).

In order not to repeat the choice of D during the iterative search for CVs at finite temperature, we fix D to twice the number of principal components obtained by the above elbow method from the zero temperature data. Even allowing for this margin, the value of D in this work varies between 4 and 24, to be compared to a number of Cartesian coordinates ranging from 3×127 to $3 \times$

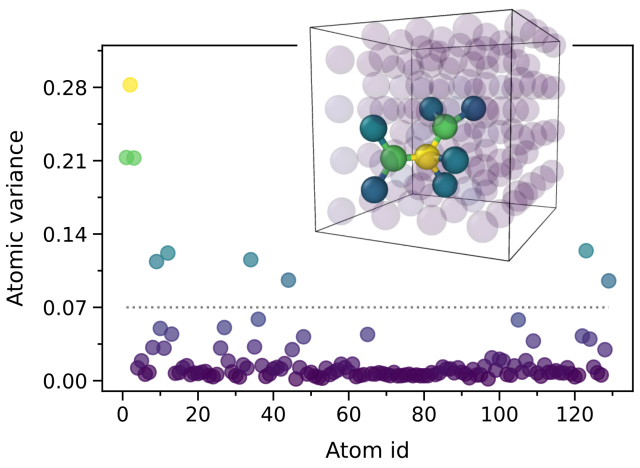

Fig. 3 Variance $var_{D,i}$ on atoms in a 129-atoms simulation cell with $\langle 110 \rangle$ self-interstitial dumbbell in bcc Fe. Each point on the plot corresponds to an atom. A large variance indicates high mobility of atoms during defect migration. In the inset structure the atoms are colored according to their variance. The atoms with $var_{D,i} < 0.07$ are set transparent.

2754.

When considering defects embedded in a large simulation cell, where the reaction coordinate is localized on few atoms around the defect center, this procedure acts as a simple linear filter eliminating the degrees of freedom with low contributions to the covariance matrix from the AE input. At the same time, it is in principle capable of capturing true collective motions of an arbitrarily large number of atoms. The latter case, however, is usually more difficult to deal with numerically, as we will discuss in Sec. 4.1.

In order to separate these two rather different cases, we represent the mobility of atom a during the migration of defect by projecting the diagonal covariance matrix $\mathbf{\Lambda}$ onto the atomic centers:

$$var_{D,a} = \sqrt{\sum_{d=1}^{D}\sum_{j=1}^{3} u_{d,aj}^2 \lambda_d}, \tag{3}$$

where $u_{d,aj}$ are the components of matrix $\mathbf{U}_D$ referring to the principal component d, atom a and Cartesian directions j. Representing atomic environments according to $var_{D,a}$ (Fig. 3) can be used to assess the level of localization of the CV.

Beyond allowing the treatment of large systems without overfitting, preprocessing the input data with PCA significantly reduces the computational cost, when compared to direct AE training. This is due both to the reduction of the number of input channels and to the extraction of linear correlations between the features before the AE training, which reduces the number of training steps. In order to perform PCA with minimum numerical cost, we use the iterative Singular Value Decomposition (SVD) algorithm[50,51], which computes $\mathbf{U}_D$ with a $\mathcal{O}(M \times N \times log(D))$ scaling instead of the $\mathcal{O}(N^3)$ scaling associated with a brute diagonalization of the sample covariance matrix.

2.3 Adaptive biasing force

The free energy in Landau sense, for a particular value of the CV $\boldsymbol{\zeta} \in \mathbb{R}^d$, can be written as follows,

$$A(\boldsymbol{\zeta}) = -\beta^{-1} \ln\left(Z^{-1} \int_{\Sigma(\boldsymbol{\zeta})} e^{-\beta U(\mathbf{q})} \mu_{\boldsymbol{\zeta}}(d\mathbf{q}) \right) \tag{4}$$

where the integration is made on the manifold $\Sigma(\boldsymbol{\zeta}) = \{\mathbf{q} \,|\, \xi(\mathbf{q}) = \boldsymbol{\zeta}\} \subset \mathbb{R}^{3N}$ with the conditional measure $\mu_{\boldsymbol{\zeta}}(d\mathbf{q})$ that is a delta measure for $\mathbf{q}$ for a given $\xi(\mathbf{q}) = \zeta$ such as $\mu_{\boldsymbol{\zeta}}(d\mathbf{q}) d\boldsymbol{\zeta} = d\mathbf{q}$, and Z is the canonical partition function[5].

The free energy profile can be integrated within different frameworks. In traditional ABF, a bias is added directly to the forces acting on the atoms, counteracting the average force along the CV direction[16]. This, unfortunately, requires the knowledge of the derivatives of the Jacobian determinant of the CV function. To avoid this, d fictitious degrees of freedom $\boldsymbol{\zeta}$ (one per dimension d of the reaction coordinate) in eABF[17,18] are coupled to the system through a harmonic potential:

$$U_k(\mathbf{q}, \boldsymbol{\zeta}) = U(\mathbf{q}) + \frac{k}{2} \| \xi(\mathbf{q}) - \boldsymbol{\zeta} \|^2, \tag{5}$$

where $\| \cdot \|$ is the euclidean distance and k a force coupling constant. In this formulation, the free energy associated to the extended potential A_k becomes:

$$A_k(\boldsymbol{\zeta}) = -\beta^{-1} \ln\left(\frac{\int_D e^{-\beta U_k(\mathbf{q},\boldsymbol{\zeta})} d\mathbf{q}}{\int_{D\times Z} e^{-\beta U_k(\mathbf{q},\boldsymbol{\zeta})} d\mathbf{q} d\boldsymbol{\zeta}} \right) \tag{6}$$

where D and Z are the domains of $\mathbf{q}$ and $\boldsymbol{\zeta}$, respectively, and the argument of the logarithm is equal to the probability $P_k(\boldsymbol{\zeta})$ to find the system at the coordinate $\boldsymbol{\zeta}$. The reaction coordinate provided by the present AE architecture $\xi(\mathbf{q}) = f_e \circ \mathrm{PCA}(\mathbf{q})$, makes the d components of the latent space almost independent and, moreover, the coupling with eABF method helps decouple the CV by the introduction of the extended potential. This formulation is close to the recent extended generalized ABF (egABF) formulation[21]. The crucial difference with classic and extended ABF is that in eABF the biasing force is applied to the fictitious particle $\boldsymbol{\zeta}$. The equation of motion for an overdamped Langevin dynamics gives:

$$\begin{aligned}
d\mathbf{q}_t &= -\nabla_{\mathbf{q}} U_k(\mathbf{q}_t, \boldsymbol{\zeta}_t) + \sqrt{2\beta^{-1}} dW_t^1 \\
d\boldsymbol{\zeta}_t &= -\nabla_{\boldsymbol{\zeta}} \left[U_k(\mathbf{q}_t, \boldsymbol{\zeta}_t) - A_k(\boldsymbol{\zeta}_t) \right] + \sqrt{2\beta^{-1}} dW_t^2 \\
\nabla_{\boldsymbol{\zeta}} A_k(\boldsymbol{\zeta}_t) &= \frac{\int_0^t \nabla_{\boldsymbol{\zeta}} U_k(\mathbf{q}_s, \boldsymbol{\zeta}_s) \mathbf{1}(\boldsymbol{\zeta}_t | \boldsymbol{\zeta}_s) ds}{\int_0^t \mathbf{1}(\boldsymbol{\zeta}_t | \boldsymbol{\zeta}_s) ds} = \mathbb{E}\left[\nabla_{\boldsymbol{\zeta}} U_k(\mathbf{q}_t, \boldsymbol{\zeta}_t) | \boldsymbol{\zeta}_t = \boldsymbol{\zeta} \right]
\end{aligned}$$

where W^1 and W^2 are $3N$-dimensional and d-dimensional Wiener process, respectively. $\mathbf{1}(\boldsymbol{\zeta}|\boldsymbol{\zeta}^\star)$ is some identity function being 1 when $\boldsymbol{\zeta}$ and $\boldsymbol{\zeta}^\star$ are in the same bin (i.e. in infinitesimal small neighbourhood of each other) and zero otherwise. For the case $d = 1$ the above equation reduces to the standard eABF case[17,18]. In the present scheme the bias, converges to an approximation, $A_k(\boldsymbol{\zeta})$, of the initial free energy, $A(\boldsymbol{\zeta})$. This biased free energy is defined as a convolution of the associated densities of states[52].

In numerical implementation the time is discrete $t \to t_n = n\delta t$ and the integration over the values of reaction coordinate is made on n_{bin} bins between $\min(\zeta) = 0$ and $\max(\zeta) = 1$ (we consider only the case $d = 1$). The eABF algorithm between the step n and $n+1$ becomes:

$$\begin{aligned}
\mathbf{q}_{n+1} &= \mathbf{q}_n - \nabla_{\mathbf{q}} U_k(\mathbf{q}_n, \zeta_n)\delta t + \sqrt{2\delta t/\beta}\mathbf{G}_n^1 \\
\zeta_{n+1} &= \zeta_n + \left[A'_n(\zeta_n) - \partial_\zeta U_k(\mathbf{q}_n, \zeta_n)\right]\delta t + \sqrt{2\delta t/\beta} G_n^2 \\
A'_{n+1}(\zeta_{n+1}) &= \frac{\sum_{i=1}^{n+1} \partial_\zeta U_k(\mathbf{q}_i, \zeta_i)\mathbf{1}(\zeta_{n+1}|\zeta_i)}{\sum_{i=1}^{n+1}\mathbf{1}(\zeta_{n+1}|\zeta_i)}
\end{aligned}$$

where $(\mathbf{G}_n^1, G_n^2)$ is Gaussian $3N+1$-dimensional vector. The step n requires the coordinates $(\mathbf{q}_n, \zeta_n)$, extended potential energy as well as the bias (A_n) and its derivative (A'_n) and predicts new coordinates $(\mathbf{q}_{n+1}, \zeta_{n+1})$ and new bias A_{n+1}/A'_{n+1}.

The above forms of ABF formalism are the most used[44]. In the present study we couple the CVs based on AE with other types of ABF, which are based on a Bayesian formulation[23–25]. Within this formalism the joint dynamics in ζ can be integrated into an implicit Bayesian formalism without computing an explicit dynamics for the extended variable. In the Bayesian procedure, we need the values of the system's free energy along the entire sampling history, because at each s^{th} step of sampling the algorithm requires a conditional probability $p_{A_s}(\zeta|\mathbf{q}_s)$ together with the free energy A_s, at the same time s. If n molecular dynamics steps are performed and we are able to provide the free energy $A_s(\zeta)$ and $p_{A_s}(\zeta|\mathbf{q}_s)$ at each s^{th} step, with $s = 0, \ldots, n-1$, then the move $s = n \to s = n+1$ can be performed using following steps.

Step 1. Firstly, $A'_n(\zeta)$ is computed as:

$$A'_n(\zeta) = \frac{\sum_{s=1}^{n-1} \nabla_\zeta U_k(\zeta, \mathbf{q}_s) p_{A_s}(\zeta|\mathbf{q}_s)}{\tau + \sum_{s=1}^{n-1} p_{A_s}(\zeta|\mathbf{q}_s)}, \tag{7}$$

where $\nabla_\zeta U(\zeta, \mathbf{q}_s)$ is easily computed from Eq. 5.

Step 2. The free energy $A_n(\zeta)$ is determined using standard integration, e.g. trapezoidal rule over the Z domain of ζ:

$$A_n(\zeta) = \int_0^\zeta A'_n(x)dx + A_n(0). \tag{8}$$

Step 3. The corresponding conditional probability for ζ for a given $\mathbf{q}_n$ is calculated as:

$$p_{A_n}(\zeta|\mathbf{q}_n) = \frac{\exp\{-\beta\left[U_k(\zeta, \mathbf{q}_n) - A_n(\zeta)\right]\}}{\int_Z \exp\{-\beta\left[U_k(\zeta, \mathbf{q}_n) - A_n(\zeta)\right]\}d\zeta} \tag{9}$$

Step 4. The "effective" force field is obtained by the equation:

$$\mathbb{F}_{A_n}(\mathbf{q}_n) = \int_Z -\nabla_{\mathbf{q}} U_k(\zeta, \mathbf{q}_n) p_{A_n}(\zeta|\mathbf{q}_n) d\zeta. \tag{10}$$

Step 5. Integration of the dynamics equation is carried out to obtain:

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \mathbb{F}_{A_n}(\mathbf{q}_n)\delta t + \sqrt{2\beta^{-1}\delta t}\mathbf{G}_n^1 \tag{11}$$

The main difficulty with this approach is in initializing the algorithm efficiently. The conditional probability in Eq. 9 depends on the free energy $A_n(\zeta)$. For a given ζ, the initial guessed values of $A_n(\zeta)$ and $p_n(\zeta)$ may be far from the ground truth, and in general they start converging towards the correct value once the full range of $\xi(q)$ has been explored. This sometimes leads to an amplification of the error on $A_n(\zeta)$ during the initial steps, which results in a slow exploration of the reaction coordinate, thereby a slow convergence of the algorithm. In order to eliminate the dependency on the initialization, we reweigh the update of the free energy derivative in Eq. 7 to reflect the growing confidence on the conditional probability values along the simulation:

$$A'_n(\zeta) = \frac{\sum_{s=1}^{n-1} \nabla_\zeta U_k(\zeta, \mathbf{q}_s) p_{A_s}(\zeta|\mathbf{q}_s) w_s}{\tau + \sum_{s=1}^{n-1} p_{A_s}(\zeta|\mathbf{q}_s) w_s}, \tag{12}$$

We use the simplest possible choice of a linearly increasing weight over the simulation, $w_s = s/N$, with N the total number of steps during the simulation.

2.4 Iterative CV discovery

The CV is defined as an application $\xi(\mathbf{q}) : \mathbb{R}^{3N} \to \mathbb{R}^d$. It is designed to encode the state of the system $\mathbf{q}$ into a manifold with low dimension d, and it is needed to couple the system dynamics with the bABF extended variable (Eq. 5). In turn, a finite-temperature CV can be extracted from the bABF simulation using an AE trained on snapshots of the dynamics.

For this reason, we employ an iterative approach alternating biased molecular dynamics simulations and AE training, similarly to previous studies of molecules[43–45]. However, our approach is designed for the specific case of materials science where the defect migration implies a much bigger number of degrees of freedom. Moreover, defect migration processes in cubic metals like Fe are highly degenerate, as the defect can move in different symmetry-equivalent directions. For this reason we fix the starting and final value of the CV to two local minima, $\mathbf{q}_0$ and $\mathbf{q}_1$, respectively, as opposed to freely exploring the free energy surface. This strategy also allows to start the iterative process from the simulations biased on the 0 K minimum energy path, which reduces the number of required MD steps and accelerates the convergence of the CV. Moreover, preconditioning the finite temperature pathway with the sequence minimum - saddle point - minimum generated at 0 K reduces the risk of instability in the iterative procedure, which are sometimes found when starting from unbiased simulations.[44] This workflow is different from the procedure adapted for molecules which requires only the initial state. Such unconstrained exploration of the free energy surface is in principle possible in a materials science context, but will require a modified workflow, including a treatment of degeneracies as well as a reweighting of the configurations in order to avoid instabilities of the iterative procedure.[44]

Furthermore, we introduce a new stopping rule for the iterative algorithm. Two CVs are considered equivalent when they describe the same isosurfaces in phase space, even if the isosurfaces do not correspond to the same value. This is equivalent to requiring $\xi^{(n)}(\mathbf{q}) = \phi(\xi^{(n)}(\mathbf{q}))$ for some scalar function ϕ. A test based on this idea, but additionally requiring that ϕ is a linear function, has recently been suggested.[44] When ϕ is a generic, unknown

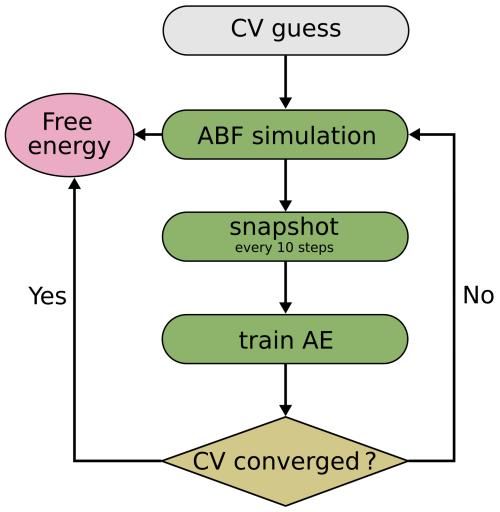

Fig. 4 Workflow of the iterative procedure for discovery of CV.

function, a test can be constructed based on $\nabla\xi$. At convergence,

$$\nabla_q \xi^{(n)}(\mathbf{q}) = \phi'(\xi^{(n-1)}(\mathbf{q}))\nabla_q \xi^{(n-1)}(\mathbf{q}) \tag{13}$$

and consequently both the following are true:

$$\nabla_q \xi^{(n)} \cdot \nabla_q \xi^{(n-1)} = \phi'(\xi^{(n-1)}) \, \|\nabla_q \xi^{(n-1)}\|^2 \tag{14}$$

$$\|\nabla_q \xi^{(n)}\| = |\phi'(\xi^{(n-1)})| \, \|\nabla_q \xi^{(n-1)}\| \tag{15}$$

where the $\mathbf{q}$ arguments have been dropped to lighten the notation. Given that we fix the value of the CV at the starting and final local minima (as discussed below) ϕ' is always positive and the modulus can be dropped, leading to the following equivalence condition between $\xi^{(n)}$ and $\xi^{(n-1)}$:

$$\frac{\nabla_q \xi^{(n)} \cdot \nabla_q \xi^{(n-1)}}{\|\nabla_q \xi^{(n)}\| \|\nabla_q \xi^{(n-1)}\|} \approx 1 \tag{16}$$

The iterative procedure can be summarized as follows:

0. An initial guess $\xi^{(0)}$ for the CV is provided, constructed in such a way that the value at the local minima is $\xi^{(0)}(\mathbf{q}_0) = \mathbf{0}$ and $\xi^{(0)}(\mathbf{q}_1) = \mathbf{1}$. This can be any reasonable guess, including the projection on the line connecting the initial and final state $\xi^{(0)}(\mathbf{q}) = (\mathbf{q}-\mathbf{q}_0)\cdot(\mathbf{q}_1-\mathbf{q}_0)/\|\|\mathbf{q}_1-\mathbf{q}_0\|\|^2$, which we use in the following. As an alternative to speed up the convergence, the 0 K minimum energy path can be used. Such a path can be found by various 0 K methods, of which we use the Nudged Elastic Band[53] approach. The AE can be trained to reproduce the minimum energy path as described in Sec. 3.2.

At each iteration n, the following steps are executed:

1. A biased MD simulation using Bayesian Adaptive Biasing Force (bABF) is performed using $\xi^{(n-1)}(\mathbf{q})$ as CV. The free energy path between $\xi^{(n-1)}(\mathbf{q}_0)$ and $\xi^{(n-1)}(\mathbf{q}_1)$ along the reaction coordinate is computed, and the atomic coordinates from full trajectory are recorded. The retained structures from MD trajectory are saved every 10 MD steps in order to avoid strong correlations between the snapshots

2. A new AE is trained on the atomic structures from the previous trajectory in order to obtain an improved CV, $\xi^{(n-1)}(\mathbf{q}) \to \xi^{(n)}(\mathbf{q})$. The output of the AE is rescaled in order to retain the same value of CV at the start and endpoint of the simulation, i.e. $\xi^{(n)}(\mathbf{q}_0) = \mathbf{0}$ and $\xi^{(n)}(\mathbf{q}_1) = \mathbf{1}$.

3. The new CV $\xi^{(n)}$ is compared with $\xi^{(n-1)}$ to assess convergence. In order to assess the condition of Eq. 16, we exploit the capability of decoders to be used as generative models to compute the mean CV path $\tilde{\mathbf{q}}_i = f_d^{(n-1)}(\zeta_i)$ for a uniform grid of $\zeta_i = i/N$ for $i = 0, \ldots N$. To satisfy the convergence criterion, the derivative of the CV path must have the same direction for the two AEs:

$$\sum_{i=0}^{N} 1 - \frac{\nabla_q \xi^{(n)}(\tilde{q}_i) \cdot \nabla_q \xi^{(n-1)}(\tilde{q}_i)}{\|\nabla_q \xi^{(n)}(\tilde{q}_i)\| \|\nabla_q \xi^{(n-1)}(\tilde{q}_i)\|} < N\varepsilon_{\mathrm{TOL}} \tag{17}$$

for some value of $\varepsilon_{\mathrm{TOL}}$.

The free energy and CV obtained in the last iteration are the results of the iterative process. The convergence of the free energy barrier during the iterative procedure can be easily verified visually or quantitatively, providing an independent check that the iterative procedure provides converged free energy profiles.

3 Computational details

3.1 Autoencoder design and training

In this study, the encoder function $f_e : \mathscr{R}^D \to \mathscr{R}$ is a dense NN with two ($L = 2$) hidden layers of dimension d_l, plus a bottleneck layer of dimension $d = 1$. In the following, the dimension d_l is chosen equal to D, except the cases when different dimension is stated. This results in a number of parameters equal to $d_l \times (D + d_l + 2) + 1$. The choice of a one-dimensional bottleneck is mainly due to the underlying physical processes studied, which can be modeled through a single reaction coordinate. Even when this is known, in the general case a larger bottleneck may be needed to account for unrelated large scale motions which may occur in the system. Preconditioning the iterative procedure with the 0 K minimal energy path, however, should help to produce a CV which correctly separates the metastable states. We use a symmetric decoder function f_d, i.e. with the same weights as the encoder, but different biases which account for $D + 2d_l$ additional parameters. We find that this constraint on weights is sufficient to regularize the AE($\mathbf{x}$) function, without including an explicit regularization term or dropout layers.

For all hidden layers, we use Rectified Linear Unit (ReLu) activation functions, and linear activations for output and bottleneck layers. Despite the ReLu activation function having a discontinuity in the derivative, the gradients of the embedding are sufficiently smooth for our purposes. In general we found ReLu provides better results and easier training of the NN with respect to smoother activation functions such as sigmoid, as is commonly found for deep learning applications. We do not use batch normalization layers, as they give different results at inference with respect to training[54,55]

We initialize the AE weights using Glorot Normal initializer. The training is always started from scratch, as one of the advantages of PCA preprocessing is a relatively fast training of the small resulting NN. The position of the center of mass is subtracted from the structures in order to eliminate any spurious drift, and the dataset is divided in batches of 1000 structures for training. We use the Adam optimizer, and the learning rate is gradually reduced on a schedule starting from 20 epochs at rate 0.02 and progressively reducing the learning rate to 0.000005 for which 200 epochs are performed. We do not use early stopping.

3.2 Encoding synthetic data

In Sec. 4.1 we test the reconstruction of 0 K reaction coordinates (i.e. minimal energy paths) from noisy data. This has two goals. The first is testing the quality of the CV and fine-tune the AE parameters and data pre-processing without coupling the AE with molecular dynamics. This makes the testing faster while allowing to identify clearly any difficulties arising from the machine learning part of the workflow, separately from molecular dynamics and free energy integration. The second is that the 0 K CV obtained in this way is a useful starting point for the iterative procedure outlined above. The minimal energy path (MEP) is relatively easy to identify in the vast majority of materials science problems, using 0 K methods such as the Replica Chain method [56], Nudged Elastic Band[53], Discrete Path Sampling [57], String method [58]. As we want this to be a fair test of the AE capabilities, we train the AE by adding a Gaussian noise to the MEP of comparable magnitude to the thermal fluctuations during a molecular dynamics simulation, as described below. Once the training of AE is achieved, we encode each NEB image structure in order to obtain the CV $\xi(\mathbf{x}_i)$ and its derivative $\nabla_{\mathbf{x}}\xi(\mathbf{x}_i)$ We denote by $\mathbf{x}_i$ the transformed atomic coordinates using PCA operator of the i^{th} NEB replica of the system along the reaction pathway $\mathbf{q}_i = \mathbf{X}(\lambda_i)$. We then compute the energy profile by integrating along the resulting CV and compare the result with the reference energy of the configurations along the path.

Training data. The minimum energy pathway at 0 K is obtained with the climbing image NEB method[59–61] from LAMMPS[62], which provides K knots replica of the system between initial $\mathbf{q}_0$ and final $\mathbf{q}_1$ state configurations. In the following we use K=28, except for the dislocation glide where K=128 knots are used. The forces are converged to 10^{-3} eV/Å.

The knot images are interpolated using akima cubic splines of atomic coordinates, $\mathbf{X}(\lambda) : [0,1] \in \mathbb{R} \rightarrow \mathbb{R}^{3N}$ with $\mathbf{X}(0) = \mathbf{q}_0$ and $\mathbf{X}(1) = \mathbf{q}_1$. This procedure has two purposes: (i) accurate evaluation of the migration barriers from the mean force[19,63] at 0 K, and (ii) use of the interpolated structures for data augmentation. To this end, we generate M_{train}+M_{test} structures, $\mathbf{q}_\lambda = \mathbf{X}(\lambda)$, by a grid of equally spaced λ in the interval $[0,1]$. For each configuration we add a Gaussian noise with a variance of 0.005 Å which is close to the average atomic displacements in bcc Fe at 300 K. Translations and the component of the noise along the spline are removed so that the data are evenly spaced along the NEB path.

Energy integration. Integration of the total energy at 0 K is performed on the minimum energy path. A free energy profile along a reaction coordinate can be obtained by integration of the average force[16,19], with an expression that simplifies for the energy at T=0 K to:

$$\partial_\zeta E(\zeta, T=0) = \left\langle \frac{\mathbf{w}(\mathbf{q})\cdot\nabla V(\mathbf{q})}{\mathbf{w}(\mathbf{q})\cdot\xi(\mathbf{q})} \right\rangle_{\xi(\mathbf{q})=\zeta} \tag{18}$$

where $\mathbf{w}(\mathbf{q})$ is any vector field for which $\mathbf{w}(\mathbf{q})\cdot\xi(\mathbf{q}) \neq 0$, of which $\xi(\mathbf{q})$ itself is the obvious choice if known as in our case. In addition, at 0 K the conditional thermodynamical average can be dropped in favor of a direct integration along the MEP, which is known at least at the discrete points provided by the NEB path, allowing a numerical integration.

In the general case, where the AE bottleneck is not one-dimensional, the components ξ_a of the embedding are not orthogonal, and the derivatives $\nabla_x\xi_a(\mathbf{x}_i)$ are orthogonalized with their Gram matrix:

$$\begin{aligned} E_i &= E_{i-1} + \frac{1}{2}(\boldsymbol{\xi}(\mathbf{x}_i) - \boldsymbol{\xi}(\mathbf{x}_i))\cdot(\mathbf{F}_i + \mathbf{F}_{i-1}) \\ \mathbf{F}_i &= \mathbf{U}_D\mathbf{S}_i^{-1}\nabla_x\boldsymbol{\xi}(\mathbf{x}_{i-1})\cdot\nabla V(\mathbf{q}) \\ \mathbf{S}_i &= \nabla_{\mathbf{x}}\boldsymbol{\xi}(\mathbf{x}_i)^\top\nabla_{\mathbf{x}}\boldsymbol{\xi}(\mathbf{x}_i) \end{aligned} \tag{19}$$

where we have explicitly expanded $\boldsymbol{\xi}(\mathbf{q}) = \mathbf{U}_D\boldsymbol{\xi}(\mathbf{x})$. In the one-dimensional case this reduces to:

$$\begin{aligned} E_i &= E_{i-1} + \frac{1}{2}[\xi(\mathbf{x}_i) - \xi(\mathbf{x}_{i-1})](F_i + F_{i-1}) \\ F_i &= \frac{\mathbf{U}_D\nabla_x\xi(x_i)\cdot\nabla V(\mathbf{q})}{\|\nabla_x\xi(x_i)\|^2} \end{aligned} \tag{20}$$

and it can be easily seen that Eq. 20 is Eq. 18 minus the thermodynamical average.

3.3 Bayesian adaptive biased force

The latent coordinate $\tilde{\xi}(\mathbf{q})$ obtained from the AE is renormalized via a linear transformation $\xi(\mathbf{q}) = (\tilde{\xi}(\mathbf{q}) - \tilde{\xi}(\mathbf{q}_0))/(\tilde{\xi}(\mathbf{q}_1) - \tilde{\xi}(\mathbf{q}_0))$ so that $\xi(\mathbf{q}_0) = 0$ and $\xi(\mathbf{q}_1) = 1$ for the start and endpoint of the simulation. The extended degree of freedom ζ for bABF is discretized in $n_\zeta = 100$ bins between 0 and 1. This discretization extends to the ζ-dependent quantities $A_n(\zeta)$, $A'_n(\zeta)$ and $p_{A_n}(\zeta|\mathbf{q})$. Following the established procedure for eABF[22], we use a temperature-dependent value of the spring constant $k = k_BTn_\zeta^2$ which results in $\langle(\zeta - \xi(\mathbf{q}))^2\rangle \sim 1/n_\zeta$. The regularization constant takes the value of $\tau = 3T^2/(N_{steps}n_\zeta)$, where the total number of steps N_{steps} is used for consistency with the choice of probability reweighting in Eq. 12. For the integration of the Langevin dynamics we use a time step of $\delta t = 1$ fs and a damping constant $\gamma = \delta t \cdot T/m_{Fe}$, which result in an average discretization step of approximately 0.01 Å. For the convergence of the iterative procedure of Sec. 2.4, we use $\varepsilon_{\mathrm{TOL}} = 0.1$, which provided well-converged free energy barriers. We stop the procedure after five iterations if the convergence threshold is not reached yet.

4 Results

4.1 Reconstruction of 0K energy barriers by encoding synthetic data

Energy barriers can be accurately reconstructed from the integration of mean force of the system, both for 0 K and finite temperature[19,63]. This requires a knowledge of the minimum energy (or free energy) pathway and its derivatives. Here we first test the quality of the CVs provided by the AE for the energy profiles at 0 K. This test is particularly informative as it does not require to couple the AE with molecular dynamics simulations, which brings additional methodological degrees of freedom and acts as possible sources of added error. To this end, we train the AE on noisy synthetic data as described in Sec. 3.2 and test its accuracy in reproducing the migration energy barriers of different defect types in bcc iron.

Our first test case is migration of $\langle 110 \rangle$ self-interstitial dumbbell. The simulation cell contains 129 atoms described by 387 Cartesian coordinates. The distortion field produced by this small defect is localized around it and impacts less than 20 atoms[64]. As a consequence, the variance of the atomic positions along the migration path is concentrated on few atoms (Fig. 5c). Therefore, this is a relatively easy case for the AE. Figure 5a compares the energy profile obtained by integrating along the CV with the NEB energy, showing a perfect agreement between the energy barriers. The migration of the dumbbell occurs by translation rotation to the first nearest neighbour position[65], which can be recognized in Fig. 5c where the atoms are colored according to the variance $var_{D^{(i)}}$.

We further consider the formation of small cluster built from two $\langle 110 \rangle$ interstitial dumbbells (Fig. 5b) in a 130-atom simulation cell. The atomic displacements in this case are concentrated around the interstitial that migrates to form the cluster, with only a minor relaxation of the other dumbbell. Consequently, the variances of the atomic positions are actually very similar to the previous case, as can be appreciated in Fig. 5c. Also in this case the CV is reconstructed accurately and we obtain an excellent agreement with the NEB barrier (Fig. 5b). Thanks to the small number of degrees of freedom, and thus of neural network parameters required, and to the relatively simple CV, these energy barriers can be obtained without reducing the dimensionality of the AE input by PCA. In addition, the result is little dependent on the choice of the hyperparameters.

A single vacancy also represents a localized point defect (Fig. 6a). The number of AE parameters is nearly equal to the one of the interstitials (Fig. 5). However, we find that the task of extracting the CV from a 127-atom simulation cell with a mono-vacancy, is more sensitive to the size of the training set and to the choice of the hyperparameters, likely due to overfitting. Preprocessing the input data with PCA, as described in Sec. 2.2, allows to overcome this problem and yields very good results (Fig. 6a).

We further increase the complexity and consider a stringent case with the transformation of a 3D C15 interstitial cluster I_2^{C15} [66] into the so-called non-parallel cluster I_2^{NP} [67]. The I_2^{C15} and I_2^{NP} are intrinsically immobile and represent important instances that impact the evolution of the microstructure in bcc Fe under irradiation[30,68]. The transformation pathway at 0 K, obtained using the activation relaxation technique (ART)[69] involves simultaneous displacement of multiple atoms (from 57 to 32 along the transition path[64]). In addition, moderately large unit cell (1 026 atoms) is needed to reduce the interactions between the periodic images of the defect clusters. An increase of the dimensionality of the input layer tends to promote overfitting of the AE. By tuning the NN parameters through cross-validation a reasonable energy barrier can be obtained (Fig. 6b). However, by performing a PCA dimensionality reduction of the input data, we easily obtain better results. Indeed, the collective variable is contained in the subspace spanned by the first ten principal components, which allows to perform an accurate integration with a small neural network (Fig. 6b).

A very complex and challenging case is the energy profile produced by dislocations. Here we consider the energy barrier of a $\frac{1}{2}\langle 111 \rangle$ screw dislocation gliding in a $\{110\}$ plane in bcc Fe. In this case, complexity arises both due to the large number of degrees of freedom (2 754 atoms in the simulation cell) and to the long range displacement field produced by the dislocation, affecting most of the atom positions in the simulation cell. In addition, the total energy barrier for the dislocation glide is relatively small, which in combination with the large unit cell results in a barrier below 0.03 meV per atom, requiring high accuracy of the integration. If the AE is trained with atomic coordinates as input, it fails to recover the reaction coordinate and reconstruct the energy profile due to strong overfitting. PCA can reduce the number of AE input channels to 20, after which the AE is easily trained to obtain a one-dimensional CV. However, with this approach, we are able to reconstruct the energy barrier underestimated at best by 5% (see Fig. 6c). This error already appears when integrating the energy profile within the PCA subspace. It is slowly reduced by increasing the number of training configurations, pointing to a difficulty to identify the pertinent CV by dimensionality reduction in the presence of random thermal noise. We note that the result shown in Fig. 6c is obtained using 8 times more training configurations with respect to the cases of vacancy migration and C15 cluster transition. The underestimation of the energy barrier is similar to the one obtained from the integration based on the distortion scores of local atomic environments[64], when including only the forces on the atoms that deviate sufficiently from the defect-free bulk. This suggests that data-driven dimensionality reduction, including PCA and AE networks, is not fully able to distinguish between the long-range elastic displacement field produced by dislocations and the thermal noise in the bulk. Increasing the size of the training set allows to better statistically separate the two, but the full long range tail of the elastic displacement would be recovered only in the limit of a very large training set.

4.2 Iterative CV refinement for reconstruction of free energy profiles at finite temperature

We apply the iterative procedure described in Sec. 2.4 to the migration of a vacancy in α-Fe at different temperatures. We note that in this case, using the 0 K CV is sufficient for integrating the

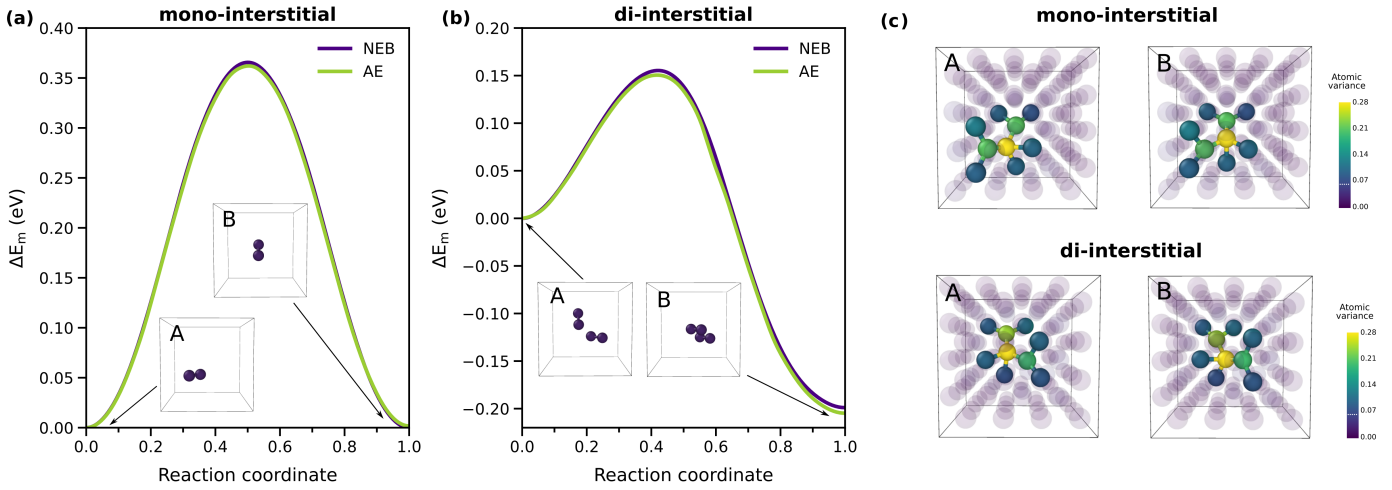

Fig. 5 Migration of $\langle 110 \rangle$ self-interstitial dumbbells in bcc Fe. (a) 0 K Migration energy of mono-interstitial dumbbell obtained using AE with 1 hidden layer with 20 ReLu units. (b) Energy profile of di-interstitial cluster formation from 2 dumbbells obtained using AE with 1 hidden layer with 80 ReLu units. The energy barriers (a,b) are obtained based on 16 000 training configurations in each case. The inset structures in (a,b) illustrate the defect structures in the initial and final states A and B, respectively. (c) The initial and final state structures A and B of mono- and di-interstitial defects with atoms colored according to the variance $var_{D^{(i)}}$. A large variance indicates high mobility of atoms during the migration process. The atoms with $var_{D^{(i)}} < 0.07$ are set transparent.

finite temperature free energy[26], as the path does not radically change with increasing temperature. However, the convergence of such approach is slow, requiring up to 10^7 force calculations[26]. For our simulations, we use between 10^5 and 10^6 MD steps (depending on the temperature) for each free energy integration, which are not sufficient to obtain accurate profiles from a non-converged CV. Indeed, when the bABF algorithm is applied at the first iteration using the CV guess, the profiles are noisy and the energy barriers are underestimated, as can be seen in Fig. 7.

At 500 K, the kinetic energy is sufficient to overcome the migration even without applying the biasing procedure (the temperature of stage III of the vacancy in the resitivity recovery experiments is between 250 K and 300 K[70]). Therefore, the vacancy could occasionally migrate in a different direction, after which the simulation would become trapped in this state degrading both the free energy profile and the training data for the AE. To avoid this, we stop the simulation when it deviates excessively from the CV path, by projecting the position on the hyperplane orthogonal to the CV derivative:

$$\left\| (\mathbf{q} - \hat{\mathbf{q}}) - \frac{\nabla\xi(\mathbf{q}) \cdot (\mathbf{q} - \hat{\mathbf{q}})}{\|\nabla\xi(\mathbf{q})\|} \right\| > \alpha \qquad (21)$$

with $\alpha = 2.5$ Å being adequate for this system and temperature. We then restart the biased Langevin dynamics from the the initial position, while preserving the average force and conditional probabilities computed up to that point.

We then train the AE on the snapshots of the MD trajectory to obtain a new CV, iterating the procedure as described in Sec. 2.4. The results are shown in Fig. 7. For each temperature, the approach returns a CV of good quality, allowing convergence in a few iterations toward accurate free energy barriers. For $T = 100$ K and $T = 300$ K, we obtain a smooth free energy profile with a limited number of force calculations. At $T = 500$ K, the thermal noise is high with respect to the migration barriers. Thereby, a larger number of steps is required to obtain smooth curves.

5 Conclusions and perspectives

We have presented the performance and limitations of deep learning AEs coupled with accelerated MD free energy sampling when applied to the evaluation of reaction coordinates and free energy profiles of defect migration in solids, taking the example of common defects in bcc Fe. Applying AEs directly on all the atomic positions implies very large number of parameters, which not only increases the computational cost and requires enlarging the size of the training dataset, but also can result in CVs that do not reflect the dynamics of the system. Therefore, the effective application of AEs to atomic-scale processes in solids requires preprocessing the input. To this end, we perform linear dimensionality reduction by PCA, which filters out the degrees of freedom associated with low covariance of the data while preserving the objective funtion minimized by the AE. The proposed strategy provides accurate reaction coordinates in the case of localized defects. However, for the defects with long-range displacement field, like dislocations, our approach exhibits a limited performance, which results in a systematic underestimation of the energy barriers. Based on these results, we recommend inspecting localization of the atomic-scale processes in the system before the application of AEs for extracting collective variables. Such a test can be performed, for instance, by taking the first principal components of the atomic displacements along a minimum energy path and projecting them back on the atomic centers, as we propose in this work.

The deep learning AEs has been successfully coupled with bayesian ABF, resulting in an iterative procedure which enables automatic discovery of reaction coordinates and reconstruction of Landau free energy profile. Considering the example of va-

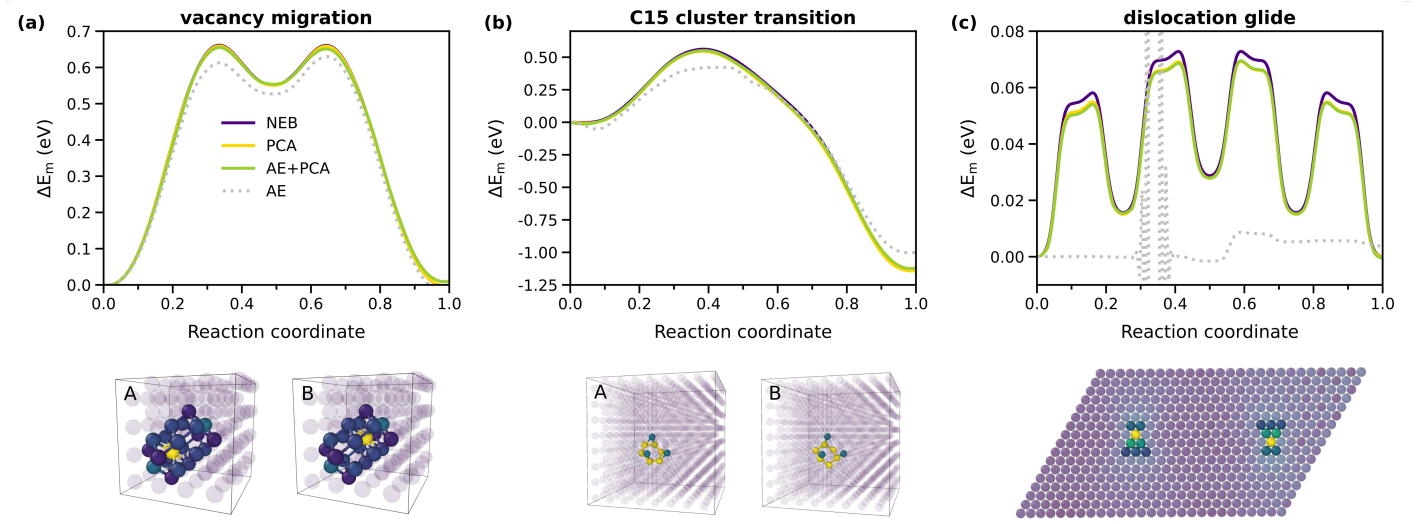

Fig. 6 (a) Energy profile of vacancy migration in bcc Fe. NEBs label the full atomistic energy profile, whilst AE and AE+PCA are the present reaction coordinate method. The PCA method of barrier energy reconstruction is used just for comparison and is not a reaction coordinate method i.e. the energy profile is obtained after reducing the dimensionality of the data to 12 PCA components. For AE+PCA the same 12 components are used to train a NN AE (2 hidden layers with 12 ReLu units, trained on 20 000 configurations) to obtain the true collective variable. (b) Energy profile of C15 cluster transition into triangular Gao cluster in bcc Fe. PCA profile is obtained by reducing the dimensionality of the data to 10 PCA components, followed by NN AE (2 hidden layers with 10 ReLu units) trained on 50 000 configurations for the AE+PCA profile. (c) Energy profile of two $\frac{1}{2}\langle 111\rangle$ screw dislocations gliding successively in the $\{110\}$ plane in bcc Fe. The AE+PCA are obtained after reducing the dimensionality of the data to 20 PCA components, followed by NN AE with 2 hidden layers with 120 ReLu units, trained on 150 000 configurations with a variance of the Gaussian noise set to 0.0025 Å. The relative difference beween the PCA+AE energy profile and the NEB target is 5±3%. The structures on the lower panel correspond to the initial and final state of the system with the atoms colored according to the variance $var_{D^{(i)}}$. Larger variance indicates high mobility of atoms during the transition process. The color scale is similar to that in Figure 5.

cancy migration at various temperatures, we demonstrate that such a coupling has very promising applications for the future studies of localized defects. The capabilities of this approach to provide appropriate CVs for phenomena such as phase transitions, crystallization or amorphization, remain to be explored but could require further methodological adaptations. Specifically, a different preprocessing of the input coordinates into functions of the interatomic distances[71] or local atomic descriptors[72,73] could be leveraged to deal with the delocalised and permutation-symmetric CVs involved in these transitions. At the same time, the excellent learning obtained with AEs for localised defects suggests that similar architectures for deep generative models, such as VAEs or invertible NNs[74], could be applied to directly sample the phase space obviating the need to simulate explicitly the system dynamics for these problems.

The present workflow `COVAEM` (COllective Variables from AutoEncoders in Materials) with examples necessary for the reproduction of the results are available at GitHub repository https://github.com/ai-atoms/covaem.

Acknowledgements

J.B. acknowledges support from the Cross-Disciplinary Program on Numerical Simulation of CEA, the French Alternative Energies and Atomic Energy Commission. For A.M.G and M.C.M., this work has been carried out within the framework of the EUROfusion Consortium, funded by the European Union via the Euratom Research and Training Programme (Grant Agreement No 101052200 — EUROfusion). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. T.D.S. gratefully recognises support from the Agence Nationale de Recherche, via the MEMOPAS project ANR-19-CE46-0006-1. J.B., A.M.G. and M.C.M. acknowledge support from GENCI - (CINES/CCRT) computer centre under Grant No. A0110906973.

Notes and references

1 C.-C. Fu, J. Dalla Torre, F. Willaime, J.-L. Bocquet and A. Barbu, *Nat. Mater.*, 2005, **4**, 68–74.
2 D. Caillard and J.-L. Martin, *Thermally activated mechanisms in crystal plasticity*, Elsevier, 2003.
3 V. Holten, C. Bertrand, M. Anisimov and J. Sengers, *J. Chem. Phys.*, 2012, **136**, 094507.
4 P. Hänggi, P. Talkner and M. Borkovec, *Rev. Mod. Phys.*, 1990, **62**, 251–341.
5 T. Lelièvre, G. Stoltz and M. Rousset, *Free energy computations: A mathematical perspective*, Imperial College Press, London, 2010.
6 C. Chipot and A. Pohorille, *Free energy calculations*, Springer-Verlag Berlin Heidelberg, 2007.
7 F. Pietrucci, *Rev. Phys.*, 2017, 32–45.
8 A. Laio and F. L. Gervasio, *Rep. Prog. Phys.*, 2008, **71**, 126601.
9 D. Frenkel and B. Smit, *Understanding molecular simulations: from algorithms to applications*, Academic press, San Diego,

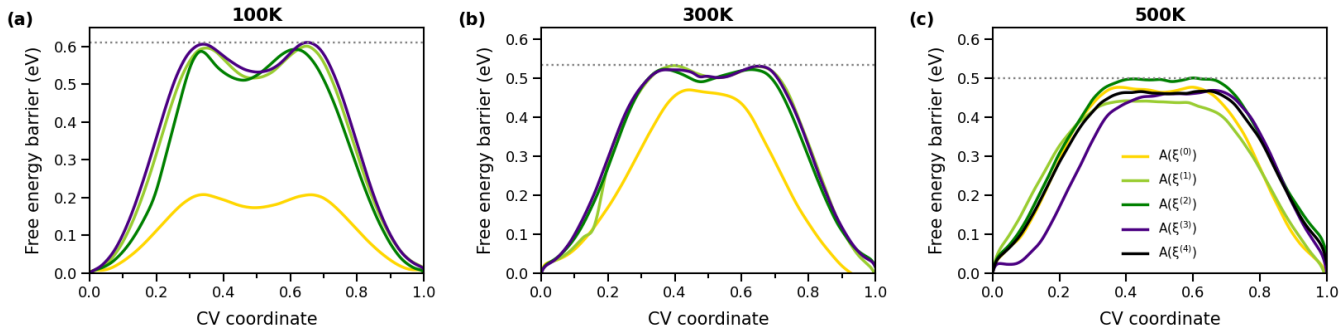

Fig. 7 Free energy profile for a vacancy migration in bcc Fe obtained with iterative procedure at (a) 100 K; (b) 300 K; and (c) 500 K. Each free energy integration uses $1 \cdot 10^5$, $5 \cdot 10^5$ and $1 \cdot 10^6$ time steps respectively for the three temperatures. Grey horizontal lines indicate the energy barriers previously computed with a traditional approach in Ref. [26].

2002.

10 G. M. Torrie and J. P. Valleau, *J. Comp. Phys.*, 1977, **23**, 187–199.

11 J. Kästner, *WIRES Comput. Mol. Sci.*, 2011, **1**, 932–942.

12 A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. USA*, 2002, **99**, 12562–12566.

13 A. Laio and F. L. Gervasio, *Rep. Prog. Phys.*, 2008, **71**, 126601.

14 B. M. Dickson, F. Legoll, T. Lelièvre, G. Stoltz and P. Fleurat-Lessard, *J. Phys. Chem. B*, 2010, **114**, 5823–5830.

15 B. M. Dickson, *Curr. Opin. Struct. Biol.*, 2017, **43**, 63–67.

16 E. Darve, D. Rodríguez-Gómez and A. Pohorille, *J. Chem. Phys.*, 2008, **128**, 144120.

17 T. Lelièvre, M. Rousset and G. Stoltz, *J. Chem. Phys.*, 2007, **126**, 134111.

18 L. Zheng, M. Chen and W. Yang, *Proc. Natl. Acad. Sci. USA*, 2008, **105**, 20227–20232.

19 T. D. Swinburne and M.-C. Marinica, *Phys. Rev. Lett.*, 2018, **120**, 135503.

20 A. Lesage, T. Lelièvre, G. Stoltz and J. Hénin, *J. Phys. Chem. B*, 2017, **121**, 3676–3685.

21 T. Zhao, H. Fu, T. Lelièvre, X. Shao, C. Chipot and W. Cai, *J. Chem. Theory Comput.*, 2017, **13**, 1566–1576.

22 H. Fu, X. Shao, C. Chipot and W. Cai, *J. Chem. Theory Comput.*, 2016, **12**, 3506–3513.

23 L. Cao, G. Stoltz, T. Lelièvre, M.-C. Marinica and M. Athènes, *J. Chem. Phys.*, 2014, **140**, 104108.

24 P. Terrier, M.-C. Marinica and M. Athènes, *J. Chem. Phys.*, 2015, **143**, 134121.

25 M. Athènes and P. Terrier, *The Journal of Chemical Physics*, 2017, **146**, 194101.

26 M. Athènes and M.-C. Marinica, *J. Comp. Phys.*, 2010, **229**, 7129–7146.

27 V. Mironov, Y. Alexeev, V. K. Mulligan and D. G. Fedorov, *J. Comp. Chem.*, 2019, **40**, 297–309.

28 J. P. Hirth and J. Lothe, *Theory of dislocations*, Wiley, New York, 1982.

29 B. Uberuaga, R. Hoagland, A. Voter and S. Valone, *Phys. Rev. Lett.*, 2007, **99**, 135501.

30 M.-C. Marinica, F. Willaime and N. Mousseau, *Phys. Rev. B*, 2011, **83**, 094119.

31 L. Bonati, G. Piccini and M. Parrinello, *Proc. Natl. Acad. Sci. USA*, 2021, **118**, e2113533118.

32 P. Gkeka, G. Stoltz, A. Barati Farimani, Z. Belkacemi, M. Ceriotti, J. D. Chodera, A. R. Dinner, A. L. Ferguson, J.-B. Maillet, H. Minoux *et al.*, *J. Chem. Theory Comput.*, 2020, **16**, 4757–4775.

33 M. Chen, *Eur. Phys. J. B*, 2021, **94**, 1–17.

34 G. E. Hinton and R. R. Salakhutdinov, *Science*, 2006, **313**, 504–507.

35 M. Scholz, M. Fraunholz and J. Selbig, Principal manifolds for data visualization and dimension reduction, 2008, pp. 44–67.

36 D. P. Kingma and M. Welling, *Auto-Encoding Variational Bayes*, 2014, `https://arxiv.org/abs/1312.6114`.

37 P. Ghosh, M. S. M. Sajjadi, A. Vergari, M. Black and B. Schölkopf, *From Variational to Deterministic Autoencoders*, 2019, `https://arxiv.org/abs/1903.12436`.

38 W. Wang, D. Yang, F. Chen, Y. Pang, S. Huang and Y. Ge, *IEEE Access*, 2019, **7**, 62421–62432.

39 D. P. Kingma and M. Welling, *Found. Trends Mach. Learn.*, 2019, **12**, 307–392.

40 J. Masci, U. Meier, D. Cireşan and J. Schmidhuber, Artificial Neural Networks and Machine Learning – ICANN 2011, 2011, pp. 52–59.

41 J. Xie, L. Xu and E. Chen, Advances in neural information processing systems, 2012, pp. 341–349.

42 C. Zhou and R. C. Paffenroth, Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017, pp. 665–674.

43 W. Chen and A. L. Ferguson, *J. Comput. Chem.*, 2018, **39**, 2079–2102.

44 Z. Belkacemi, P. Gkeka, T. Lelièvre and G. Stoltz, *Journal of Chemical Theory and Computation*, 2022, **18**, 59–78.

45 J. M. L. Ribeiro, P. Bravo, Y. Wang and P. Tiwary, *J. Chem. Phys.*, 2018, **149**, 072301.

46 M. M. Sultan, H. K. Wayment-Steele and V. S. Pande, *J. Chem. Theory Comput.*, 2018, **14**, 1887–1894.

47 C. X. Hernández, H. K. Wayment-Steele, M. M. Sultan, B. E. Husic and V. S. Pande, *Phys. Rev. E*, 2018, **97**, 062412.

48 C. Wehmeyer and F. Noé, *J. Chem. Phys.*, 2018, **148**, 241703.
49 R. L. Thorndike, *Psychometrika*, 1953, **18**, 267–276.
50 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, *JMLR*, 2011, **12**, 2825–2830.
51 N. Halko, P. G. Martinsson and J. A. Tropp, *SIAM Review*, 2011, **53**, 217–288.
52 J. Comer, J. C. Gumbart, J. Hénin, T. Lelièvre, A. Pohorille and C. Chipot, *J. Phys. Chem. B*, 2015, **119**, 1129–1151.
53 G. Henkelman, *Annual Review of Materials Research*, 2017, **47**, 199–216.
54 S. Ioffe and C. Szegedy, International conference on machine learning, 2015, pp. 448–456.
55 S. Ioffe, Advances in Neural Information Processing Systems, 2017.
56 R. Elber and M. Karplus, *Chem. Phys. Lett.*, 1987, **139**, 375 – 380.
57 D. J. Wales, *Mol. Phys.*, 2002, **100**, 3285–3305.
58 E. Weinan, W. Ren and E. Vanden-Eijnden, *Phys. Rev. B*, 2002, **66**, 052301.
59 H. Jónsson, G. Mills and K. W. Jacobsen, in *Classical and Quantum Dynamics in Condensed Phase Simulations*, WORLD SCIENTIFIC, 1998, ch. Nudged elastic band method for finding minimum energy paths of transitions, pp. 385–404.
60 G. Henkelman and H. Jónsson, *J. Chem. Phys.*, 2000, **113**, 9978–9985.
61 G. Henkelman, B. P. Uberuaga and H. Jónsson, *J. Chem. Phys.*, 2000, **113**, 9901–9904.
62 S. Plimpton, *J. Comput. Phys.*, 1995, **117**, 1–19.
63 T. D. Swinburne and J. R. Kermode, *Phys. Rev. B*, 2017, **96**, 144102.
64 A. M. Goryaeva, C. Lapointe, C. Dai, J. Dérès, J.-B. Maillet and M.-C. Marinica, *Nat. Commun.*, 2020, **11**, 4691.
65 C.-C. Fu, F. Willaime and P. Ordejón, *Phys. Rev. Lett.*, 2004, **92**, 175503.
66 M.-C. Marinica, F. Willaime and J.-P. Crocombette, *Phys. Rev. Lett.*, 2012, **108**, 025501.
67 D. A. Terentyev, T. P. C. Klaver, P. Olsson, M.-C. Marinica, F. Willaime, C. Domain and L. Malerba, *Phys. Rev. Lett.*, 2008, **100**, 145503.
68 A. Chartier and M.-C. Marinica, *Acta Mater.*, 2019, **180**, 141 – 148.
69 N. Mousseau, L. K. Béland, P. Brommer, J.-F. Joly, F. El-Mellouhi, E. Machado-Charry, M.-C. Marinica and P. Pochet, *J. At. Mol. Opt. Phys.*, 2012, **2012**, 1.
70 P. Ehrhart, P. Jung, H. Schultz and H. Ullmaier, *Atomic Defects in Metals*, Springer Nature, 1991, vol. 25.
71 L. Bonati and M. Parrinello, *Phys. Rev. Lett.*, 2018, **121**, 265701.
72 A. P. Bartók, R. Kondor and G. Csányi, *Phys. Rev. B*, 2013, **87**, 184115.
73 A. M. Goryaeva, J.-B. Maillet and M.-C. Marinica, *Comp. Mater. Sci.*, 2019, **166**, 200–209.
74 F. Noé, S. Olsson, J. Köhler and H. Wu, *Science*, 2019, **365**, eaaw1147.