



HAL
open science

Efficient and transferable machine learning potentials for the simulation of crystal defects in bcc Fe and W

Alexandra Goryaeva, Julien Dérès, Clovis Lapointe, Petr Grigorev, T D Swinburne, James Kermode, Lisa Ventelon, Jacopo Baima, Mihai-Cosmin Marinica

► To cite this version:

Alexandra Goryaeva, Julien Dérès, Clovis Lapointe, Petr Grigorev, T D Swinburne, et al.. Efficient and transferable machine learning potentials for the simulation of crystal defects in bcc Fe and W. *Physical Review Materials*, 2021, 5 (10), pp.103803. 10.1103/PhysRevMaterials.5.103803 . hal-03395915

HAL Id: hal-03395915

<https://hal.science/hal-03395915v1>

Submitted on 3 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Efficient and transferable machine learning potentials for the simulation of crystal defects in bcc Fe and W

Alexandra M. Goryaeva,^{1,*} Julien Dérès,¹ Clovis Lapointe,¹ Petr Grigorev,^{2,3} Thomas D. Swinburne,² James R. Kermode,³ Lisa Ventelon,¹ Jacopo Baima,¹ and Mihai-Cosmin Marinica^{1,†}

¹Université Paris-Saclay, CEA, Service de Recherches de Métallurgie Physique, 91191, Gif-sur-Yvette, France

²Aix-Marseille Université, CNRS, CINaM UMR 7325, Campus de Luminy, 13288 Marseille, France

³Warwick Centre for Predictive Modelling, School of Engineering,
University of Warwick, Coventry CV4 7AL, United Kingdom

(Dated: October 23, 2021)

Data-driven, or machine learning (ML), approaches have become viable alternatives to semi-empirical methods to construct interatomic potentials, due to their capacity to accurately interpolate and extrapolate from first principles simulations if the training database and descriptor representation of atomic structures are carefully chosen. Here, we present highly accurate interatomic potentials suitable for the study of dislocations, point defects and their clusters in bcc iron and tungsten, constructed using a linear or quadratic input-output mapping from descriptor space. The proposed quadratic formulation, called Quadratic Noise ML, differs from previous approaches, being strongly preconditioned by the linear solution. The developed potentials are compared to a wide range of existing ML and semi-empirical potentials, and are shown to have sufficient accuracy to distinguish changes in the exchange-correlation functional or pseudopotential in the underlying reference data, while retaining excellent transferability. The flexibility of the underlying approach is able to target properties almost unattainable by traditional methods, such as the negative divacancy binding energy in W or the shape and the magnitude of the Peierls barrier of the $\frac{1}{2}\langle 111 \rangle$ screw dislocation in both metals. We also show how the developed potentials can be used to target important observables that require large time-and-space scales unattainable with first principles methods, though we emphasize the importance of thoughtful database design and degrees of non-linearity of the descriptor space to achieve the appropriate passage of information to large-scale calculations. As a demonstration, we perform direct atomistic calculations of the relative stability of $\frac{1}{2}\langle 111 \rangle$ dislocations loops and three-dimensional C15 clusters in Fe and find the crossover between the formation energies of the two classes of interstitial defects occurs at around 40 self-interstitial atoms. We also compute the kink-pair formation energy of the $\frac{1}{2}\langle 111 \rangle$ screw dislocation in Fe and W, finding good agreement with density functional theory informed line tension models that indirectly measure those quantities. Finally, we exploit the excellent finite temperature properties to compute vacancy formation free energies with full anharmonicity in thermal vibrations. The presented potentials thus open up many avenues for systematic investigation of free energy landscape of defects with *ab initio* accuracy.

I. INTRODUCTION

Enabled by the constant increase in computational power, *ab initio* simulations regularly reveal new, previously hidden, aspects of defects in crystalline materials. Empirical potentials should constantly improve and integrate these new findings. In metals where the functional form of the density of states is relatively simple, embedded atom method (EAM) potentials [1–6] are widely applied. However, attempts to improve the performance of these potentials on some specific properties, very often, imply a degradation of others. Over the last three decades, efforts of the community to improve the empirical force fields yielded various potentials, such as modified EAM (MEAM) [7], bond-order potentials (BOP) [8], reactive force field (ReaxFF) [9], charge-optimized many-body potentials (COMB) [10], etc. Although these po-

tentials outperform the EAM formalism in terms of accuracy with respect to *ab initio* calculations, they also face transferability problems due to the rigid form of underlying physical formalism. In this context, proposing new fitting solutions and support functions with increased flexibility of the functional form inspired by the artificial intelligence (AI) and machine learning (ML) opens up new avenues to overcome the limitations of classical interatomic potentials. Moreover, AI/ML methods provide a possibility to bridge the gap between the less accurate, empirical potentials that scale as N^2 or lower (N being the number of atoms) and the more accurate electronic structure calculations that scale as N^3 or higher. Although electronic structure methods, such as tight binding or hybrid electronic structure-empirical methods (e.g., QM/MM), attempt to bridge this gap, they are not always successful and retain the unfavorable scaling. In this context, ML potentials are good candidates to enable large length-scale and time-scale calculations where the computational cost of *ab initio* methods does not allow for direct atomistic simulations and accuracy of empirical potentials is not sufficient.

* alexandra.goryaeva@cea.fr

† mihai-cosmin.marinica@cea.fr

The first attempt to couple AI and high-dimensional problems in atomic-scale materials science was proposed by Behler and Parrinello in 2007 [11]. In contrast to classical force fields where the performance and limitations of the potential are mainly defined by the physical formalism, performance and accuracy of ML potentials is determined by three equally important components: database, its representation in descriptor space, and the regression algorithm.

ML potentials require an extensive training database because its content has a strong impact on the accuracy and transferability of the potential. The design of the database, which includes the choice of relevant information, as well as the selection of pertinent instances [12–14], called sparsification, is a crucial step for obtaining a relevant ML potential.

The atomic descriptors enable a specific numerical representation of the atomic structures from the database and its invariant description with respect to symmetries of the Hamiltonian of the system (e.g., permutation of like atoms, rotation, translation). Thus, instead of using \mathbb{R}^{3N} -dimensional description of the local atomic environments, one employs a space \mathbb{R}^D . The dimension of descriptor space commonly ranges from few tens to few thousands [15–17]. Most commonly, atomic descriptors encode the local geometry on neighboring atoms using the distances and/or angles between atoms [11, 15, 18], spectral analysis of local atomic environments [15, 18] or a tensorial description of atomic coordinates [19, 20]. A systematic basis that preserves the symmetry of the potential energy function with respect to rotations and permutations can also be developed by writing the total energy as a sum of atomic body-ordered terms giving atomic body-ordered permutation-invariant polynomials [21, 22]. Some innovative descriptors, e.g., proposed by Mallat *et al.* [23, 24], are based on the scaling wavelets transformation. Quantum mechanics informed descriptors can be built on physical observables, such as Mulliken charges [25] or partial histograms of electronic density of states [26]. The similarity distance descriptors are based on the distances between pairs of atomic environments, e.g., smooth overlap of atomic positions (SOAP) [15] or graph version [27, 28] defined through a functional representation of atomic positions. The atomic cluster expansion (ACE) can be used in order to build a complete basis of invariant polynomials [29, 30] by combination of radial and spherical harmonic functions. In some cases, the framework of deep learning neural networks (NN) with a special design can be used to construct a pertinent descriptor of the system [31–34].

The fitting of ML potentials is performed in descriptor space and the (statistical) ML procedure of the fit defines the performance and limitations of the potential. The relationship between atomic energies and components of the descriptors can be linear [35–41] or non-linear [11, 41–53]. The linear model does not imply a linear relation between the phase space and the observable. Any non-linear regression becomes linear if the domain of

the function is projected into a space with a sufficiently large number of dimensions [54, 55]. Non-linear models are most commonly based on NN [11, 41–44] or kernel methods [46–53]. Using a linear kernel is equivalent to performing a linear regression while a polynomial kernel is equivalent to linear regression with a basis set formed from outer products of the elements of the feature vectors [56]. Some of the kernel models are formalized in the ever-growing field of the statistical on-the-fly learning methods [51, 53, 57], while the others are built in the form of potentials such as Gaussian approximation potentials (GAP) [18, 58]. The GAP is the widely used version of kernel potentials. Generally speaking, highly non-linear methods are suitable to interpolate multivariate functions while in the extrapolation regime they tend to give poor performance [54, 55]. One can partially overcome this inconvenience with the help of well-chosen regularization, constant augmentation of the database or by using on-the-fly active learning techniques [53, 57, 59] in order to constantly increase the boundaries of interpolation regime. Alternatively to ML force fields, many other ML approaches and surrogate models are designed to characterize defects in crystalline materials [14, 60–64].

In this work we aim to design ML force fields that are suitable for modeling radiation-induced defects in Fe and W and allow performing large-scale calculations. The energy landscape of defects in bcc Fe and W is extremely complex [65–73] and its accurate description at the atomic scale requires using appropriate force-field models that provide a correct description of atomic systems beyond the equilibrium conditions. The new potentials should have higher accuracy than traditional potentials for the essential properties of defects and, in addition to that, correctly reproduce the peculiar behavior of some small defects known from *ab initio* calculations, such as negative binding energy of the di-vacancies in W [74–80], the distinct energy landscape of C15 interstitial clusters in Fe [68, 72, 81–83], dislocation core structures as well as the shape and the magnitude of the Peierls barrier [84–87]. In addition to the improved accuracy, the developed ML potentials should be reasonably fast in order to enable large-scale calculations of defects. Therefore, we focus on linear ML (LML) models and its non-linear variants that are strongly preconditioned by the linear fit. Although these methods are not as accurate as non-linear kernel or NN methods for interpolation of the database, linear models have interesting advantages related to their robust behavior outside the fitting domain and computational cost that allows large-scale modeling.

The paper is organized as follows. The next Section II describes the methods used in this work, including the description of the density functional theory (DFT) database calculations (Sec. II A) and the statistical ML approaches used to fit the force fields and the details on the representation of the databases in the descriptor space (Sec. II B). Then, the Sec. III summarizes the main results of modeling formation energies and stability (Sec. III A) as well as the migration barriers (Sec.

III B) of small SIA and vacancy clusters in bcc Fe and W. In Section IV we further test the developed ML potentials for three challenging cases where the accuracy of the existing empirical potentials is not sufficient to provide reliable results and the computational cost of *ab initio* calculations does not allow for direct atomistic calculations. In Section IV A we investigate the $\frac{1}{2}\langle 111 \rangle$ screw dislocation in Fe and W and compute its Peierls barrier and kink-pair formation energy employing direct atomistic simulations. The computed formation energies of the kink-pair are compared with those from line tension (LT) models parametrized using *ab initio* calculations [86–88]. Further, Sec. IV B presents the results of direct atomistic calculations of the relative stability of large dislocation loops and C15 clusters in Fe. Finally, in Section IV C we present the first calculations of anharmonic free energy of the mono-vacancy formation in Fe and W using accurate ML potentials. This quantity is essential for the parametrization of large-scale calculations such as kinetic Monte-Carlo, cluster dynamics, etc. The calculations presented in Section IV are nowadays impossible to perform directly with *ab initio* as they either require simulation boxes larger than 100 000 atoms (Secs. IV B and IV A) or millions of force evaluations in order to accurately converge the sampling of the thermodynamic integration (Sec. IV C).

II. METHODS

A. DFT calculations of the database

The designed DFT databases for Fe and W contain atomic environments relevant for the physics of defects in materials under irradiation. For both materials, we take into account configurations of the perfect and distorted bcc structures; point, linear and extended defects, such as self-interstitial atoms (SIAs), vacancies, free surfaces, γ -surfaces, dislocations; and the liquid state. For most of atomic systems, we compute energies, forces and the virial stress. The full content of the databases and total number of observables to fit are detailed in Appendix A and Table V.

The databases are calculated with VASP [89] using projector augmented wave (PAW) pseudopotentials that account for 8 valence electrons [Ar]3d⁷4s¹ for Fe and 14 valence electrons [Xe4f¹⁴]5s²5p⁶6s¹5d⁵ for W. The magnetic state of bcc and liquid Fe is ferromagnetic at 0 K as well as at finite temperature. The exchange-correlation energy is evaluated using the Perdew-Burke-Ernzerhof (PBE) parametrization of the generalized gradient approximation (GGA). The plane wave energy cutoff is set to 500 eV and the Hermite-Gaussian broadening-width for Brillouin zone integration is 0.1 eV. The k -point grid of the Brillouin zone was chosen such that each configuration in the database has a similar density of k -points and corresponds to that of the cubic unit cells of Fe with $a_0 = 2.8327$ Å and W with $a_0 = 3.1854$ Å with

Monkhorst-Pack (MP) $20 \times 20 \times 20$ grid. The databases contain information from three types of DFT calculations: (i) structural optimization at 0 K; (ii) minimum energy pathways at 0 K; and (iii) finite temperature MD calculations. The 0 K minimization is performed using conjugate gradients until the maximum magnitude of the atomic forces becomes lower than 0.01 eV/Å. The minimum energy pathway (MEP) calculations are performed using the climbing image version of Nudge Elastic Band (NEB) method [90–92] with 7-9 images and the same criterion on the maximum force as above. The MD-DFT simulations sample finite temperature trajectories of bcc (perfect bulk or perfect bulk with few defects) and the liquid state. High temperature MD-DFT simulations are used to sample the Maxwell-Boltzmann distribution of positions at a given temperature. Including these calculations to the database are compulsory in order to have reliable vibrational properties for α -Fe [40, 93, 94]. The MD-DFT calculations are performed in the NVT ensemble. The time step of MD integration is set to 1.0 fs. The shape of the simulation boxes used for MD is cubic and is fixed to $4a_0 \times 4a_0 \times 4a_0$. The value of a_0 is set to 0 K lattice parameter from DFT calculations or (when it is specified in Table V) to the experimental lattice parameter, at the corresponding temperature, re-scaled with the ratio between the very low temperature experimental and DFT values. The chosen temperatures are 300 K and 800 K for Fe and 300 K, 1 000 K and 3 000 K for W.

B. ML force fields using linear and quadratic noise regression in the descriptor space

The foundation of any empirical potential concept states a correlation between the local atomic environment (LAE) and local atomic energy. Here, we use the local energy term as the energy accounted for from the atoms located in the neighborhood / LAE of a central atom within a cutoff distance R_c . The link between the total energy and the local atomic energy was established in the early days of atomistic materials science. For metals, the tight binding approximation [95–97] has formalised the basis of this relation. According to this formalism, the total energy E^s of the system s containing N^s atoms can be written as a sum of local energies ϵ_a of the a th atom:

$$E^s = \sum_{a=1}^{N^s} \epsilon_a. \quad (1)$$

It should be noted that the above form of the total energy is a crude approximation for systems where the electronic correlations are important or in the case where the charge screening is not very effective (, insulators) and the charge interaction between ions is not negligible. Such systems require more sophisticated formalism [98], e.g., taking separately into account long-range interactions beyond R_c , which is beyond the scope of this study.

1. Linear ML formalism

The present-day ML potentials propose a direct multivariate regression, in the descriptor space, between the LAE and the atomic energy. This linear proportionality was originally proposed by Thompson *et al.* [35] and used in many other studies [19, 20, 23, 24, 37–40, 99]. In the linear ML model, the target *ab initio* total energy E_{DFT}^s can be approached by the linear-ML energy E_{LML}^s through the linear local contributions $\epsilon_a^{(1)}$ [23, 35, 40]:

$$\epsilon_a^{(1)} = \beta_0 + \sum_{k=1}^K \beta_k D_k^{s,a}$$

$$E_{LML}^s = \left(N^s, \sum_{a=1}^{N^s} \mathbf{D}^{s,a} \right) \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix},$$

where $\mathbf{D}^{s,a} = (D_1^{s,a}, \dots, D_K^{s,a})$ is the descriptor vector with K components of the a th atom and $(\beta_0, \beta_1, \dots, \beta_K)^T$ are the parameters of the fit (β_0 is the constant energy contribution). The total energy descriptor for the system s becomes $\mathbf{D}_E^s = (N^s, \sum_a \mathbf{D}^{s,a})$ vector in the $\mathbb{R}^{1 \times (K+1)}$ space. The energy of the corresponding system is obtained by a linear regression using the $K+1$ parameters mentioned above. The atomic forces and the virial stress are obtained from the derivatives, with respect to the coordinates, of the energy, Eq. 2 (more details about derivatives are given in Ref. [40]).

2. Quadratic noise ML formalism

In order to keep the advantages of the linear interpolation that can behave wrong outside their fitting range, we put a precondition on the quadratic regime imposed by the linear interpolations. Here we propose to fit in quadratic regime only the deviation of the LML values from DFT. The reference DFT energy of the system s can be written:

$$E_{DFT}^s = E_{LML}^s + \Delta E \simeq E_{LML}^s + E_{QNML}^s, \quad (2)$$

with the last term E_{QNML}^s having a quadratic form in the descriptors elements, which will be called hereafter as quadratic noise ML (QNML):

$$E_{QNML}^s = \sum_{a=1}^{N^s} \epsilon_a^{(2)} = \sum_{a=1}^{N^s} \sum_{k=1}^K \sum_{k'=1}^K \alpha_{kk'} D_k^{s,a} D_{k'}^{s,a}, \quad (3)$$

and the local atomic energy can be written:

$$\epsilon_a = \epsilon_a^{(1)} + \epsilon_a^{(2)} \quad (4)$$

$$\epsilon_a^{(2)} = \sum_{k=1}^K \sum_{k'=1}^K \alpha_{kk'} D_k^{s,a} D_{k'}^{s,a}. \quad (5)$$

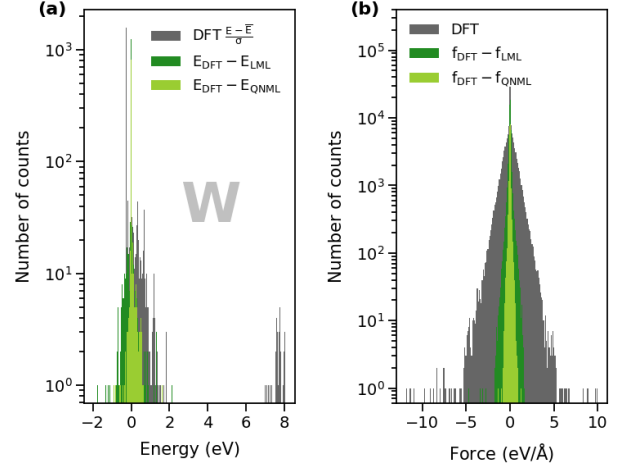


FIG. 1: Histogram distribution of the DFT data for energy (in eV), force (in eV/Å), and error noise deviation of LML and QNML force fields on W database. The absolute DFT energy is presented using re-scaled form $(E - \bar{E}) / \sigma$ where \bar{E} is the average energy of the database and σ is the standard deviation. The error distribution for energy and force has Gaussian shape. The distribution is narrower for QNML force field.

The above QNML development has a well-defined preconditioning, imposed by the linear fitting, i.e., the parameters α are determined after the parameters β of linear fit are fixed. The solution is given by the least-square estimate of the α parameters $\alpha = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}_{QNML}$. $\Phi \in \mathbb{R}^{M \times K^2}$, $\Phi_{s,kk'} = \sum_{a=1}^{N^s} D_k^{s,a} D_{k'}^{s,a}$ where s is an order number for the system s between the 1st and M th system of the training dataset. \mathbf{y}_{QNML} is a column vector $\in \mathbb{R}^{M \times 1}$, which contains the differences between the DFT and LML values of the total energies. As stated above, the quadratic fit in the $(K+1)$ dimensional descriptor space becomes a linear fit in higher dimensional descriptor space $\mathbb{R}^{1 \times (1+K+K^2)}$ for which the energy descriptor is $\mathbf{D}_E^s = (N^s, \sum_a \mathbf{D}^{s,a}, \sum_a \mathbf{D}^{s,a} \otimes \mathbf{D}^{s,a})$. Similar to the linear case [36, 40, 45], the forces and virial stress can be included in the quadratic formalism. The force acting on atom b is computed from the total ML energy, Eq. 2. For example, in QNML case, the descriptor associated with the force acting on atom b in the direction α becomes: $\mathbf{D}_F^{s,b\alpha} = -(0, \sum_a \nabla_{ba} \mathbf{D}^{s,a}, \sum_a (\nabla_{ba} \mathbf{D}^{s,a} \otimes \mathbf{D}^{s,a} + \mathbf{D}^{s,a} \otimes \nabla_{ba} \mathbf{D}^{s,a}))$. The above force descriptors lie in the same $\mathbb{R}^{1 \times (1+K+K^2)}$ space and require non-trivial evaluation of derivatives of each atomic descriptor with respect to all atomic Cartesian coordinates. In practical implementation, the descriptor functions have a finite cutoff radius and thus, the sum over N atoms is replaced by the sum of the neighbors of the b th atom within the cutoff. The techniques for regularization or weighted fit useful for the linear case [40] can be also employed in the case of

QNML.

We emphasize that the proposed QNML formalism is different from quadratic SNAP (qSNAP) [45]. While qSNAP potentials are fitted as a polynomial of second degree with explicit linear and quadratic form, in QNML only the error of linear fit (LML) is treated as a quadratic form of atomic descriptors. Moreover, QNML procedure is inspired by the nearly Gaussian shape of error distribution of the linear fit. As showed in Figure 1, the distribution of noise $E_{\text{DFT}}^s - E_{\text{LML}}^s$ is close to uni-modal Gaussian, while absolute DFT energies in our database exhibit a bi-modal shape. Such distribution of DFT data, without any treatment, can lead to overfitting or highly heterogeneous parametrization. This error of the LML fit is the target of the QNML formalism. In order to avoid the inherent overfit that is associated to non-linear approach appropriate regularization (e.g., \mathcal{L}^2 ridge, Bayesian etc.) techniques can be applied [54, 55, 100].

Using the QNML procedure implies a significant increase in the number of parameters (which becomes in the order of K^2) compared to the linear case. For the descriptors with large dimensions, such as MTP or SOAP (with K higher than 10^3), this approach is not practical: an important numerical effort should be made in the fitting procedure in order to perform regularization of parameters and to handle the design matrix, which will require a huge amount of memory. For such large-dimensional descriptors, the linear or kernel formalism are much better adapted as the range of the design matrix and the number of parameters are order of K or M , respectively. The suggested QNML approach is well adapted for compact descriptors with $D < 100$ components, such as angular Fourier series (AFS), bispectrum SO(4) [15], hybrid descriptors [40], or quantum mechanics informed descriptors [25].

3. Fitting of the potentials

In this study, Fe ML potentials are trained on 761 and tested on 300 DFT structures, providing, respectively, 111 683 and 41 496 observables with energy, forces, and stress ($n_E + n_F + n_S$). In case of W, ML potentials are trained on 2 132 and tested on 607 DFT structures, providing 131 057 and 22 543 observables, respectively. More details on the database content and the observables in each structural class of the database are provided in Appendix A and Table V. The setup of DFT calculations is described in Sec. II A.

The database is represented using bispectrum SO(4) descriptor with the angular moment $j = 4$, leading to $K = 55$ components. The cutoff function is the same as in the implementation from Ref. [40]. The cutoff distance R_c is set to 4.7 Å and 5.3 Å for Fe and W, respectively.

The fit is performed using Moore-Penrose pseudoinversion of the design matrix [101]. In order to establish a hierarchy in the importance of the target observables, a weighted fit is performed. Accounting for the

TABLE I: Elastic properties of bcc Fe and W provided by the developed LML and QNML potentials and their comparison with the reference DFT values, obtained from DB class 2 in Table V. The bcc lattice parameter a_0 and the elastic constants are reported in Å and GPa, respectively. MAEs of LML and QNML denote the corresponding fitting mean squared error for the energy per atom (MAE_E , in meV), forces (MAE_F , in meV/Å) and stress (MAE_S , in meV/Å³). MAE listed for DFT correspond to the estimated accuracy of the given exchange-correlation functional, due to the mismatch in k -point meshing, Fourier grid, etc. between the simulation cells of different size.

	Fe			W		
	LML	QNML	DFT	LML	QNML	DFT
a_0	2.8325	2.8327	2.8327	3.1855	3.1854	3.1854
B	194.9	193.8	193.6	303.2	304.1	304.5
C_{11}	287.7	293.6	292.3	509.8	516.7	516.6
C_{12}	148.5	144.0	144.3	200.0	197.8	198.5
C_{44}	120.5	102.3	102.1	144.5	137.2	140.2
MAE_E	33	10	3	5	3	3
MAE_F	70	33	5	95	50	5
MAE_S	25	19	4	10	10	4

weights (that are the diagonal elements of the diagonal matrix \mathbf{W}), the objective function can be written as $J(\boldsymbol{\alpha}) = \|\mathbf{W}^{1/2}(\mathbf{y} - \Phi\boldsymbol{\alpha})\|^2$ and the solution of the potential parameters is taken accordingly [40]. The magnitude of weight associated with a subset of the database (indicated as DB class in Table V) controls the accuracy of the fit for that particular property [40]. For some components of the database, the weights are fixed a large value (e.g., lattice parameter and elastic constants), while others are optimized, using a genetic algorithm, through an objective loss function that includes the MAEs for energy, forces and stress. This optimization should avoid the risk of over-fitting as it is described in [40]. The solution is regularized using \mathcal{L}^2 Tikhonov procedure, i.e., $J(\boldsymbol{\alpha})^{reg} = J(\boldsymbol{\alpha}) + \lambda \|\boldsymbol{\alpha}\|^2$ employing a unique regularization constant of $\lambda = 10^{-6}$ for energy, force, and stress. This value was found using a grid search in the range from 10^{-8} to 10^2 .

We note that in the framework of present ML potentials short-range interactions are learned by including various configurations of self-interstitial defects and strongly deformed and compressed bulk structures (see Appendix A and Table V) and no additional potentials for very short-range interactions are included into the model. In order to ensure robust performance of the potentials for modeling displacement cascades or Frenkel pair accumulation, it is recommended to couple the ML potentials with a short-range interaction model.

The comparison of computed lattice parameters and elastic constants with the reference DFT values is provided in Table I. Compared to LML potentials, QNML systematically provides the values closer to DFT. Overall, the LML and QNML fit provides reasonably low values

for the mean average error (MAE) of energy per atom, force and virial stress (see Table I). The QNML potentials have MAEs that are 30-50% lower than those from LML fit. To the best of our knowledge, there are no traditional potentials able to reach such low values of MAE. However, the parameter-free ML kernel formalism, e.g., GAP potentials, can outperform this accuracy of fit being up to five times lower in MAE than LML.

The present LML and QNML potentials are developed using Machine Learning Dynamic (MILADY) package [40]. The potentials together with the LAMMPS-MILADY module necessary to perform the simulations in LAMMPS [102] are available at GitHub repository [103].

III. RESULTS

In bcc metals, and in particular in Fe and W, the defect energy landscape is very complex [70, 73]. Both metals accommodate a rich morphology of defects with size-dependent stability and mobility. This section firstly profiles the performance of the developed ML potentials for computing small defect clusters with interstitial and vacancy character. The 0 K formation and migration energies of small defects are reported in Secs. III A and III B, respectively. Further, the performance of the ML force fields is examined for some very challenging cases. We test the ML potentials ability to perform large-scale calculations, far beyond the limits in length scale and time scale of DFT calculations, in order to study the kink-pair formation energy (Sec. IV A), size-dependent stability of C15 clusters in Fe (Sec. IV B), and the anharmonic free energy landscape of bulk and mono-vacancies (Sec. IV C) in bcc Fe and W.

A. Formation energies and relative stability of small defect clusters

This section summarizes the performance of the fitted ML potentials to compute the formation energies and relative stability of small vacancy and self-interstitial defect clusters in bcc Fe and W.

1. Vacancy clusters

The energy landscape of small defect clusters in Fe and W significantly differs from each other. The binding energy of V_n vacancy clusters is always positive in Fe, while di-, tri- and quadri-vacancy clusters in W can have a negative binding energy. In this work, the formation energies of vacancy clusters V_n calculated with LML/QNML potentials are close, within ± 0.15 eV in both cases, to the DFT values. Such good performance of the fitted ML potentials results from the carefully constructed database in combination with the low values of MAEs (Table I) that

ensure the accuracy of potentials with respect to the data included in the fit.

The di-vacancy in tungsten, a bcc metal of the VI B group, has an unusual energy landscape [77]. According to the *ab initio* calculations, the first nearest neighbor (1NN) di-vacancies in W can be slightly repulsive or attractive, depending on the approximation used for the DFT exchange-correlation functional, while the second nearest neighbor (2NN) configuration is strongly repulsive [77, 78]. The same tendency is observed for all elements in the VI B group. The classical mechanism of void formation implies firstly the formation of small vacancy clusters like V_2 . The unstable character of 2NN vacancies in W might imply the revision of the void formation mechanism in this material. Vacancy clusters are evidenced in the low temperature limit by direct observations using transmission electronic microscopy [104, 105]. Those clusters form by precipitation of vacancies during quenching and subsequent annealing during the experiments. Thus, an interatomic potential that accurately predicts the stability of V_2 clusters in various configurations is necessary to clarify the mechanism of void formation in W. The computed binding energies of 1NN and 2NN di-vacancies in W are negative for all ML potentials developed until now, including LML, QNML or GAP, in accordance with the predictions of the DFT calculations [77, 78]. It is worth mentioning that there are no EAM potentials, except [79], which are able to reproduce such behavior. In this work, information about the repulsive character of V_2 is directly included in the fitting database. For this particular problem of vacancy clusters in W, we have tested the ability of ML potentials to accurately compute the clusters of up to 10 vacancies. The tested V_n , $n \leq 10$ clusters are similar to the configurations presented in Mason *et al.* [79]. The results in Fig. 2a demonstrate that the formation energies predicted by all tested ML potentials correlate very well with the present DFT GGA-PBE calculations. The potentials GAP14 [94] and GAP19[106] slightly deviate from the perfect correlation. The predictions from LML and QNML potentials fall between the results of the two GAP potentials. The ML potentials studied here are designed employing databases (Tab. V) that contain information about the structural geometry of vacancies, namely, minimum energy configurations of $V_{1,2,3}$ and short molecular dynamics trajectories around those minima, as well as larger V_{8-16} clusters and free surfaces with various orientations.

The analysis of the binding energy of vacancy clusters in W shows that all the ML potentials follow the trend given by DFT (Fig. 2a). The main differences are imposed by the value of the formation energy of the V_1 , which is 3.16 eV for LML, 3.19 eV for QNML and 3.24 eV for DFT. It worth noting also the differences between the ML potentials and the DFT reference (Fig. 2b) are close to those between the DFT calculations that use *the same* exchange-correlation functional and *k*-points sampling but employ different pseudopotentials (DFT-

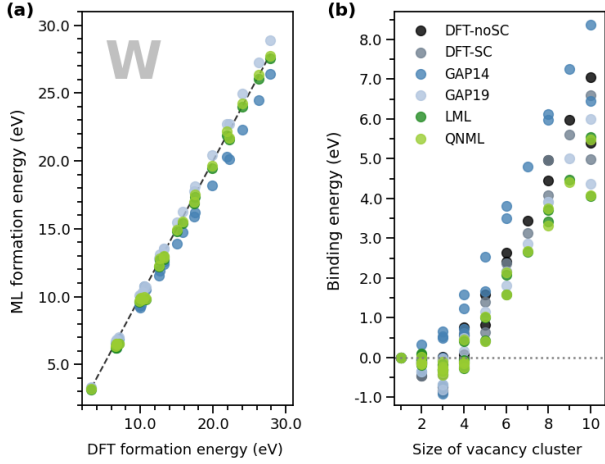


FIG. 2: Formation energies of vacancy clusters V_n (from $n=1$ to 10 vacancies) in bcc W computed with LML, QNML, GAP14 [94], GAP19 [106] potentials compared with the DFT GGA-PBE values from this work. (a) Correlation between the formation energies computed with ML potentials and with DFT. The dashed black line indicates a perfect correlation. (b) Binding energies of the vacancy clusters $E_b^n = E_f^n - n \times E_f^1$, where E_{for}^n and E_f^1 are the formation energies of a cluster with n defects and of a single defect, respectively. A negative value of the binding energy implies energetic instability and dissociation of the defect cluster. DFT-SC and DFT-noSC correspond to the DFT calculations using the pseudopotentials with and without semi-core states, respectively.

SC and DFT-noSC in Fig. 2b with and without semi-cores states, respectively).

2. Interstitial clusters

As in the case of vacancies, the characteristics of SIA clusters in Fe and W differ considerably from each other. For single SIAs, resistivity recovery experiments [69, 107], DFT calculations [66, 71, 108–110], EAM [72, 77, 81, 110–113], ML-GAP [93, 94] and the present ML potentials consistently indicate that the I_1 dumbbells with $\langle 110 \rangle$ orientation in Fe and $\langle 111 \rangle$ orientation in W are the most stable. In W, the energy landscape of SIAs predicted by DFT calculations or accurate physics-informed models, such as the discrete-continuum model (DC) [71], suggests that variants of $\frac{1}{2}\langle 111 \rangle$ dislocation loops [67, 114] are the lowest energy configurations at all sizes. Similar results are provided by the present LML/QNML potentials, as well as by the other ML potentials investigated here and, with few exceptions, by the EAM potentials (for more details see Ref. [71]).

The energy landscape of SIA clusters in Fe is more complex [71, 110, 111] than that in W. With increasing

TABLE II: Formation energy of SIA in Fe, I_n clusters (from $n=1, 2$ and 5). LE stands for the lowest energy interstitial defects. The value in parenthesis is the relative energy with respect to corresponding formation energy of the $I_n^{\langle 110 \rangle}$. $\Delta I_n^{\langle 111 \rangle}$ and $\Delta I_n^{\langle C15 \rangle}$ columns give the difference between the formation energy of $I_n^{\langle 110 \rangle}$ and the formation energy of the $\langle 111 \rangle$ and $C15$ clusters, respectively. When a reference work is not specified, the DFT values are provided by the present calculations.

Fe mono-interstitial, I_1 (eV)			
	LE	$\Delta I_1^{\langle 111 \rangle}$	
AM04[115]	$\langle 110 \rangle$	0.49	
MA07[81, 83]	$\langle 110 \rangle$	0.72	
GAP18[93]	$\langle 110 \rangle$	0.82	
LML	$\langle 110 \rangle$	0.69	
QNML	$\langle 110 \rangle$	0.72	
DFT	$\langle 110 \rangle$	0.74	
Fe di-interstitial, I_2 (eV)			
	LE	$I_2^{\langle 111 \rangle}$	I_2^{C15}
AM04[115]	$\langle 110 \rangle$	0.68	1.34
MA07[81, 83]	$\langle 110 \rangle$	1.15	0.35
GAP18[93]	$\langle 110 \rangle$	1.57	1.98
LML	Gao (-0.03)	1.05	0.64
QNML	Gao (-0.18)	1.12	0.62
DFT	Gao (-0.13)	1.21	0.64
Fe quadri-interstitial, I_4 (eV)			
	LE	$I_4^{\langle 111 \rangle}$	I_4^{C15}
AM04[115]	C15	0.51	-1.51
MA07[81, 83]	C15	0.63	-3.70
GAP18[93]	C15	1.23	-0.90
LML	C15	0.92	-3.42
QNML	C15	1.07	-2.45
DFT	C15	0.83	-2.49

size, SIA clusters in Fe change their relative stability. For the traditional 2D loops with more than 4 SIAs, the DFT studies reveal that $\langle 111 \rangle$ clusters become more stable than $\langle 110 \rangle$ family [69, 71, 110]. The present LML / QNML potentials predict the same crossover as DFT between the two families of loops. The ML GAP18 potential [93] gives slightly overestimated formation energies with respect to the same DFT calculations. Overall, for the small 2D loops, GAP18 [93] provides results very similar to those from the MA07 EAM potential [81, 83] and predicts the transition between the $\langle 111 \rangle$ and $\langle 110 \rangle$ loops after accumulation of 5 SIAs, which is slightly larger than the cluster sizes obtained by DFT and AM04 EAM potential [115].

DFT calculations (Tab.II) predict that the lowest I_2 configuration in Fe is the so-called triangle or Gao configuration I_2^{Gao} [116]. These clusters are known to be immobile [82]. Moreover, I_2^{Gao} represents an elementary building brick for complex 3D SIA clusters with C15 structure [72, 83]. Until now, there are no empirical potentials that are able to predict the Gao configuration as SIAs with the lowest energy. The LML and QNML potentials developed here predict the correct relative stability of I_2^{Gao}

clusters (Tab. II).

Formation of 3D interstitial clusters with the C15 structure is specific to bcc Fe [70, 71, 83, 117]. The fitted ML potentials accurately reproduce the relative stability of the 3D C15 clusters and 2D dislocation loops predicted by DFT [71, 83, 117] (Tab. II). The C15 are the most stable configurations of the I_n clusters with $n > 3$ SIAs. The fitted ML potentials predict that the C15 energy relative to the parallel cluster, for 2 and 4 SIAs, is close to the DFT counterpart, which ensures that the characteristics of the formation and the concentration of C15 clusters are similar to those found with DFT. For I_2^{C15} clusters, the GAP18 potential [93] predicts formation energies that are much higher than those of dumbbell configurations. The predicted formation energy of I_2^{C15} is about 2.0 eV higher than that of $\langle 110 \rangle$ dumbbells, yielding an impossible formation of C15 in Fe. Moreover, for larger clusters, GAP18 [93] predicts less stable C15 clusters. These results are not consistent with the DFT calculations [72, 83]. The EAM AM04 potential poorly reproduces the energy landscape of C15, while MA07, is much closer to the DFT results [72, 112, 118]. In contrast to the 2D loops, the tested GAP18 potential [93] exhibits a limited transferability for the C15 clusters and provides results close to those from AM04. It is interesting to note that, at variance with the present LML and QNML potentials, both traditional and ML GAP interactions are not designed for C15 clusters (i.e., the C15 clusters are not explicitly included in the potential databases). However, the fitting databases of both potentials contained the Gao di-interstitial configuration I_2^{Gao} [72]. A detailed analysis based on outlier detection and distortion score shows that the components related to C15 clusters are missing in the ML GAP database [14]. As a possible solution to improve the performance of the GAP potential for C15 clusters, one may consider including 3D cluster structures into the training databases in order to enrich the variety of atomic environments known by the model. In Section IV B, we will further test the performance of the LML and QNML potentials to predict size-dependent relative stability of large C15 clusters and 2D dislocation loops in Fe.

B. Migration energies of small defect clusters

In this section we outline the performance of the fitted ML potentials to compute the migration barriers of small vacancy V_{1-3} and self-interstitial I_{1-3} clusters.

1. Vacancy clusters

The migration energy barriers of mono-vacancies V_1 in Fe computed using the new LML/QNML potentials and previously developed GAP18 potential [93] are remarkably accurate compared to those provided by existing EAM potentials (Table III). Moreover, the shape of

TABLE III: The migration barriers, ΔE_m (in eV), of the mechanisms with lowest DFT barriers for the small SIAs (I_{1-3}) and vacancy clusters (V_{1-3}). The barriers marked with * have a double hump shape.

Fe self-interstitials ΔE_m (eV)			
	I_1	I_2	I_3
AM04[115]	0.30	0.33	0.32
MA07[81, 83]	0.27	0.32	0.34
GAP18[93]	0.32	0.37	0.31
LML	0.35	0.35	0.49
QNML	0.33	0.42	0.37
DFT [81]	0.34	0.43	0.43
Fe Vacancy ΔE_m (eV)			
	V_1	V_2	V_3
AM04[115]	0.64*	0.71	0.55
MA07[81, 83]	0.68	0.71	0.43
GAP18[93]	0.67	0.54	0.14
LML	0.67	0.65	0.33
QNML	0.68	0.69	0.33
DFT [81]	0.67	0.62	0.35
W Vacancy ΔE_m (eV)			
	V_1	V_2	V_3
MN17[79]	1.52	1.63	1.79
GAP14[94]	1.77*	1.75	0.70*
GAP19[106]	1.73	1.42	0.69*
LML	1.82	1.74	1.31
QNML	1.83	1.77	1.13
DFT [79]	1.76	1.83	1.15
DFT [80]		1.68	

the mono-vacancy barrier computed with the ML potentials exhibits a clear single saddle point, which is not the case for the EAM potentials that have ‘double humped’ migration profiles. Such a shape occurs because the jumping atom migrates along the $\langle 111 \rangle$ direction through the center of two equilateral triangles (at $\frac{1}{3}a_0\langle 111 \rangle$ and $\frac{2}{3}a_0\langle 111 \rangle$) lying in (111) plane. The traditional force fields, such as EAM, do not favor this symmetric triangle geometry, whilst the ML potentials are able to fit correctly these subtle structures. Similar effects are observed for the migration of the mono-vacancy in W.

For migration of di-vacancy clusters V_2 , the ML potentials also predict single saddle-point curves. The energy barriers computed for V_2 both with ML potential and traditional potentials compare well with the DFT calculations. The remarkable agreement between LML/QNML and DFT calculations are reported in Tab. III.

In order to examine the transferability of fitted ML potentials we test their ability to compute migration barriers that are not included in the training database, such as migration of tri-vacancies V_3 in Fe and W. A big error in these migration barriers will have a strong impact on the predictions of defect kinetics under irradiation and interpretation of processes during resistivity recovery experiments. The experimental stage IV of the vacancy in Fe and W will be strongly impacted by the fast diffusion of the vacancy clusters V_n with $n > 2$ [69, 107], which

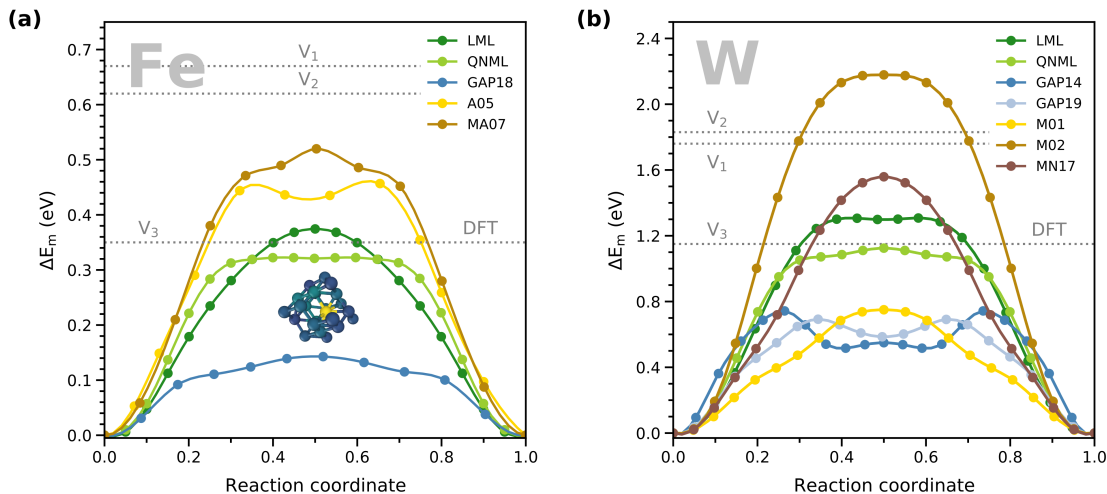


FIG. 3: Migration barriers of tri-vacancy clusters V_3 in (a) bcc Fe and (b) bcc W. The comparison is made between the LML and QNML potentials, GAP [93, 94, 106] and some commonly used EAM potentials for Fe [81, 83, 115] and W [79, 83]. The gray dotted lines labeled with V_{1-3} refer to the energy barriers for migration of mono-, di- and tri-vacancies from DFT calculations in Fe[81] and W [79]. The saddle point structure in the inset plot of (a) is detected using the distortion score of local atomic environments [14]. The atoms are colored according to the magnitude of the distortion score (yellow - high, green - medium, blue - low).

will affect the predictions of the size and the density of vacancy clusters above 300 K.

The saddle point of 1.15 eV for V_3 migration in W, as it was previously mentioned in Mason *et al.* [79], is very low compared to the migration barrier of $V_{1,2}$, around 1.7 eV. For tri-vacancies V_3 , LML and QNML have rather good transferability and provide correct migration barriers (Fig. 3), both in terms of energy and single saddle-point shape, whilst for the GAP class potentials, the error reaches up to 60% with the barrier having a nonphysical shape (Fig. 3). In this case, the highly non-linear character of GAP formalism yields a poor transferability. The results reported in Fig. 3 and Table III demonstrate that the performance of some of the EAM potentials for W and Fe is comparable with that of GAP potentials in the extrapolation regime. In order to improve the performance of GAP potentials in such cases, it is essential to enrich the training database and to include the saddle point configurations of V_3 migration in the training data set. The missing configurations can be revealed at the stage of the database design using the distortion scores based on outlier analysis of the database. This method was recently demonstrated to be very efficient for a similar problem with GAP potential for Fe [14].

2. Interstitial clusters

The migration of SIAs in W occurs along the $\langle 111 \rangle$ direction and the order of magnitude of this migration barrier is tens of meV [67, 114, 119–122]. Such low migra-

tion barriers do not play a critical role in the simulation of SIAs, except at cryogenic temperatures [123, 124]. In the high temperature limit, the movement of $\langle 111 \rangle$ loops is dominated by stochastic motion along the axis of the surrounding cylinder glide [66, 120]. These characteristics are well reproduced both by the present and existing interatomic potentials for W.

In bcc Fe, the dumbbells can migrate from their initial position to the next site via several different jump mechanisms [69, 109, 125]. The lowest migration barriers for $I_{1,2,3}$ in Fe correspond to Johnson’s mechanism of nearest-neighbor translation-rotation [69, 110, 125]. The computed migration barriers for this mechanism are reported in Table III and compared with other interatomic potentials. Among the potentials compared, the LML/QNML and GAP potentials perform better than any EAM potential and predict energy barriers within 10% of the DFT values. The QNML potential performs slightly better than the LML and GAP18 [93] for I_2 and I_3 clusters. The good performance of LML/QNML for the migration of small clusters in Fe results from the thoughtful design of the training database, where we have explicitly introduced the saddle point configuration for I_1 migration. In case of the GAP18 [93], the exact trajectories of the jump mechanisms are not explicitly included into the training database and the good performance of the potential was ensured by including a rich variety of single SIA atomic environments.

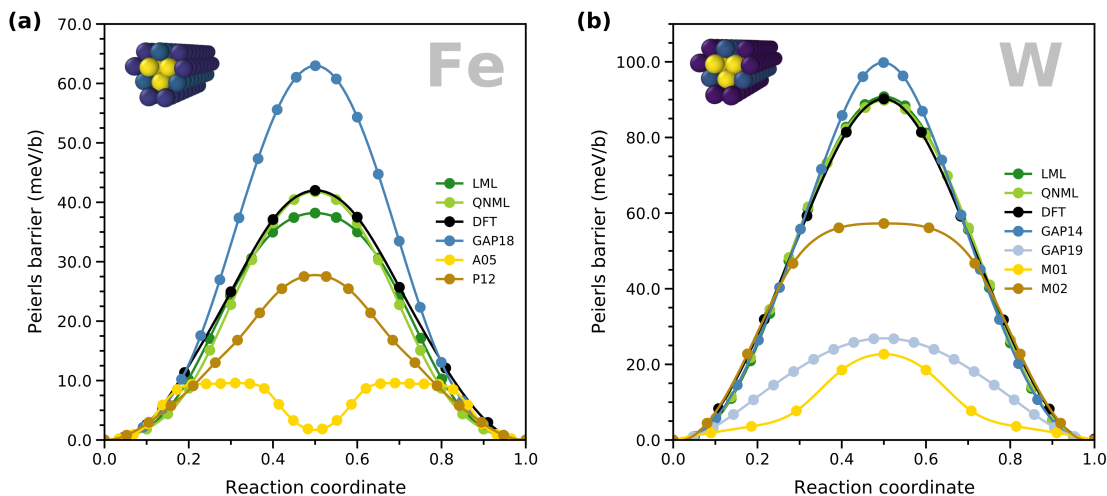


FIG. 4: Peierls barriers of the $\frac{1}{2}\langle 111 \rangle$ screw dislocation gliding in the $\{110\}$ plane (a) in Fe and (b) in W. The comparison is made between the LML potentials, DFT calculations [86, 87], GAP [93, 94, 106] and some commonly used EAM potentials for Fe [115, 126] and W [83]. The higher Peierls barrier in bcc Fe predicted by GAP18 [93] results from different DFT calculations in the training database of the potential.

The dislocation core structures in the inset plot are detected using the distortion score of local atomic environments [14]. The color of the atoms is set according to the magnitude of the distortion score (yellow - high, green - medium, blue - low).

IV. LARGE-SCALE APPLICATIONS

In this section, we address large-scale calculations that are challenging or impossible to perform using first principles calculations and compare them with the results from accurate multiscale models that were previously designed to overcome the limitations of *ab initio* methods. These large length scale calculations employ systems containing 10^{5-6} atoms. Section IV A provides the essential characteristics of the screw dislocations in bcc Fe and W, such as the formation energy of the kink-pair, obtained directly from atomistic calculations. Section IV B reports calculations of relative stability of large C15 clusters and dislocation loops. The large time scale calculations are described in Section IV C, which provide the accurate anharmonic free energy calculations of perfect bulk and mono-vacancy that require millions of force evaluations.

A. Peierls barrier and kink-pair formation energy of the screw dislocation

Here we present the performance of the developed ML potentials to compute the essential properties of the $\frac{1}{2}\langle 111 \rangle$ screw dislocation in bcc Fe and W: the core structure, the Peierls barrier, and the kink-pair formation energy. Direct DFT calculations can provide the dislocation core structure and the Peierls barrier [84, 85], whilst the formation energy of the kink-pair requires large simulation cells with more than 100 000 atoms [87, 88]. In order to compute the kink-pair formation enthalpy in bcc metals with DFT, Proville *et al.* [86, 88] proposed a method

based on the parametrization of one-dimensional line tension (LT) model from atomistic calculations performed in small simulation cells. This model was applied to bcc Fe and W [87, 88] in order to determine the kink-pair formation enthalpy at different applied stresses. Our direct calculations of the kink-pair with ML potentials will be compared to the results of the LT models.

The developed ML potentials yield a non-degenerate dislocation core structure, as predicted by DFT calculations [77, 84, 87, 127–129]. The structure of the dislocation core is analyzed using the distortion score of local atomic environments [14]. The dislocation cores in Fe and W are represented in the inset of Fig. 4. The geometry of the cores together with the magnitude of the distortion scores indicate the non-degenerated symmetry of the core structure.

The Peierls barriers of the $\frac{1}{2}\langle 111 \rangle$ screw dislocation in Fe and W gliding in the $\{110\}$ plane are reported in Fig. 4. In order to remain consistent with the reference DFT calculations, the barriers are calculated in systems containing 135 atoms with a quadripolar periodic array of dislocation dipoles, similar to that used in Refs. [86, 87]. The two dislocations were displaced simultaneously in the $\{110\}$ plane in the same $\langle 112 \rangle$ direction such as their separation distance remained constant. The minimal energy pathway is computed using the NEB implementation available in LAMMPS [91, 102]. All the Peierls barriers from our ML potentials have a single-humped profile, in agreement with DFT results [86, 87]. This shape of the barriers is also predicted by the GAP potentials [93, 94, 106]. In bcc W, the magnitude of the Peierls barrier predicted by LML and QNML is very close to the reference DFT values (Fig. 4b). In bcc Fe, LML exhibits

a slightly underestimated barrier compared to QNML potential. The higher Peierls barrier in bcc Fe predicted by the GAP18 potential [93] results from different DFT calculations that were used to train the potential and is not related to the kernel formalism of the potential. The ML algorithm that underlays the GAP potential, Gaussian Process regression, is non-parametric and can integrate all the information provided by the projection of database into the descriptor space \mathbb{R}^D .

As it was previously pointed out in Ref. [14] based on the distortion score analysis, the screw dislocation structure, including the core and saddle point configurations, can be accurately predicted by any kernel method that is fitted on a training database that contains the appropriate stacking faults, as was done for GAP18 [93]. The disagreement of the GAP19 potential for W with the present ML and DFT calculations is related to the fact that the database of that potential does not contain any instance that is close, in the descriptor space, to the structure of the saddle point of the $\frac{1}{2}\langle 111 \rangle$ dislocation along the glide plane. Consequently, the core of the saddle point is an outlier with respect to the training database, which yields poor transferability for the given structure. This transferability issue has the same origin as the case of V_3 migration, described in Section III B.

The kink-pairs are computed directly in large simulation cells that contain a dipole of $\frac{1}{2}\langle 111 \rangle$ straight screw dislocations with and without a kink-pair on each dislocation line. Half of the difference of the total energies of the two systems gives the kink-pair formation energy. The simulation cell is oriented such that the glide plane is a horizontal $[1\bar{1}0]$ plane, the dislocation line is along the $[111]$ direction and the glide occurs along the $[\bar{1}\bar{1}2]$ axis. In order to reduce the finite-size effects [88], we use simulation cells with the length $200\mathbf{b}$ ($\mathbf{b} = a_0\sqrt{3}/2$) along the dislocation line, which contain 243 000 atoms. The relaxation is performed using conjugate gradients until the maximum force of the system is lower than 10^{-3} eV/Å. The obtained values for the kink-pair energies are $E_{Fe}^{kp} = 0.77$ eV (LML) / 0.84 eV (QNML) and $E_W^{kp} = 1.42$ eV (LML) / 1.65 eV (QNML), for Fe and W, respectively. The comparison with the experimental values is not straightforward. The available experimental results [130–132] are extracted from three different regimes with various magnitudes of stress and temperature. Those experimental regimes are interpreted using elastic models to extract the different values for the kink-pair enthalpy. The extended ranges over the three regimes are $0.6 - 0.91$ eV [130, 131] and $1.3 - 2.05$ eV [132] for Fe and W, respectively. As in Ref.[86], here we retain as the reference values the kink-pair formation enthalpies measured at the lowest stresses and higher temperatures without any additional theoretical assumption from elasticity, providing 0.91 eV and 2.05 eV for Fe and W, respectively. However, it is worth noting that the high temperature regime can change the magnitude of the kink-pair formation enthalpy [133] biasing the comparison.

Further we compare the present values of the kink-

TABLE IV: Kink pair formation energies (in eV) computed with LML and QNML potentials in comparison with line tension (LT) models parametrized with DFT calculations

	LML	QNML	LT-Models
Fe	0.77	0.84	0.73-0.91 [86, 88, 134]
W	1.42	1.65	1.54 [86]

pair formation energy with the 0 K values obtained from the line tension (LT) models [88, 134] parametrized with the *ab initio* calculations. In Fe, Proville *et al.* [88] parametrized the LT model using localized orbital basis as implemented in the SIESTA code, resulting in $E_{Fe}^{kp} = 0.86$ eV. To construct the training database, in this study we used VASP and PBE-GGA for exchange-correlation functional and equivalent of 20^3 MP k -point mesh for the cubic bcc unit cell. Itakura *et al.*[134], using a plane wave basis (VASP) and different LT method, obtained $E_{Fe}^{kp} = 0.73$ eV, employing the same PBE-GGA and a MP k -point grid close to 20^3 . Based on the same LT model as in Ref. [88], but with different *ab initio* parametrization, Dezerald *et al.* [86] computed $E_{Fe}^{kp} = 0.91$ eV and $E_W^{kp} = 1.54$ eV for Fe and W, respectively. For both metals, the k -point grid in Ref.[86] is equivalent to 16^3 MP k -points mesh. Moreover, for W there are no semi-core electrons. Thus, in Fe, the differences between various DFT calculations used to parametrize the LT models provide a range of 0.73 eV - 0.91 eV for the kink-pair energies predicted theoretically (Table IV). Using GAP18 potential [93] for Fe, Maresca *et al.*[135] predicted a kink-pair energy of 1.0 eV. This overestimation is likely related to the magnitude of Peierls barrier, which is higher than all the other DFT predictions (but in accord with the DFT calculations on which the GAP18 potential was fitted). According to the LT model considering a sinusoidal shape of Peierls barrier [86, 126, 136], E^{kp} is proportional to the square root of the Peierls barrier. Then, assuming that a LT of dislocation line predicted by DFT and GAP18 is similar, having a 40% lower Peierls barrier in agreement with other DFT calculations (Fig. 4) would result in the value of the kink-pair formation energy around 0.80 eV. This value falls in the range of $0.73 - 0.91$ eV predicted by the DFT parametrized LT models [86, 88, 134].

The values of the kink-pair energy in Fe and W found with the ML potentials developed here are within the limits of the existing models and our results are close to the LT model predictions [86]. These results demonstrate that fitting the potentials with LML and QNML formalism based on database with small periodic cells provides a robust force field that can be used for larger systems with similar atomic environments. The accuracy of ML potentials for kink-pairs opens up many perspectives for future studies of physics of dislocations. Another convincing example of LML and QNML transferability is presented in Secs. IV B and IV C.

B. Relative stability of dislocation loops and C15 clusters in Fe

The traditional mechanism of point defect clustering in bcc metals suggests that the elementary building bricks, like SIA dumbbells and vacancies, pack together in bundles, forming small 2D dislocation loops with well defined Burgers vector. These 2D clusters gradually grow and reach observable nanometric sizes. In bcc Fe, vacancies can bundle in 3D voids [119] that, at sufficient large size, collapse into dislocation loops with vacancy character. The works of Marinica *et al.* [70, 71, 83, 117] demonstrated that interstitials also can cluster into 3D objects with a specific crystallographic structure, which subsequently dissociate into 2D dislocation loops after reaching the critical size. The structure of such 3D clusters corresponds to the C15 Laves phase. Using DFT and empirical potentials, it was demonstrated [71, 83, 117] that in α -Fe these C15 aggregates have four important characteristics: (i) they are highly stable and (ii) they are immobile; (iii) they exhibit large antiferromagnetic moments; and (iv) they form directly under irradiation, in displacement cascades [118] or Frenkel pair accumulations [70] and they can grow by absorbing self-interstitials. Therefore, the correct characterization of the crossover between the 2D dislocation loops and 3D C15 clusters is of primary importance for understanding the microstructural evolution of iron-based materials under extreme conditions.

Investigations of the crossover between 2D and 3D defects require atomistic simulations with nanometric-sizes atomic systems. Such calculations can be performed using traditional empirical potentials, which are numerically fast but not always accurate and the crossover size depends on the choice of empirical potential. Recently, EAM [72] and bond order [112] potentials were developed to address this problem. These potentials, as well as older ones [71], provide a crossover between the C15 and $\frac{1}{2}\langle 111 \rangle$ clusters in the range of 20 - 35 SIAs. The computational cost of *ab initio* calculations prevents direct first principle investigations of C15 clusters and dislocation loops at nanometer sizes. To overcome this problem, Alexander *et al.* [71] developed a discrete-continuum (DC) model with *ab initio* accuracy to predict formation energies of defect clusters with various geometries and sizes. This combines the discrete nature of interstitial clusters and continuum elasticity for crystalline solid matrix. The model was applied to interstitial dislocation loops with $\frac{1}{2}\langle 111 \rangle$ Burgers vector, and to the C15 clusters in bcc Fe, W, and V to determine their relative stability as a function of defect size. In bcc Fe, C15 clusters in Fe smaller than 1.5 nm were found to be more stable than dislocation loops, i.e., the critical number of self-interstitial atoms in the 3D clusters is in the range of 40 - 49 SIAs. The accurate parametrization of such models is difficult and requires a considerable amount of computational resources. Alexander *et al.* [71] used large-scale DFT calculations of clusters with up to 20 SIAs embedded in a matrix with more than thousand bcc atoms.

Here, using our ML potentials for Fe, we perform large-scale atomistic calculations in order to determine the cross over between the formation energies of $\frac{1}{2}\langle 111 \rangle$ loops and C15 clusters. Inserting dislocation loops and C15 clusters from 1 to 300 SIAs in a simulation box of $50a_0 \times 50a_0 \times 50a_0$ (250 000 bulk atoms), we compute the formation energy of defect clusters. Atomic relaxation is performed until the maximum atomic force is lower than 10^{-3} eV/Å. Using this procedure, for the cluster sizes smaller than 140 SIAs, the error for formation energy is lower than 0.3 %. This error is estimated with respect to the calculations in much larger cells containing more than 10^6 atoms. The largest C15 cluster I_{101}^{C15} is computed using $80a_0 \times 80a_0 \times 80a_0$ (1 024 101 atoms) and $50a_0 \times 50a_0 \times 50a_0$ (250 101 atoms) boxes, which results in formation energies of 169.98 eV and 170.48 eV, respectively. In order to explore the energy landscape of C15 clusters, we generate the defect configurations using the *selection rules* described in [72]. This approach facilitates the construction of the lowest energy C15 clusters with minimal effort. The results are reported in Fig. 5. The present LML and QNML potentials for iron predict a crossover in the range of 35 - 45 SIAs. This range is in good agreement with the DC model [71] and provides a considerable improvement with respect to the traditional potentials.

Small differences in the crossover size range between LML/QNML and DC approach [71] can be explained by the differences in DFT pseudopotentials that were used for the design of DC and ML databases. The DC model is fitted using a database computed using a PAW Fe pseudopotential having semi-core states, while the database for fitting ML potentials in this work accounts for only 8 valence electrons without semi-core state. Moreover, the plane wave energy cutoff is set to 350 eV in Ref. [71], while in this study we use 500 eV. The MP scheme for k -point mesh of the Brillouin zone of the cubic unit cell is $16 \times 16 \times 16$ in Ref. [72] compared to $20 \times 20 \times 20$ for the present DFT calculations. These differences in the setup of DFT calculations impact the formation energies of defects.

It is worth to note that our ML database for Fe contains only two C15 cluster configurations: I_2^{C15} (one Z16 polyhedron) and I_4^{C15} (two joint Z16 polyhedra). More details about the construction of the C15 clusters can be found in Refs. [71, 72, 83]. With the employed PAW pseudopotential, we find the lowest energy configuration with the magnetic state having the spin of atoms centering the Z16 polyhedra anti-ferromagnetic state with respect to the bcc matrix with magnetization 2.2 and $2.4 \mu_B$ for I_2^{C15} and I_4^{C15} , respectively. In the calculations from Ref. [71], the central atom of the Z16 cage is ferromagnetic with respect to the matrix. This has an impact on the formation energies of those clusters: formation energies of I_2^{C15} and I_4^{C15} from [71] are 8.03 eV and 11.23 eV, respectively. In the present paper, the same C15 clusters have formation energies of 7.84 eV and 11.0 eV, respectively. Those relatively small differ-

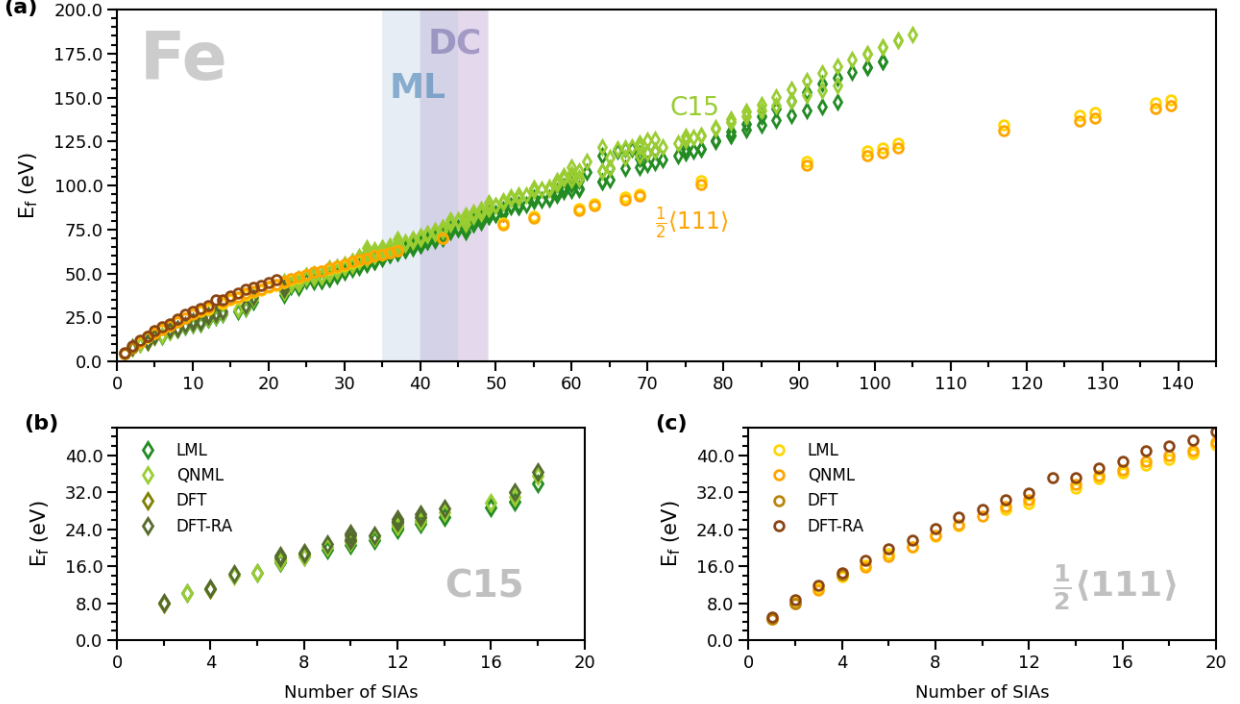


FIG. 5: Formation energies of C15 clusters and interstitial $\frac{1}{2}\langle 111 \rangle$ loops in bcc Fe from LML and QNML potentials and DFT calculations. (a) The relative stability of C15 clusters and $\frac{1}{2}\langle 111 \rangle$ loops. The shaded gray area between 35 and 45 SIA represents the crossover size predicted by our ML potentials, the purple shaded area between 40 and 49 SIA is the crossover predicted by the discrete-continuum model (DC) [71]. (b) Formation energies of C15 clusters. (c) Formations energies of $\frac{1}{2}\langle 111 \rangle$ loops. In all the subplots, the x axis correspond to the number of SIA; open circles and diamonds represent the energies of $\frac{1}{2}\langle 111 \rangle$ loops and C15 clusters, respectively. The color of points is explained in the legends in (b) and (c). The DFT-RA calculations are taken from Ref. [71] and are the calculations used in the parametrization of the DC model.

ences in DFT formation energies of small clusters become more important for larger clusters and exhibit a systematic trend. The comparison of the present LML/QNML results with the DFT values for the clusters smaller than 20 SIA are presented in Fig. 5b and c for dislocation loops and C15 clusters, respectively. A similar tendency of the present DFT formation energies being lower than those from [71] is also observed for the dislocation loops. Here, it is worth noting that our ML training database includes only three configurations of 2D clusters: $\frac{1}{2}\langle 111 \rangle$, $I_n^{(111)}$ with the number of SIA $n = 1, 2, 4$. The DFT values of the formation energy for the corresponding loops are 4.90, 8.63, and 14.47 eV in [71] and 4.64, 7.90, 14.07 eV in the present study.

Despite some differences in formation energies caused by differences in the reference DFT calculations, the two approaches, DC and present ML potentials, exhibit a remarkable agreement for this complex case of the energy landscape of C15 clusters. As such, the developed ML potentials represent a real solution for exploration of the energy landscape of C15 clusters and interstitial dislocation loops in Fe.

C. Anharmonic free energy calculations of bcc bulk and mono-vacancy in Fe and W

The thermal properties of defect populations in a material are driven by the underlying free energy landscape of defect formation and migration. Accurate characterization of the free energy landscape (e.g., with less than 1 meV error for small systems with few hundreds of atoms) requires considerable numerical effort, with at least 10^5 force evaluations. This makes direct simulations with *ab initio* methods very impractical. As such, ML potentials with numerical cost between the traditional potentials and DFT calculations open up many perspectives for such calculations. In the classical regime above one-third of the Debye temperature T_D , the evaluation of free energy using classical Boltzmann partition function can be performed with various methods for free energy calculations [137, 138].

Direct thermodynamic integration of the free energy variation ΔF between an initial state $\mathcal{U}_i(\mathbf{q})$ and a final state $\mathcal{U}_f(\mathbf{q})$, where $\mathbf{q} \in \mathbb{R}^{3N}$ represents a given atomic configuration, is numerically challenging for large N giv-

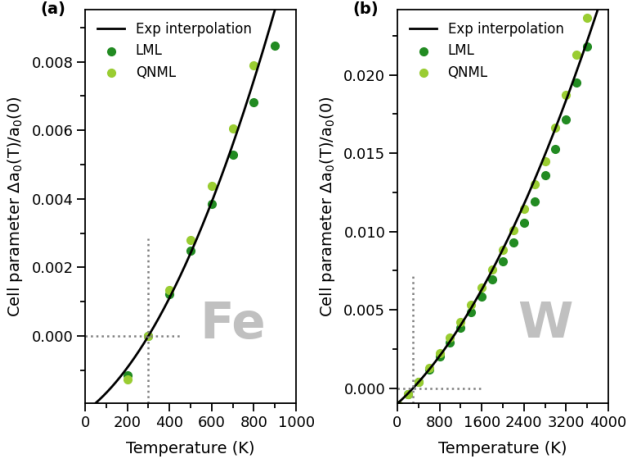


FIG. 6: The variation $\Delta a_0(T)$ of the lattice parameter $a_0(T)$ in bcc Fe and W at the temperature relative to the experimental value $a_0^{exp}(0)$ at $T = 0$ K [139]. The variation $\Delta a_0(T) = a_0^{ML}(T) - a_0^{exp}(300)$. We report the interpolation of the experimental values [139] (black lines) and the corresponding values of the present LML/QNML potentials (green circles). The numerical values for ML potentials are computed using the evaluation of the minimal free energy of the bcc systems using thermodynamic integration.

ing rise to a prohibitively large variance. Here we use alchemical integration [137, 140, 141] where the initial state is well known and can be easily evaluated, e.g., Einstein or harmonic crystal $\mathcal{U}_i(\mathbf{q}) = U_{HA}(\mathbf{q})$, and the final state is the real state of the system $\mathcal{U}_f(\mathbf{q}) = U(\mathbf{q})$. The two states, initial and final, are linearly coupled by introducing an alchemical reaction coordinate $\zeta \in [0, 1]$. The new system has the energy $U_s(\mathbf{q}, \zeta) = \zeta U(\mathbf{q}) + (1 - \zeta)U_{HA}(\mathbf{q})$, where ζ characterizes the free energy coupling between the system's harmonic state and real state. This coupling makes direct sampling tractable [142–144]. From the canonical partition function Z , one can deduce a probability of finding the system in a given state characterized by ζ , i.e., the system's free energy in the Landau sense, $A(\zeta) = -\beta^{-1} \ln Z(N, V, T, \zeta)$. The difference between the full free energy ($\zeta = 1$) and the reference free energy ($\zeta = 0$) becomes:

$$A(1) - A(0) = \int_0^1 \langle U(\mathbf{q}) - U_{ref}(\mathbf{q}) \rangle_{d\mu_\zeta} d\zeta, \quad (6)$$

where $A(0)$ is the free energy of the reference potential $U_{ref}(\mathbf{q})$ for $\zeta = 0$, and $A(1)$ gives the free energy of the system at $\zeta = 1$ with the potential energy given by $U(\mathbf{q})$. The average of the thermodynamic integration, on the energies difference, is made using the Boltzmann measure $d\mu_\zeta$ associated with the alchemical potential $U_s(\mathbf{q}, \zeta)$. Then, the difference in free energy is calculated as:

$$F(T, V) = F_{ref}(T, V) + \int_0^1 \langle U(\mathbf{q}) - U_{ref}(\mathbf{q}) \rangle_{d\mu_\zeta} d\zeta. \quad (7)$$

Having derived Eq. 6 and Eq. 7, the determination of free energy of a given system is subject to two requirements. Firstly, a reference system, for which free energy can be calculated analytically, should exist. Secondly, a reversible artificial pathway between the system of interest and a reference crystal should be provided to carry out a thermodynamic integration of the potential energy on the right-hand-side of Eq. 7 or Eq. 6. This method is also known as alchemical transition because a reaction coordinate or coupling parameter enables a smooth transition from a realistic potential energy of the system to a known reference state that is easy to compute. Thermodynamic integration can be performed using various methods, such as perturbation methods [145, 146], adaptive biasing potential [147], or adaptive biasing force [148–151] to determine the free energy difference. In this work we use adaptive biasing force method.

Here we present the evaluation of the free energy basin of the bcc bulk minima and the minima of mono-vacancies in bcc Fe and W. Both simulations are performed at zero pressure. Assuming zero-pressure condition implies that the thermodynamic functions like Gibbs free energy $G = E + PV - TS$ and enthalpy $H = E + PV$ are replaced by Helmholtz free energy $F = E - TS$ and energy E of the system, respectively. The Helmholtz free energy surface $F(T, V, N)$, a function of number of atoms N , volume V , and temperature T , is a central quantity and the intensive parameters, as pressure, become: $P(V, T) = -(\frac{\partial F}{\partial V})_T$. The equilibrium volume of the system at zero pressure is evaluated from the systems free energy around equilibrium state using the Birch-Murnaghan (BM) [152, 153] equation of state. BM interpolations with between 7 and 11 points are used. The maximum applied isostrain is less than 0.9%. The initial guess of the equilibrium volume is not obvious. Here, for each temperature, the guess is taken based on the experimental value of thermal expansion rescaled by 0 K ratio between DFT and experimental volume. Using the initial guess from BM interpolation, a new value of the equilibrium volume is computed. Then, several iterations are performed until convergence is reached, i.e., when the lattice parameter variation becomes less than 10^{-4} Å.

1. Thermal expansion

To compute the thermal expansion of bcc Fe and W, their lattice parameters are calculated at various temperatures. Figure 6 reports the computed thermal expansion of ML potentials in comparison with experimental values [139]. The reference cell parameter is taken at the experimental value [139] at 300 K. For temperatures lower than 300 K, the direct comparison between experimental and theoretical values of lattice parameters is not applicable. The present thermodynamic integration is based on classical mechanics framework within which the contribution of quantized phonons cannot be taken into account. Recently, in the context of physics

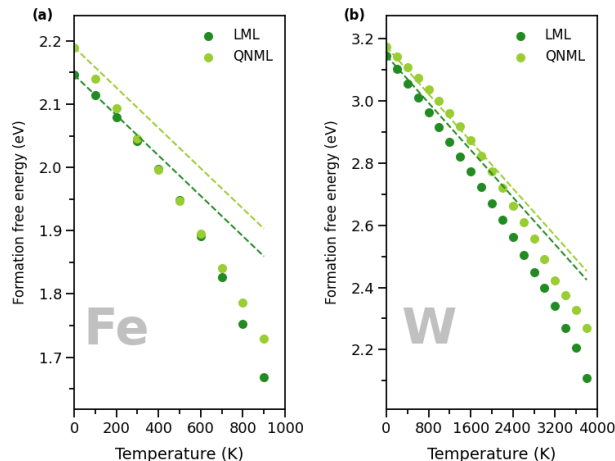


FIG. 7: The anharmonic free energy of mono-vacancy formation at zero pressure as a function of temperature computed using LML and QNML potentials. The dashed lines are the vibrational formation free energy based on the values of zero pressure vacancy formation entropy from self diffusion experiments [119] (see the text for more details).

of defects, it was demonstrated that those discrepancies are important below Debye temperature T_D [123, 124] even for metals with large atomic mass, like W. However, for temperatures where the classical regime is dominant, there is an excellent agreement between theoretical predictions and experiments. At high temperature (Fig. 6), the predictions of QNML potential are slightly more accurate than those the LML potential. As reported in Table I, the MAE of forces within QNML formalism are almost twice smaller compared to LML. The better accuracy of atomic forces results in better performance for free energy landscape calculations, as it was previously attested by the free energy calculations using the mean force [14, 133, 154]. Consequently, the vibrational free energy obtained with QNML is closer to DFT and experiments.

2. Free energy of mono-vacancy formation

The formation free energy of mono-vacancies is an important quantity for parametrization of multi-scale models like kinetic Monte Carlo, cluster dynamics, dislocation dynamics, etc. Its magnitude controls the density of vacancies, which acts as a driving force of self-diffusion [119], vacancy flux coupling [155–158], Ostwald ripening [159], climb mechanisms [160–162], phase transformations [160], etc. Until now, in bcc metals there are no studies that investigate the anharmonic contribution to the formation free energy of vacancies in Fe and W. The vast majority of multiscale models, cited above, use the values of vibrational entropy suggested by experience

or the values computed using *ab initio* techniques within the harmonic approximation. Here, using the developed ML potentials, we compute this quantity directly using thermodynamic integration. The process of forming a vacancy includes the creation of an internal free surface. The exchange-correlation functional plays here an important role: LDA tends to describe internal surfaces better than GGA. The reason is a well-known cancellation effect: LDA largely overestimates the exchange energy of a free metal surface, but underestimates by approximately the same magnitude the correlation energy. This results in a reasonable net total value of the surface energy. GGA results can vary depending on the parametrization of the exchange-correlation functional in use. For example, for Perdew-Burke Ernzerhof (PBE) the exchange surface energy is underestimated but the correlation surface energy is only slightly overestimated, resulting in too small vacancy formation energies. In the case of the phonon spectrum, it has been shown that in the case of the fcc metals the GGA functional gives the lower limit and the LDA functional the higher limit of spectrum [163–165]. Moreover, variants of PBE exchange-correlation functional, such as AM05 [166, 167], were designed in order to overcome this problem. However, for finite temperature properties of vacancies in fcc metals it is commonly stated that is better to use the standard GGA PBE functional [165]. For those reasons we inspect the free energy landscape of the mono-vacancies only with GGA PBE. Direct comparison of simulations with experimental data is not easy to perform. Self-diffusion experiments can provide only the activation enthalpy, the sum of migration and formation energy. Moreover, the change of magnetic structure in α -Fe (from ferromagnetic to paramagnetic) at almost $\frac{1}{2}T_m$ (1,043 K), and the large magnitude of vacancy migration barrier in W impede the acquisition of accurate activation enthalpies. The accepted experimental values [119, 168] of the formation entropy are $3.7 k_B$ and $3.2 k_B$ for Fe and W, respectively. In the case of W, we use the formation entropy of $2.2 k_B$ rather than $3.2 k_B$. As it was mentioned in Satta *et al.* [169], we subtract the electronic contribution (even though this contribution is averaged only at high temperatures).

Figure 7 reports the computed zero pressure formation free energy of mono-vacancy with full anharmonicity in thermal vibrations up to the melting temperature T_m . The experimental data are presented there as $F_{ML}(0) - TS_{exp}$. In the case of Fe, constant volume PBE-GGA DFT calculations within harmonic approximation (HA) predict larger values of the formation entropy: $4.1 k_B$ [170] or $4.9 k_B$ [171, 172]. Using the present ML potentials for the similar simulations, we find $5.9 k_B$ and $5.5 k_B$ with LML and QNML, respectively. These values are in good agreement with DFT predictions [171, 172], taking into account the differences in the DFT set up between the cited studies and the present DFT database of ML potentials. For W, constant volume DFT calculations with HA predict $2.6 k_B$, while the ML potentials

predict $4.1 k_B$ (LML) and $2.9 k_B$ (QNML). As in the case of Fe, QNML values in W are closer to DFT than LML predictions.

The results in Figure 7 emphasise the importance of anharmonicity for $T > \frac{1}{3}T_m$. Free energy of mono-vacancy formation in Fe is more impacted by anharmonicity in thermal vibrations. Most likely, it can be explained by the lower mass of Fe. The computed anharmonic free energies offer a significant improvement of theoretical predictions, which can be of great help for accurate multiscale simulations that imply characterization of high-temperatures regimes. Moreover, the precision of the present predictions encourages further studies to explore the impact of different exchange-correlation functional on finite-temperature properties.

V. CONCLUSIONS

In this work we have developed and tested ML interatomic potentials for accurate modeling of radiation defects and dislocations in bcc Fe and W. The generalization power and confidence limit of these potentials enable large-scale atomistic simulations that are not feasible with first principles methods. Beyond the standard linear ML potentials (LML) [35, 40], which assume a linear proportionality between the descriptors and local atomic energy, we have proposed a different version of ML potentials, called quadratic noise ML (QNML). In this approach the error of linear fit is treated as a quadratic form of atomic descriptors. This noise has a Gaussian distribution and, consequently, the fit is strongly predetermined by the linear solutions. Regression of the error noise for QNML potentials improves the accuracy of standard LML approach without significant loss of transferability and generalization power. Moreover, the QNML potentials increase the learning capacity with respect to LML and make the present ML approach suitable for active learning [20, 59, 173]. The QNML potentials are well adapted for atomic descriptors with reasonably small dimension (preferably $D < 100$). In this work we have demonstrated the performance of these ML potentials in conjunction with the bispectrum $SO(4)$ feature set.

The developed LML and QNML potentials reproduce most of the essential *ab initio* and experiment-based knowledge about the defects in Fe and W. The potentials accurately predict many difficult issues related to the stability and mobility of defects that are determined by electronic structure and, therefore, very challenging for semi-empirical potentials. Among such cases, the ML potentials correctly predict: (i) the triangle Gao configuration as the most stable configuration for di-interstitial in Fe; (ii) the magnitude and single saddle point shape of the migration barriers of mono-vacancy and of the Peierls barrier of the $\frac{1}{2}\langle 111 \rangle$ screw dislocation; (iii) the magnitude of the saddle point for V_3 migration; (iv) the binding energy of di-vacancies in W in accord with the DFT calculations.

In addition to modeling small defects with DFT accuracy, the ML potentials developed here are suitable for large-scale simulations with computational cost beyond the scope of *ab initio* methods. The potentials enable direct atomistic calculations of important properties such as formation and migration energy of large defects, e.g., kink-pairs, large 2D loops and 3D clusters that require large simulation cells containing $10^5 - 10^6$ atoms in order to eliminate finite size effects (e.g., elastic interaction of defects).

In this work we have provided two important examples of such large length scale calculations. The first example considers direct atomistic calculations of the relative stability of large interstitial dislocation loops and C15 clusters in Fe. The energy landscape of SIA in Fe is very complex. The revisited scenario for formation of 2D dislocation loops in Fe [70, 71, 83, 117] predicts that interstitial atoms first cluster into 3D objects with C15 crystallographic structure, which subsequently dissociate into dislocation loops after reaching the critical size. However, the accuracy of traditional empirical potentials is not sufficient to find this critical size [71, 72] and the numerical cost of *ab initio* methods does not allow for direct atomistic simulations. Here, using the ML force fields, we find the crossover between the formation energies of the two families of defects around 40 SIAs. This result is in very good agreement with a physics-informed discrete-continuum (DC) model [71]. Detailed analysis of the calculations emphasizes that the small difference in the order of 5 SIAs between the ML and DC predictions is determined by differences in the DFT databases used for the parametrization of the models (the pseudopotential and k -point grid sampling). It is interesting to note that the present ML potential was trained on small defect clusters with less than 5 SIAs and small simulation cells ($4a_0 \times 4a_0 \times 4a_0$), whilst the accurate fit of the DC model required a significant numerical effort to build the database with the clusters up to 20 SIAs and large DFT simulation cells (up to $8a_0 \times 8a_0 \times 8a_0$). Thus, the generalisation power of designed ML force fields is sufficient to ensure transferability and reconstruct the geometry and energy landscape of large defects. A similar conclusion is reached in the second example with direct atomistic calculations of the kink-pair formation energy of the $\frac{1}{2}\langle 111 \rangle$ screw dislocation. Kink-pairs were not directly included to the training database, but the presence of straight screw dislocations and Peierls barriers were enough to correctly predict kink-pair formation energies. The values predicted by the LML and QNML, 0.77 eV – 0.84 eV for Fe and 1.42 eV – 1.65 eV for W, are in excellent agreement with LT models [86, 88, 134] parametrized using DFT calculations. Based on these two large length scale physical problems, it can be concluded that the design of the database, the type of the exchange-correlation functional, and the setup of *ab initio* calculations are all crucial for the accuracy of ML potentials. Consequently, development of robust ML potentials requires not only an extensive training database but also reliable and co-

herent *ab initio* content.

Finally, we used the developed ML potentials to characterize the anharmonic free energy landscape of bulk and mono-vacancy formation in Fe and W. The accurate characterization of the free energy landscape (less than 1 meV error) requires a considerable numerical effort, e.g., 10^5 to 10^6 force evaluations for small systems containing few hundreds of atoms. Enabled by accuracy and numerical performance of LML and QNML potentials, we have computed the thermal expansion of the perfect bulk and zero-pressure formation free energy in bcc Fe and W. The results are in excellent agreement with the experimental data. The anharmonic free energy of mono-vacancy formation in Fe and W was computed and we demonstrate that the effect of anharmonicity is important above $\frac{2}{3}T_m$ in Fe and W and cannot be neglected. The obtained free energies can be used as a reliable input for multiscale simulations like kinetic Monte-Carlo, cluster dynamics, etc. These promising applications combined with various sampling strategies [14, 133] open up many perspectives for accurate characterization of defects at finite temperature in the materials science community.

The developed LML and QNML potentials for Fe and W together with the LAMMPS-MILADY module necessary for the reproduction of the results are available at GitHub repository [103].

ACKNOWLEDGMENTS

This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014-2018 and 2019-2020 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission. This work also received funding from the Euratom research and training programme 2019-2020 under grant agreement No. 755039. TDS and PG gratefully recognise support from the Agence Nationale de Recherche, via the MEMOPAS project ANR-19-CE46-0006-1. This work was granted access to the HPC resources of IDRIS under the allocations A0070910965 and A0090910965 attributed by GENCI. JRK and PG acknowledge support from the UK Engineering and Physical Sciences Research Council (EPSRC) under grant numbers EP/R012474/1 and EP/R043612/1. Additional support was provided by the Leverhulme Trust under grant RPG-2017-191. We are grateful for computational support from the UK national high performance computing service, ARCHER, for which access was obtained via the UKCP consortium and funded by EPSRC grant reference EP/P022065/1. Additional computing facilities were provided by the Scientific Computing Research Technology Platform of the University of Warwick. JD, AMG, CL and MCM acknowledge the support from GENCI - (CINES/CCRT) computer center under Grant No. A0090906973. MCM and LV acknowledge the support from GENCI-CINES

and GENCI-CCRT computer centres under Grant No. A0070906821. The authors thank Daniel R. Mason and Duc Nguyen-Manh for pointing the problem with modeling tri-vacancy migration in W relevant for the fusion community.

Appendix A: Content of DFT database

Content of ML database has a strong impact on the accuracy and transferability of the potential. Here we detail the the databases that were used for fitting of LML and QNLM potentials for bcc Fe and W. The database is organised in subsets, called DB classes in Table V. Different classes contain the simulations relevant for different properties, e.g., elasticity, formation and migration energy of the certain defects, thermal properties, etc. Aiming to model structural defects in bcc Fe and W, we include to the database point defects and their small clusters, including C15 interstitial clusters in bcc Fe; as well as extended defects like dislocations, free surfaces and γ -surfaces, also called generalized stacking faults (GSF). Defect-free structures of the database contain MD configurations of bcc bulk, elastically deformed systems and few thousands of highly deformed bcc systems. In addition to that, the databases contain MD configurations of liquid state, which ensures good statistical sampling of atomic environments, drastically different from those of bcc. We empirically observed that the presence of liquid in the database ensures a dense sampling of the descriptor space and improves the stability of ML potential for predictions of atomic environments that were not explicitly included into the database. Moreover, it was recently demonstrated in Ref. [22] that including liquid structures acts as regularization of the fitting solution, which improves the transferability and prediction power of ML potentials.

For most atomic systems, we compute energies (E), forces (F) and the virial stress (S). A system with N atoms provides at most $7+3N$ observables to fit: 1 energy (n_E), $3N$ force components (n_F) and 6 independent stress components (n_S). Fe potentials were trained on 111 683 energies, forces and stress components ($n_E + n_F + n_S$) from 761 configurations and tested on 41 496 data instances from 300 configurations. In case of W, training was performed on 131 057 data points from 2 132 atomic systems, and test was carried out on 22 543 instances from 607 systems. The number of instances for train and test per each DB class are given in Table V.

Figure 8 emphasizes the variety of atomic environments different from defect-free bcc in the potential database. The atomic environments different from bcc are selected using the distortion score of local atomic environments [14], which is based on outlier detection with Minimum Covariance Determinant (MCD) [174, 175]. In this approach, each atom is characterized by a distortion score that describes a statistical distance from a reference distribution in the descriptor space. Here reference

TABLE V: Details of the different classes used in the Fe and W databases. The “DB class” column gives a short description of the calculations, the “Atoms per cell” column lists the number of atoms in the cell used for DFT calculations, the “Properties to fit” column corresponds to the fitted properties for training/testing (energy E , force F , stress S), the fourth column corresponds to the total number of train/test data that are used to fit/test the potential n_E, n_F, n_S stands for number of energies, forces and stresses, respectively, included in the fit. Finally, the last column “Configurations train/test” gives the total number of configurations that were used for the train/test purposes.

Fe Database					
DB class	Content	Atoms per cell	Properties to fit	$n_E + n_F + n_S$ train/test	Configurations train/test
1	Bcc bulk	128	E	1/0	1/0
2	Elasticity (bcc)	2	ES	315/0	45/0
3	Deformations	2	ES	2 800/1 351	400/193
4	SIA $I_1 - I_4$	129-132	E	9/0	9/0
5	Vacancies $V_1 - V_4$	124-127	E	11/0	11/0
6	Free surfaces	30	E	6/0	6/0
7	MD - bcc bulk (300 K, 800 K)	128	$EF S$	7 820/2 737	20/7
8	MD - bcc bulk (a_0 rescaled 800 K)	128	$EF S$	3 519/1 173	9/3
9	MD - liquid (400 K, 800 K)	100	EF	18 060/6 020	60/20
10	MD - SIA $I_1 - I_4$ (800 K)	129-132	$EF S$	35 745/18 304	90/46
11	MD - vacancies $V_1 - V_4$ (800 K)	127-124	$EF S$	30 659/11 911	80/31
12	Migration of mono-vacancy V_1	127	E	7/0	7/0
13	Migration of mono-interstitial I_1	129	$EF S$	2 364/0	6/0
14	Migration of $\frac{1}{2}\langle 111 \rangle$ screw dislocation	135	$EF S$	3 708/0	9/0
Total				111 683/41 496	761/300
W Database					
DB class	Content	Atoms per cell	Properties to fit	$n_E + n_F + n_S$ train/test	Configurations train/test
1	Bcc bulk	128	E	1/0	1/0
2	Elasticity (bcc)	2	ES	525/0	75/0
3	Elasticity (fcc)	4	E	33/0	33/0
4	Deformations	1	ES	10 500/3 500	1 500/500
5	SIA $I_1 - I_4$	129-132	E	12/0	12/0
6	Vacancies $V_1 - V_4$	124-128	E	4/0	4/0
7	Vacancy clusters $V_8 - V_{16}$	240-248	EF	12 636/0	17/0
8	Free surfaces	12	E	45/15	45/15
9	γ -surfaces	12	E	100/22	100/22
10	MD - bcc bulk (300 K, 1 000 K, 3 000 K)	128	$EF S$	17 595/5 865	45/15
11	MD - liquid (1 000 K, 3 000 K)	100	EF	9 632/2 408	32/8
12	MD - vacancies $V_1 - V_2$ (1 000 K)	126-127	$EF S$	11 980/3 477	31/9
13	MD - vacancies $V_2 - V_3$ (300 K)	125-126	EF	6 330/1 886	16/5
14	MD - vacancies V_5 (300 K)	123	EF	4 440/1 110	12/3
15	MD - (110) γ -surface with vacancy V_1 (300 K)	47	EF	14 200/ 4 260	100/30
16	MD - $\frac{1}{2}\langle 111 \rangle$ screw dislocation (300 K, 1 000K)	135	EF	40 600/0	100/0
17	Migration of mono-vacancy V_1	53	EF	800/0	5/0
18	Migration of $\frac{1}{2}\langle 111 \rangle$ screw dislocation	135	EF	1 624/0	4/0
Total				131 057/22 543	2 132/607

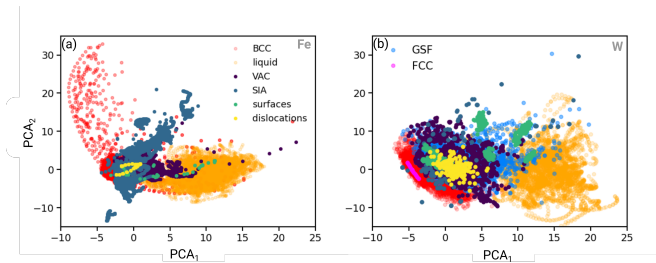


FIG. 8: Principal component analysis (PCA) representation of the databases for (a) bcc Fe and (b) bcc W. Each point on the plot corresponds to an atomic environment. The depicted atomic environments are outliers with respect to defect-free bcc bulk structures (see text), as provided by the distortion scores of LAEs [14]. Different colors represent different classes of structures in the database. The same color coding is used for Fe and W, with two additional classes, fcc and γ -surfaces, present in W.

distribution is be constructed from the local atomic environments of defect-free bcc systems (Table V, DB class 7 for Fe and 10 for W at 300K). Figure 8 depicts the atoms that were identified as outliers deviating from the bcc structure. The selected data is represented using principal components analysis (PCA) and visualized along the directions with maximum variance. The Fe and W databases (Fig. 8) have two major components given by the atoms belonging to the point defects embedded into bcc matrix and by the atoms of MD liquid at various temperatures. Between those two main clouds, there are defects that favor the physical transition between bcc and liquid phase such as vacancies. The database of W also includes GSF. In the Fe database, some highly deformed bcc configurations (few hundreds atoms among 2800 atomic environments) clearly deviate from the reference defect-free bulk.

-
- [1] M. S. Daw and M. I. Baskes, *Phys. Rev. Lett.* **50**, 1285 (1983).
- [2] M. S. Daw and M. I. Baskes, *Phys. Rev. B* **29**, 6443 (1984).
- [3] M. W. Finnis and J. E. Sinclair, *Philos. Mag. A* **50**, 45 (1984).
- [4] V. Rosato, M. Guillope, and B. Legrand, *Philos. Mag. A* **59**, 321 (1989).
- [5] J. Luo and B. Legrand, *Phys. Rev. B* **38**, 1728 (1988).
- [6] M. S. Daw, S. M. Foiles, and M. I. Baskes, *Mater. Sci. Reports* **9**, 251 (1993).
- [7] M. I. Baskes, *Phys. Rev. B* **46**, 2727 (1992).
- [8] D. G. Pettifor, *Phys. Rev. Lett.* **63**, 2480 (1989).
- [9] A. Van Duin, S. Dasgupta, F. Lorant, and W. Goddard, *J. Phys. Chem. A* **105**, 9396 (2001).
- [10] T.-R. Shan, B. D. Devine, T. W. Kemper, S. B. Sinnott, and S. R. Phillpot, *Phys. Rev. B* **81**, 125328 (2010).
- [11] J. Behler and M. Parrinello, *Phys. Rev. Lett.* **98**, 146401 (2007).
- [12] M. W. Mahoney and P. Drineas, *Proc. Natl. Acad. Sci. U. S. A.* **106**, 697 (2009).
- [13] A. P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, *Phys. Rev. X* **8**, 041048 (2018).
- [14] A. M. Goryaeva, C. Lapointe, C. Dai, J. Dérès, J.-B. Maillet, and M.-C. Marinica, *Nat. Commun.* **11**, 4691 (2020).
- [15] A. P. Bartók, R. Kondor, and G. Csányi, *Phys. Rev. B* **87**, 184115 (2013).
- [16] H. Huo and M. Rupp, [arXiv:1704.06439 \[cond-mat, physics:physics\]](https://arxiv.org/abs/1704.06439) (2018), arXiv: 1704.06439.
- [17] B. Onat, C. Ortner, and J. R. Kermode, *J. Chem. Phys.* **153**, 144106 (2020).
- [18] A. P. Bartók, *Gaussian Approximation Potential: an interatomic potential derived from first principles Quantum Mechanics*, Ph.D. thesis, University of Cambridge (2009).
- [19] A. Shapeev, *Multiscale Model. Sim.* **14**, 1153 (2016).
- [20] E. V. Podryabinkin and A. V. Shapeev, *Comput. Mater. Sci.* **140**, 171 (2017).
- [21] A. E. A. Allen, G. Dusson, C. Ortner, and G. Csányi, *Mach. Learn.: Sci. Technol.* **2**, 025017 (2021).
- [22] C. van der Oord, G. Dusson, G. Csányi, and C. Ortner, *Mach. Learn.: Sci. Technol.* **1**, 015004 (2020).
- [23] M. Eickenberg, G. Exarchakis, M. Hirn, and S. Mallat, in *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017) p. 65406549.
- [24] M. Hirn, S. Mallat, and N. Poilvert, *Multiscale Model. Simul.* **15**, 827863 (2017).
- [25] F. Bruneval, I. Maliyov, C. Lapointe, and M.-C. Marinica, *J. Chem. Theory Comput.* **16**, 4399 (2020).
- [26] B. C. Yeo, D. Kim, C. Kim, and S. S. Han, *Sci. Rep.* **9**, 5879 (2019).
- [27] G. Ferré, J.-B. Maillet, and G. Stoltz, *J. Chem. Phys.* **143**, 104114 (2015).
- [28] G. Ferré, T. Haut, and K. Barros, *J. Chem. Phys.* **146**, 114107 (2017).
- [29] R. Drautz, *Phys. Rev. B* **99**, 014104 (2019).
- [30] R. Drautz, *Phys. Rev. B* **102**, 024104 (2020).
- [31] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, *Nat. Commun.* **8**, 1 (2017).
- [32] K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer, *Nat. Commun.* **10**, 5024 (2019).
- [33] F. Noé and C. Clementi, *J. Chem. Theory Comp.* **11**, 5002 (2015).
- [34] F. Noé, S. Olsson, J. Köhler, and H. Wu, *Science* **365**, eaaw1147 (2019).
- [35] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker, *J. Comp. Phys.* **285**, 316 (2015).
- [36] M. A. Wood and A. P. Thompson, [arXiv:1702.07042v1 \[physics.comp-ph\]](https://arxiv.org/abs/1702.07042v1) (2017).
- [37] A. Seko, A. Takahashi, and I. Tanaka, *Phys. Rev. B* **90**, 024101 (2014).
- [38] A. Seko, A. Takahashi, and I. Tanaka, *Phys. Rev. B* **92**, 054113 (2015).
- [39] C. Chen, Z. Deng, R. Tran, H. Tang, I.-H. Chu, and S. P. Ong, *Phys. Rev. Materials* **1**, 043603 (2017).
- [40] A. M. Goryaeva, J.-B. Maillet, and M.-C. Marinica,

- Comp. Mater. Sci.* **166**, 200 (2019).
- [41] H. Yanxon, D. Zagaceta, B. C. Wood, and Q. Zhu, *J. Chem. Phys.* **153**, 054118 (2020).
- [42] J. Behler, *J. Chem. Phys.* **134**, 074106 (2011).
- [43] N. Artrith and A. Urban, *Comp. Mater. Sci.* **114**, 135 (2016).
- [44] J. Behler, *J. Chem. Phys.* **145**, 170901 (2016).
- [45] M. A. Wood and A. P. Thompson, *J. Chem. Phys.* **148** (2018).
- [46] A. P. Bartók and G. Csányi, *Int. J. Quantum Chem.* **115**, 1051 (2015).
- [47] T. Hofmann, B. Schlkopf, and A. J. Smola, *Ann. Statist.* **114**, 1171 (2008).
- [48] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, *J. Phys. Chem. C* **121**, 511 (2017).
- [49] V. Botu and R. Ramprasad, *Int. J. Quantum Chem.* **115**, 1074 (2015).
- [50] V. Botu and R. Ramprasad, *Phys. Rev. B* **92**, 094306 (2015).
- [51] Z. Li, J. R. Kermode, and A. De Vita, *Phys. Rev. Lett.* **114**, 096405 (2015).
- [52] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, *Sci. Adv.* **3**, e1701816 (2017).
- [53] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, *Phys. Rev. Lett.* **114**, 105503 (2015).
- [54] C. C. Aggarwal, *Neural Networks and Deep Learning* (Springer, Berlin, Heidelberg, 2018).
- [55] C. E. Rasmussen, *Gaussian Processes in Machine Learning* (Springer, Berlin, Heidelberg, 2004).
- [56] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, *Chem. Rev.* **121**, 1014210186 (2021).
- [57] B. Grabowski, Y. Ikeda, P. Srinivasan, F. Krmann, C. Freysoldt, A. I. Duff, A. Shapeev, and J. Neugebauer, *npj Comput. Mater.* **5**, 80 (2019).
- [58] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, *Phys. Rev. Lett.* **104**, 136403 (2010).
- [59] N. Bernstein, G. Csányi, and V. L. Deringer, *npj Comput. Mater.* **5**, 99 (2019).
- [60] C. Lapointe, T. D. Swinburne, L. Thiry, S. Mallat, L. Proville, C. S. Becquart, and M.-C. Marinica, *Phys. Rev. Materials* **4**, 063802 (2020).
- [61] L. Huber, R. Hadian, B. Grabowski, and J. Neugebauer, *npj Comput. Mater.* **4**, 64 (2018).
- [62] T. A. Sharp, S. L. Thomas, E. D. Cubuk, S. S. Schoenholz, D. J. Srolovitz, and A. J. Liu, *Proc. Natl. Acad. Sci. U. S. A.* **115**, 10943 (2018).
- [63] T. Tamura, M. Karasuyama, R. Kobayashi, R. Arakawa, Y. Shihara, and I. Takeuchi, *Model. Simul. Mater. Sci. Eng.* **25**, 075003 (2017).
- [64] Y. Cheng, L. Zhu, G. Wang, J. Zhou, S. R. Elliott, and Z. Sun, *Comp. Mater. Sci.* **183**, 109803 (2020).
- [65] J. Marian, B. D. Wirth, and J. M. Perlado, *Phys. Rev. Lett.* **88**, 255507 (2002).
- [66] S. L. Dudarev and P. M. Derlet, *J. Phys. Condens. Mat.* **17**, 7097 (2005).
- [67] S. L. Dudarev, R. Bullough, and P. M. Derlet, *Phys. Rev. Lett.* **100**, 135503 (2008).
- [68] D. Terentyev, L. Malerba, P. Klaver, and P. Olsson, *J. Nucl. Mater.* **382**, 126 (2008), proceedings of the Symposium on Microstructural Processes in Irradiated Materials.
- [69] C.-C. Fu, J. D. Torre, F. Willaime, J.-L. Bocquet, and A. Barbu, *Nat. Mater.* **4**, 68 (2005).
- [70] A. Chartier and M. C. Marinica, *Acta Mater.* **180**, 141 (2019).
- [71] R. Alexander, M.-C. Marinica, L. Proville, F. Willaime, K. Arakawa, M. R. Gilbert, and S. L. Dudarev, *Phys. Rev. B* **94**, 024103 (2016).
- [72] R. Alexander, L. Proville, C. S. Becquart, A. M. Goryeava, J. Deres, C. Lapointe, and M.-C. Marinica, *J. Nucl. Mater.* **535**, 152141 (2020).
- [73] P. M. Derlet and S. L. Dudarev, *Phys. Rev. Materials* **4**, 023605 (2020).
- [74] C. S. Becquart and C. Domain, *Nucl. Instrum. Meth. B* **255**, 23 (2007).
- [75] C. S. Becquart, C. Domain, U. Sarkar, A. DeBacker, and M. Hou, *J. Nucl. Mater.* **403**, 75 (2010).
- [76] C. S. Becquart and C. Domain, *Curr. Opin. Solid State Mater. Sci.* **16**, 115 (2012).
- [77] L. Ventelon, F. Willaime, C.-C. Fu, M. Heran, and I. Ginoux, *J. Nucl. Mater.* **425**, 16 (2012).
- [78] M. Muzyk, D. Nguyen-Manh, K. J. Kurzydłowski, N. L. Baluc, and S. L. Dudarev, *Phys. Rev. B* **84**, 104115 (2011).
- [79] D. R. Mason, D. Nguyen-Manh, and C. S. Becquart, *J. Phys. Condens. Matter* **29**, 505501 (2017).
- [80] K. Heinola, F. Djurabekova, and T. Ahlgren, *Nucl. Fusion* **58**, 026004 (2017).
- [81] L. Malerba, M.-C. Marinica, N. Anento, C. Björkas, H. Nguyen, C. Domain, F. Djurabekova, P. Olsson, K. Nordlund, A. Serra, D. Terentyev, F. Willaime, and C. S. Becquart, *J. Nucl. Mater.* **406**, 19 (2010).
- [82] M.-C. Marinica, F. Willaime, and N. Mousseau, *Phys. Rev. B* **83**, 094119 (2011).
- [83] M.-C. Marinica, F. Willaime, and J.-P. Crocombette, *Phys. Rev. Lett.* **108**, 025501 (2012).
- [84] C. Domain and G. Monnet, *Phys. Rev. Lett.* **95**, 215506 (2005).
- [85] L. Ventelon and F. Willaime, *J. Comp.-Aided Mater. Des.* **14**, 85 (2007).
- [86] L. Dezerald, D. Rodney, E. Clouet, L. Ventelon, and F. Willaime, *Nat. Commun.* **7**, 11695 (2016).
- [87] L. Dezerald, L. Proville, L. Ventelon, F. Willaime, and D. Rodney, *Phys. Rev. B* **91**, 094105 (2015).
- [88] L. Proville, L. Ventelon, and D. Rodney, *Phys. Rev. B* **87**, 144106 (2013).
- [89] G. Kresse and J. Furthmüller, *Phys. Rev. B* **54**, 11169 (1996).
- [90] G. Henkelman, G. Jóhannesson, and H. Jónsson, “Methods for finding saddle points and minimum energy paths,” in *Theoretical Methods in Condensed Phase Chemistry*, edited by S. D. Schwartz (Springer Netherlands, Dordrecht, 2002) pp. 269–302.
- [91] G. Henkelman, B. P. Uberuaga, and H. Jónsson, *J. Chem. Phys.* **113**, 9901 (2000).
- [92] G. Henkelman, *Annual Review of Materials Research* **47**, 199 (2017).
- [93] D. Dragoni, T. D. Daff, G. Csányi, and N. Marzari, *Phys. Rev. Materials* **2**, 013808 (2018).
- [94] W. J. Szlachta, A. P. Bartók, and G. Csányi, *Phys. Rev. B* **90**, 104108 (2014).
- [95] J. Friedel, *Adv. Phys.* **3**, 446 (1954).
- [96] F. Cyrot-Lackmann, *Analyse de la structure électronique des systèmes désordonnés dans l’approximation des liaisons fortes*, Ph.D. thesis,

- Université Paris-Sud (1968).
- [97] F. Ducastelle and F. Cyrot-Lackmann, *J. Phys. Chem. Solids* **31**, 1295 (1970).
- [98] C. Freysoldt, B. Grabowski, T. Hickel, J. Neugebauer, G. Kresse, A. Janotti, and C. G. Van de Walle, *Rev. Mod. Phys.* **86**, 253 (2014).
- [99] M. Benoit, J. Amodeo, S. Combettes, I. Khaled, A. Roux, and J. Lam, *Mach. Learn.: Sci. Technol.* **2**, 025003 (2021).
- [100] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University, Cambridge, 2005).
- [101] G. H. Golub and C. F. van Loan, *Matrix Computation, Fourth Edition* (The Johns Hopkins University Press, 2013).
- [102] S. Plimpton, *J. Comp. Phys.* **117**, 1 (1995).
- [103] “LAMMPS-MILADY,” (2020).
- [104] J. N. Mundy, *Philos. Mag. A* **46**, 345 (1982).
- [105] J. Y. Park, H. C. W. Huang, R. W. Siegel, and R. W. Balluffi, *Philos. Mag. A* **48**, 397 (1983).
- [106] J. Byggmästar, A. Hamedani, K. Nordlund, and F. Djurabekova, *Phys. Rev. B* **100**, 144105 (2019).
- [107] S. Takaki, J. Fuss, H. Kuglers, U. Dedek, and H. Schultz, *Radiat. Eff.* **79**, 87 (1983).
- [108] C. Domain and C. S. Becquart, *Phys. Rev. B* **65**, 024103 (2001).
- [109] C.-C. Fu, F. Willaime, and P. Ordejón, *Phys. Rev. Lett.* **92**, 175503 (2004).
- [110] F. Willaime, C. C. Fu, M.-C. Marinica, and J. D. Torre, *Nucl. Instrum. Meth. B* **228**, 92 (2005).
- [111] M.-C. Marinica and F. Willaime, *Solid State Phen.* **129**, 67 (2007).
- [112] J. Byggmästar and F. Granberg, *J. Nucl. Mater.* **528**, 151893 (2020).
- [113] M.-C. Marinica, L. Ventelon, M. R. Gilbert, L. Proville, S. L. Dudarev, J. Marian, G. Bencteux, and F. Willaime, *J. Phys.: Cond. Matter.* **25**, 395502 (2013).
- [114] S. L. Dudarev, *Comptes Rendus Physique* **9**, 409 (2008).
- [115] G. J. Ackland, M. I. Mendelev, D. J. Srolovitz, S. Han, and A. V. Barashev, *J. Phys.: Condens. Matter* **16**, S2629 (2004).
- [116] F. Gao, D. J. Bacon, Y. N. Osetsky, P. E. J. Flewitt, and T. A. Lewis, *J. Nucl. Mater.* **276**, 213 (2000).
- [117] L. Dézerald, M.-C. Marinica, L. Ventelon, D. Rodney, and F. Willaime, *J. Nucl. Mater.* **449**, 219 (2014).
- [118] J. Byggmästar, F. Granberg, A. E. Sand, A. Pirttikoski, R. Alexander, M.-C. Marinica, and K. Nordlund, *J. Phys.: Condens. Matter* **31**, 245402 (2019).
- [119] P. Ehrhart, P. Jung, H. Schultz, and H. Ullmaier, *Atomic Defects in Metals* (Springer-Verlag, Berlin, 1991).
- [120] P. M. Derlet, D. Nguyen-Manh, and S. L. Dudarev, *Phys. Rev. B* **76**, 054107 (2007).
- [121] S. Fitzgerald and D. Nguyen-Manh, *Phys. Rev. Lett.* **101**, 115504 (2008).
- [122] P.-W. Ma and S. L. Dudarev, *Phys. Rev. Materials* **3**, 043606 (2019).
- [123] T. D. Swinburne, P.-W. Ma, and S. L. Dudarev, *New J. Phys.* **19**, 073024 (2017).
- [124] 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, T. Amino, S. Arai, Y. Yamamoto, K. Higuchi, N. Tanaka, H. Yasuda, T. Yasuda, and H. Mori, *Nat. Mater.* **19**, 508 (2020).
- [125] R. A. Johnson, *Phys. Rev.* **134**, A1329 (1964).
- [126] L. Proville, D. Rodney, and M.-C. Marinica, *Nat. Mater.* **11**, 845 (2012).
- [127] L. Romaner, C. Ambrosch-Draxl, and R. Pippan, *Phys. Rev. Lett.* **104**, 195503 (2010).
- [128] H. Li, S. Wurster, C. Motz, L. Romaner, C. Ambrosch-Draxl, and R. Pippan, *Acta Mater.* **60**, 748 (2012).
- [129] D. Cereceda, M. Diehl, F. Roters, D. Raabe, J. M. Perlado, and J. Marian, *Int. J. Plast.* **78**, 242 (2016).
- [130] D. Brunner and J. Diehl, *Phys. Status Solidi A* **124**, 455 (1991).
- [131] D. Brunner and J. Diehl, *Phys. Status Solidi A* **124**, 155 (1991).
- [132] D. Brunner, *Mater. Trans.* **41**, 152 (2000).
- [133] T. D. Swinburne and M.-C. Marinica, *Phys. Rev. Lett.* **120**, 135503 (2018).
- [134] M. Itakura, H. Kaburaki, and M. Yamaguchi, *Acta Mater.* **60**, 3698 (2012).
- [135] F. Maresca, D. Dragoni, G. Csányi, N. Marzari, and W. A. Curtin, *npj Comput. Mater.* **4**, 69 (2018).
- [136] J. D. Eshelby and G. V. Raynor, *P. Roy. Soc. Lond. A Mat.* **266**, 222 (1962).
- [137] T. Lelièvre, G. Stoltz, and M. Rousset, *Free energy computations: A mathematical perspective* (Imperial College Press, London, 2010).
- [138] C. Chipot and A. Pohorille, *Free energy calculations* (Springer-Verlag Berlin Heidelberg, 2007).
- [139] Y. S. Touloukian and C. Y. Ho, *Thermal expansion, metallic elements and alloys. Thermophysical properties of matter: the TPRC data series; a comprehensive compilation of data, vol. 12.* (IFI/Plenum, New York, 1970).
- [140] D. Frenkel and A. J. C. Ladd, *J. Chem. Phys.* **81**, 3188 (1984).
- [141] D. Frenkel and B. Smit, *Understanding molecular simulations: from algorithms to applications* (Academic Press, San Diego, 2002).
- [142] T. Lelièvre, M. Rousset, and G. Stoltz, *Nonlinearity* **21**, 1155 (2008).
- [143] G. M. Torrie and J. P. Valleau, *J. Comp. Phys.* **23**, 187 (1977).
- [144] A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.* **99**, 12562 (2002).
- [145] L. D. Landau, *Statistical Physics* (The Clarendon Press, Oxford, 1996).
- [146] R. W. Zwanzig, *J. Chem. Phys.* **22**, 1420 (1954).
- [147] B. M. Dickson, F. Legoll, T. Lelièvre, G. Stoltz, and P. Fleurat-Lessard, *J. Phys. Chem. B* **114**, 5823 (2010).
- [148] E. Darve and A. Pohorille, *J. Chem. Phys.* **115**, 9169 (2001).
- [149] E. Darve, M. A. Wilson, and A. Pohorille, *Mol. Simul.* **28**, 113 (2002).
- [150] E. Darve, D. Rodríguez-Gómez, and A. Pohorille, *J. Chem. Phys.* **128** (2008).
- [151] L. Cao, G. Stoltz, T. Lelièvre, M.-C. Marinica, and M. Athènes, *J. Chem. Phys.* **140**, 104108 (2014).
- [152] F. Birch, *Phys. Rev.* **71**, 809 (1947).
- [153] F. D. Murnaghan, *Proc. Natl. Acad. Sci. USA* **30**, 244 (1944).
- [154] T. D. Swinburne and D. Perez, *Phys. Rev. Materials* **2**, 053802 (2018).
- [155] T. Schuler and M. Nastar, *Phys. Rev. B* **93**, 224101 (2016).
- [156] T. Schuler, M. Nastar, and F. Soisson, *Phys. Rev. B*

- [95, 014113 \(2017\)](#).
- [157] T. Schuler, L. Messina, and M. Nastar, *Comp. Mater. Sci.* **172**, 109191 (2020).
- [158] M. Nastar, L. T. Belkacemi, E. Meslin, and M. Loyer-Prost, *Comm. Mater.* **2**, 1 (2021).
- [159] S. Moll, T. Jourdan, and H. Lefaix-Jeuland, *Phys. Rev. Lett.* **111** (2013).
- [160] H. Mehrer, *Diffusion in Solids: Fundamentals, Methods, Materials, Diffusion-Controlled Processes*, Springer Series in Solid-State Sciences (Springer, Berlin Heidelberg, 2007).
- [161] T. D. Swinburne, K. Arakawa, H. Mori, H. Yasuda, M. Isshiki, K. Mimura, M. Uchikoshi, and S. L. Dudarev, *Sci. Rep.* **6**, 30596 (2016).
- [162] M. Landeiro Dos Reis, L. Proville, M.-C. Marinica, and M. Sauzay, *Phys. Rev. Materials* **4**, 103603 (2020).
- [163] B. Grabowski, T. Hickel, and J. Neugebauer, *Phys. Status Solidi B* **248**, 1295 (2011).
- [164] A. Glensk, B. Grabowski, T. Hickel, and J. Neugebauer, *Phys. Rev. X* **4** (2014).
- [165] A. Glensk, B. Grabowski, T. Hickel, and J. Neugebauer, *Phys. Rev. Lett.* **114**, 195901 (2015).
- [166] R. Armiento and A. E. Mattsson, *Phys. Rev. B* **72**, 085108 (2005).
- [167] A. E. Mattsson, R. Armiento, J. Paier, G. Kresse, J. M. Wills, and T. R. Mattsson, *J. Chem. Phys.* **128**, 084714 (2008).
- [168] O. Seydel, G. Frohberg, and H. Wever, *Phys. Status Solidi A* **144**, 69 (1994).
- [169] A. Satta, F. Willaime, and S. de Gironcoli, *Phys. Rev. B* **57**, 11184 (1998).
- [170] G. Lucas and R. Schublin, *Nucl. Instrum. Methods Phys. Res., B* **267**, 3009 (2009).
- [171] D. Murali, M. Posselt, and M. Schiwarth, *Phys. Rev. B* **92**, 064103 (2015).
- [172] N. Sandberg, Z. Chang, L. Messina, P. Olsson, and P. Korzhavyi, *Phys. Rev. B* **92**, 184102 (2015).
- [173] M. Hodapp and A. Shapeev, *Mach. Learn.: Sci. Technol.* **1**, 045005 (2020).
- [174] M. Hubert and M. Debruyne, *WIRES Comp. Stat.* **2**, 36 (2010).
- [175] M. Hubert, M. Debruyne, and P. J. Rousseeuw, *WIRES Comp. Stat.* **10**, e1421 (2018).