

Achieving DFT accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron

Daniele Dragoni,^{1,2} Thomas D. Daff,³ Gábor Csányi,³ and Nicola Marzari¹

¹*Theory and Simulation of Materials (THEOS), and National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*

²*Dipartimento di Scienza dei Materiali, Università di Milano-Bicocca, Via R. Cozzi 55, I-20125 Milano, Italy*

³*Engineering Laboratory, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, United Kingdom*



(Received 23 August 2017; published 30 January 2018)

We show that the Gaussian Approximation Potential (GAP) machine-learning framework can describe complex magnetic potential energy surfaces, taking ferromagnetic iron as a paradigmatic challenging case. The training database includes total energies, forces, and stresses obtained from density-functional theory in the generalized-gradient approximation, and comprises approximately 150,000 local atomic environments, ranging from pristine and defected bulk configurations to surfaces and generalized stacking faults with different crystallographic orientations. We find the structural, vibrational, and thermodynamic properties of the GAP model to be in excellent agreement with those obtained directly from first-principles electronic-structure calculations. There is good transferability to quantities, such as Peierls energy barriers, which are determined to a large extent by atomic configurations that were not part of the training set. We observe the benefit and the need of using highly converged electronic-structure calculations to sample a target potential energy surface. The end result is a systematically improvable potential that can achieve the same accuracy of density-functional theory calculations, but at a fraction of the computational cost.

DOI: [10.1103/PhysRevMaterials.2.013808](https://doi.org/10.1103/PhysRevMaterials.2.013808)

I. INTRODUCTION

Iron is the most abundant element at the Earth's core, it is responsible for the generation of the geomagnetic field, it is the main component of the most widely used structural engineering material (steel), and in its atomic form is a component of, e.g., oxygen-binding proteins. In its crystalline form, it can host impurities that improve its mechanical properties and make it a formidably strong material suitable for many applications in the fields of construction, automotive, machinery, and energy production. It is a metal with partially filled d electronic bands and has a complex phase diagram which presents transformations driven by the interplay of magnetic, electronic, and vibrational degrees of freedom. As a consequence, the modeling of iron is highly nontrivial. Density-functional theory (DFT) provides a relatively good description of its zero-temperature properties [1–6] although, even in this regime, theory shows discrepancies with respect to experimental data [7]. The finite-temperature behavior of thermodynamical quantities of the bulk crystal can be well described up to a good fraction of the Curie point considering only vibrational effects [7,8]. Nonetheless, as temperature approaches and crosses the Curie point, magnetic fluctuations and magnetic disorder become crucial for a correct description of such thermodynamic properties [9–11]. In fact, despite the progress achieved in the past years, a satisfactory description of the thermodynamic phase transitions and of the paramagnetic phases of iron from first principles remains a formidable task. Even more complex is the study of iron alloys and steels that, on top of the challenges mentioned above, requires in many cases the capability to deal with length and time scales which are beyond the reach of any *ab initio* technique.

For this reason, empirical interatomic potentials have been developed, fitted typically to a mixture of experimental and *ab initio* data, that are capable of simulating systems containing thousands or millions of atoms for thousands or millions of time steps. These models allowed a detailed study of the microscopic processes at the origin of macroscopic mechanical properties of iron and iron alloys under different conditions. Embedded Atom Models (EAM) [12], and other similar approaches such as the Finnis-Sinclair model [13], local volume potentials [14], and the glue model [15] have proved to be successful. In particular, the Mendeleev family of parametrizations [16–18] are able to reproduce many fundamental properties of elemental bcc iron at zero temperature. These models, however, are not always fully satisfactory in reproducing the energetics of defective configurations such as self-interstitials [19] and divacancies [18], the Peierls potentials associated with screw dislocations [18,20] or even fundamental bulk properties at finite temperature within the range of stability of the ferromagnetic α phase [21]. Additionally, due to their fixed functional form, these potentials are not easily generalizable to the modeling of bonds with mixed metallic and covalent character as can be found, for example, in Fe-C alloys. More recently, new approaches such as the modified EAM [22], the (analytic) bond order potentials [23–25], magnetic EAM [26], or metallic-covalent interatomic potentials [27] have been developed in order to overcome some of these limitations.

In this work, we follow an alternative approach, generating a Gaussian approximation potential [28] (GAP) for the α phase of iron. GAP is a highly flexible machine-learning model that allows to fit directly and accurately first-principles potential energy surfaces (PES). Transferability is ensured by

regular and smooth basis functions (kernels, in the language of machine learning), and by an extended training database which covers here roughly 150 000 local atomic environments (LAEs). Similar machine-learning approaches, such as neural networks, have been successful recently in modeling materials where previous, more empirical strategies have run out of steam [29–38]. GAP uses Gaussian process regression [39,40], whose advantages are that (i) its hyperparameters (that control the kernel function and linear algebra regularization) make physical sense and rarely need adjusting, (ii) the fit itself is determined by simple linear algebra, rather than iterative nonlinear optimization of a highly multimodal function as in the case of neural networks, and (iii) input data such as energies, forces, and stresses are treated in a consistent manner, with appropriate error estimates that allow the inclusion of variable accuracy data. In the machine-learning literature, Gaussian process regression is often thought of as scaling poorly (cubically) with the size of the input data, but we find that well-known heuristics allow us to limit the number of basis functions to be much smaller than the number of input configurations, leading to training times of about a day on a single multicore server and to prediction costs similar to that of neural network-based potentials. The key to the success of Gaussian process regression is an appropriate kernel function that captures the symmetries and describes the spatial correlation structure of the target function. We use the “smooth overlap of atomic positions” (SOAP) kernel [41] that has been shown previously to lead to excellent results for other materials [42–45].

II. METHOD

We start by assuming that the Born-Oppenheimer potential energy surface of a set of atoms is a smooth function of the atomic coordinates. As it is usually done when constructing interatomic potentials, we write the total energy as a sum of atomic contributions

$$E = \sum_i \epsilon(\mathbf{q}_i), \quad (1)$$

where the short-ranged local atomic energy $\epsilon(\mathbf{q}_i)$ is assumed to depend explicitly on the positions of the atoms within a sphere of radius r_{cut} centered on atom i . The list of such atomic positions defines the local atomic environment of atom i and is represented by a suitable set of descriptors, here denoted by the vector \mathbf{q}_i . (Standard terms representing electrostatic and van der Waals interactions can be added as needed to account for long-range interactions.) Empirical interatomic potentials are designed using functional forms derived from physical intuition to approximate $\epsilon(\mathbf{q}_i)$, and parameters are fitted to experimental or computational data. The moderate flexibility of these functional forms limits their scope to be systematically improved by increasing the fitting datasets; on the other hand, their qualitative description of the essential physical interactions ensures a good degree of transferability. In the GAP framework, Gaussian process regression is used instead to define a model for the local atomic energy function ϵ as a linear combination of nonlinear kernel functions

$$\epsilon(\mathbf{q}^*) = \sum_s \alpha_s K(\mathbf{q}_s, \mathbf{q}^*) \equiv \mathbf{K}(\mathbf{q}^*)^T \boldsymbol{\alpha}, \quad (2)$$

TABLE I. Hyperparameters for the SOAP kernel and the GAP model.

Atomic environment kernel	SOAP
r_{cut}	5.0 Å
r_{Δ}	1.0 Å
σ_v^{energy} DB1	1.0×10^{-4} eV/atom
σ_v^{energy} DB2	1.0×10^{-3} eV/atom
σ_v^{energy} default	5.0×10^{-3} eV/atom
σ_v^{force} DB1	1.0×10^{-2} eV/Å
σ_v^{force} DB2	5.0×10^{-2} eV/Å
σ_v^{force} default	2.0×10^{-1} eV/Å
σ_v^{virial}	1.0×10^{-2} eV/atom
σ_w	1.0 eV
σ_{atom}	0.5 Å
ξ	4
n_{max}	12
l_{max}	12
GAP software version	1469201250
Representative environments	4500
sparse method	CUR

where the sum runs over some representative subset s of training configurations, usually far fewer than the total training set. The kernel function $K(\mathbf{q}_i, \mathbf{q}_j)$ of two local atomic environments, represented by their sets of descriptors \mathbf{q}_i and \mathbf{q}_j , corresponds to the expected covariance of their respective local atomic energies $\epsilon(\mathbf{q}_i)$ and $\epsilon(\mathbf{q}_j)$, and can be interpreted as a measure of similarity of the two local atomic environments. In the present work, we use the “smooth overlap of atomic positions” (SOAP) kernel developed by Bartók *et al.* [41], which is equivalent to choosing a polynomial kernel function

$$K(\mathbf{q}_i, \mathbf{q}_j) = \sigma_w^2 |\hat{\mathbf{q}}_i \cdot \hat{\mathbf{q}}_j|^\xi, \quad (3)$$

where the descriptor $\hat{\mathbf{q}}$ is the rotational power spectrum of the local atomic environment, which is a smooth and regular function, invariant to rotation and permutation of like atoms. All hyperparameters, including those inherent in the definition of the rotational power spectrum, are shown in Table I and their role is extensively discussed in Ref. [42]. The physically motivated hyperparameters include the energy scale σ_w , which roughly corresponds to the expected standard deviation of the atomic energy, and the length scale σ_{atom} , which controls the regularity of the potential. The power spectrum of the local environment includes a cutoff function that is zero for $r > r_{\text{cut}}$, a parameter whose choice is governed by the decay of the force constant matrix, since the potential will give exactly zero force constants for $r > 2r_{\text{cut}}$ by construction.

The vector of coefficients $\boldsymbol{\alpha}$ is obtained by substituting the training data into Eq. (2) and solving the linear system. We briefly outline the necessary steps, see Refs. [42,46] for further detail. Since the decomposition into atomic energies is not available from electronic structure calculations, the training data comprises total energies, and its derivatives (forces and virial stresses) corresponding to collections of atoms. Let us define \mathbf{y} as the vector with D components containing the target data: all total energies, forces, and virial stress components in the training database, and \mathbf{y}' as the vector with N components

containing the *unknown* atomic energies of the N atomic environments in the database, and \mathbf{L} as the linear differential operator of size $N \times D$, which connects \mathbf{y} with \mathbf{y}' such that $\mathbf{y} = \mathbf{L}^T \mathbf{y}'$. After selecting M representative atomic environments (with $M \ll N$), the expression for the coefficients in Eq. (2) is given by [47]

$$\boldsymbol{\alpha} = [\mathbf{K}_{MM} + \mathbf{K}_{MN} \mathbf{L} \Lambda^{-1} \mathbf{L}^T \mathbf{K}_{NM}]^{-1} \mathbf{K}_{MN} \mathbf{L} \Lambda^{-1} \mathbf{y}, \quad (4)$$

where K_{MM} is the covariance matrix between the M representative atomic environments, and K_{MN} is the covariance matrix between the M representative environments and all N environments in the training data. (In the Gaussian process literature, using a subset of the data to construct the basis is called *sparsification*.) While taking $\Lambda = \sigma_v^2 \mathbf{I}$ as the *regularization* matrix would be sufficient to solve the linear system (corresponding to simple L_2 regularization), the probabilistic interpretation of Gaussian process regression suggests that the elements of Λ are the tolerances, or expected errors in components of the training data vector \mathbf{y} , with even different units for different types of input data. Note that the expected errors are not just due to lack of numerical convergence in the electronic structure calculations, but also include the *model error* of the GAP representation, e.g., due to the finite cutoff of the local environment. Our informed choices for these parameters are reported in Table I. The representative local environments are chosen by the CUR matrix decomposition procedure [48] applied to the matrix of descriptor vectors in the input dataset which essentially finds a subset of the atomic environments that would lead to a good low-rank approximation of the full covariance matrix. The upshot of using only a small number of representative atomic environments is that the computational cost to train the model scales as $O(NM^2)$ rather than $O(N^3)$, and the cost of evaluating a single local atomic energy scales as $O(M)$ rather than $O(N)$. Typically, we find that $M < 10\,000$ is sufficient (in the sense that prediction results do not improve when a larger M is used) even when $N > 150\,000$.

We trained the GAP model using the QUIP software code, which is publicly available [49], and the full set of command line parameters as follows:

```
at_file=data.xyz gap={soap l_max=12 n_max=12
cutoff=5.0 cutoff_transition_width=1.0 delta=1.0
atom_sigma=0.5 zeta=4 config_type_n_sparse=
{slice_sample_high:500:phonons_54_high:500:
phonons_128_high:500:default:3000}
sparse_method=cur_points
covariance_type=dot_product}
sparse_jitter=1e-12 default_sigma={0.005 0.2
1.0 0.0} config_type_sigma={slice_sample_high:
0.0001:0.01:0.01:0.0:phonons_54_high:0.001:
0.05:1.0:0.0:phonons_128_high:0.001:0.05:
1.0:0.0}
```

III. DATABASE

A large training database of electronic structure calculations is required in order to ensure transferability of flexible GAP models to a wide range of atomic environments. In what follows we discuss the details of how we generated such database.

A. Generation protocol

We choose to include in the database only first-principles data. Although computationally costly, this approach allows for a direct control and propagation of the accuracy and the degree of convergence of the data entering the training procedure. The database generation protocol that we adopt can be rationalized as follows. (1) We start by selecting the physical properties that we require to be well reproduced or predicted by our model. For each material property of interest, we select a number of representative *small* periodic configurations (with varying cell parameters and atomic positions) that are amenable for first-principles calculations and covers the relevant local atomic environments needed for the potential to reproduce that property. (2) We sample the configurational space associated to each unit cell selected in (1) by means of Monte Carlo or molecular dynamics techniques using density functional theory calculations that are configured to have only a moderate level of convergence. (3) From each sampling run, we extract a weakly correlated subset of configurations. Each of these subdatabases is denoted as DB x . (4) Finally, we recompute total energies, forces, and stresses for each configuration in each subdatabase using highly converged parameters in order to minimize the stochastic and systematic errors due to the finite k -point sampling and plane-wave cutoff. Even so, it is not possible to use the same (consistent) k -point sampling across the entire database due to resource limitations, and the resulting varying degrees of convergence are used to inform the magnitude of the regularization terms corresponding to each subdatabase, as shown in Table I.

B. Training configurations

The complete database consists of eight subdatabases, which include 12193 configurations, equivalent approximately to 1.5×10^5 atomic environments. The details of each subdatabase are described below (see also Ref. [50]) and also summarized in Table II for simplicity.

DB1 aims at training around the bcc equilibrium geometry and the elastic response of the bulk. It consists of energies and stresses computed for one-atom cells whose vectors are distorted with respect to the equilibrium bcc primitive cell geometry. The distortions are randomly obtained using a slice-sampling MC algorithm and performed with respect to various reference volumes which are compressed or expanded with respect to the 0 K DFT equilibrium value as reported in Table II.

DB2 is used to teach bulk vibrational properties and consists of total energies and forces computed from $3 \times 3 \times 3$ and $4 \times 4 \times 4$ conventional cubic supercells containing 54 and 128 atoms, respectively. The configurations are extracted from MD runs equilibrated at the volumes and temperatures shown in Table II.

DB3 similarly to DB2 consists of total energies and forces computed from $3 \times 3 \times 3$ cubic supercells generated from MD runs also equilibrated at various volumes and temperatures reported in Table II. This subdatabase is used to teach bulk monovacancy energetics. As such, the unit cells contain 53 atoms.

DB4 provides information on the divacancy energetics. Divacancy environments up to third-nearest neighbor are explicitly included. This subdatabase consists of total energies and

TABLE II. Database details used for training the α -Fe GAP. For each subdatabase (DB), we report the name of the physical properties focus of the training, the physical quantities explicitly used for training, the number of training local environments, the volume (expressed in percentage variation with respect to the electronic equilibrium value), the temperature, the number of atoms, the simulation box used for the generation of the configurations, the k spacing used for the accurate calculations, and other details concerning the type of environments within the DB. The notation for trivacancy identification is taken from Ref. [60].

	Target property	Total number of LAEs	V (% V_0)	T (K)	Number atoms in unit cell	Simulation box	k spacing (\AA^{-1})	notes
DB1	bulk elastic constants	6001	-0.81/1.08/3.55	300	1	primitive bcc distorted	0.015	-
DB2	bulk phonons	12474 11520	0.0/ \pm 2.08/3.55	400–1400 800	54 128	$3 \times 3 \times 3$ $4 \times 4 \times 4$	0.03	-
DB3	bulk monovacancies	20193	0.0/ \pm 2.08	400–1000	53	$3 \times 3 \times 3$	0.03	-
DB4	bulk divacancies	10836	0.0	800	126	$4 \times 4 \times 4$	0.03	1-, 2-, 3- <i>nn</i>
	trivacancies	9375	0.0	800	125	$4 \times 4 \times 4$	0.03	[112],[113], [223],[333],[339]
DB5	vacancy clusters	1736 1476	0.0 0.0	800–1000 600	124 123	$4 \times 4 \times 4$ $4 \times 4 \times 4$	0.03 0.03	4 vac. 5 vac.
		2709	0.0	100	129	$4 \times 4 \times 4$	0.03	dumbbell ₁₀₀
		1548	0.0	300	129	$4 \times 4 \times 4$	0.03	dumbbell ₁₁₀
DB6	self-interstitials	4773	0.0	100–300	129	$4 \times 4 \times 4$	0.03	crowdion ₁₁₁
		3225	0.0	100–300	129	$4 \times 4 \times 4$	0.03	tetrahedral
		2064	0.0	100	129	$4 \times 4 \times 4$	0.03	octahedral
	di-interstitials	2340	0.0	300	130	$4 \times 4 \times 4$	0.03	nonparallel
	bulk	660	0.0	300	12	$1 \times 1 \times 6$	0.03	(100)
DB7	terminated surfaces	588	0.0	300	12	$1 \times 1 \times 6$	0.025	(110)
		516	0.0	300	12	$1 \times 1 \times 6$	0.04	(111)
		648	0.0	300	12	$1 \times 1 \times 12$	0.025	(211)
DB8	γ surfaces	30000	0.0	300	12	primitive xy	0.03	(110)
		29388	0.0	300	12	primitive xy	0.025	(211)
DB		152070						

forces of $4 \times 4 \times 4$ conventional cubic supercells containing 126 atoms and obtained from MD equilibrated at 800 K and at the equilibrium volume.

DB5 embodies selected trivacancies and small vacancy clusters such as tetravacancies and pentavacancies (see Fig. 1) that should provide a starting point for describing nanovoids. We choose those trivacancy configurations, which lie in low Miller index crystallographic planes $\{100\}$, $\{110\}$, and $\{111\}$ and that, in those planes, are most localized [51]. Total energies and forces from $4 \times 4 \times 4$ cubic supercell configurations obtained from MD are used as training quantities.

DB6 embraces relevant self-interstitial environments, including the $\langle 100 \rangle / \langle 110 \rangle$ dumbbell, $\langle 111 \rangle$ crowdion, and the tetrahedral and octahedral configurations. A type of nonparallel di-interstitial configuration (see Fig. 1) is also considered to cover further defective environments beyond simple self-interstitials. The configurational space of all these point defects is sampled by means of MD performed on cubic $4 \times 4 \times 4$ supercells containing 129/130 atoms at the theoretical equilibrium bulk volume at 0 K. Training is from total energies and forces.

DB7 consists of total energies and forces of bulk-terminated surface configurations with $\{100\}$, $\{110\}$, $\{111\}$, and $\{211\}$ crystallographic orientations. For this subdatabase, we choose supercells that are elongated along c and primitive in the sur-

face plane (at the equilibrium lattice parameter). We simulate slabs that are 12 atomic layers thick to minimize interactions between the two surfaces of the slabs. A vacuum separation region of 16 \AA is also used to avoid replica interactions in the c direction. Molecular dynamics is performed on these cells with the atoms allowed to move only along z in order to gain insight mainly on the out-of-plane surface relaxation of the atoms at the vacuum-slab interface.

DB8 is generated to train on γ surfaces, to be able to ensure coverage of local environments found typically around dislocation cores. In particular, we consider $\{110\}$ and $\{211\}$ crystallographic orientations that are the most important slip planes for bcc metals. As with the bulk-terminated surfaces, we use supercells elongated along c , which contain 12 atomic layers. Configurations are created in a 10×10 grid of slips in directions in the glide plane of the gamma surface. Total energies and forces are used as training quantities.

C. Computational details

MD simulations are all performed in a NVT ensemble with time steps ranging from 2 to 4 fs and a Berendsen thermostat [52]. Temperatures and volumes are varied as specified in the section above (details are reported in Table II). This is done to ensure a good coverage of the physical properties of

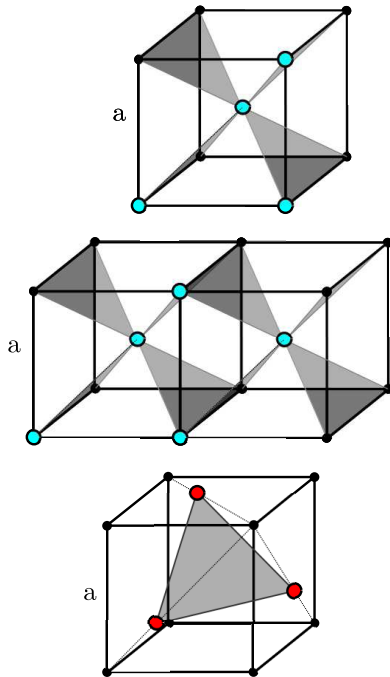


FIG. 1. From top to bottom we show the schematics of tetra-, pentavacancy, and nonparallel di-interstitials of DB5/DB6. Azure circles represent the missing atoms in an otherwise perfect bcc structure. Red circles schematically represent the atomic arrangement of a nonparallel di-interstitial defect.

interest across the theoretical thermodynamic range of stability of the α -phase of iron [7]. Sampling of self-interstitial defects requires some attention since most of them are metastable states which tend to rapidly relax to more stable configurations during the MD. In those cases, we perform very short MD runs at low temperatures trying to capture the transition pathway to lower energy states.

Monte-Carlo sampling was originally performed in Ref. [42] for bcc tungsten at 300 K and exploiting a slice-sampling technique. Here, we simply take the same periodic cells and rescale them to account for the differences in the lattice parameter and elastic constants of ferromagnetic iron and tungsten.

All quantum-mechanical calculations of this work are performed in a collinear spin-polarized plane waves DFT framework as implemented in the QUANTUM ESPRESSO distribution [53], employing an ultrasoft GGA PBE [54] pseudopotential from the *0.2.1 pslibrary* [55] with semicore electrons in valence. This pseudopotential has been carefully tested and has proved to be able to reproduce the correct all-electrons behavior for a number of ground state properties. The exchange-correlation functional used provides a relatively good description of thermomechanical properties for the α phase [7]. At the same time, it is a reliable choice for reproducing point defect properties [56].

For the accurate calculations mentioned at step (4) of the generation protocol, all input parameters are chosen to ensure convergence to 1 meV/at, 0.01 eV/Å, and 0.01 GPa for the energy difference, forces, and stresses, respectively. In particular, a value of 90 Ry on the wave function (dual of 12 [57]) is required for the convergence of energy differences and

forces. The convergence of stresses instead requires a cutoff value of 144 Ry (dual of 12). It is important to stress that none of these values is, however, sufficient to ensure proper convergence of the total energy. As a consequence, in order to avoid inconsistencies (that would affect the training procedure) between total energies of DB1 and those of DB x with $x > 1$, we have built DB1 as a combination of stresses computed at 144 Ry and of total energies computed at 90 Ry. The BZ is integrated with a Monkhorst-Pack grid and a Marzari-Vanderbilt smearing scheme [58] at an effective temperature of 0.01 Ry. In order to ensure the level of convergence mentioned above, we found critical to choose a sampling density so that for all the subdatabase cells the largest k spacing along any reciprocal cell vector is below 0.03 \AA^{-1} . Exact k spacing values for each DB are reported in Table II for completeness. Note in the table that slightly different k -point densities are used for some of the data. This is due to computational costs (for primitive unit cells one can afford higher densities than for large supercells) but also due to the incommensurate nature of the simulation boxes used in the different DBs. The k spacing has a typical value of 0.025 \AA^{-1} with a standard deviation of 0.005 \AA^{-1} . The AiiDA materials' informatics infrastructure [59] has been partially used as a tool to automate submission of accurate calculations of the generation protocol, and to provide easy access to provenance information of all the data of the training database.

For the calculations at step (2) of the generation protocol, i.e., those related to the sampling of the quantum-mechanical PES, we do not require such level of accuracy. In fact, in this case, we use lower cutoff values of 60 Ry (dual 8) with a reduced k sampling of the Brillouin zone.

IV. RESULTS

In this section, we present the GAP model for bcc iron, which has been trained on the database of Table II with generation details reported in Table I. Validation is performed in the following sections through an analysis of the energetics and of the thermomechanic properties of the α phase by comparing with DFT data; comparisons with experiments (when possible) are also reported. The DFT calculations that are used for comparison are either taken from the literature or computed in this work with input parameters consistent with those described in Sec. III C. The latter are considered part of a testing set and are not used for training.

A. Fundamentals

We start our analysis showing in Table III the lattice parameter a_0 , the bulk modulus B_0 , and the elastic constants C_{11} , C_{12} , and C_{44} calculated with GAP at zero temperature (with and without zero-point contributions). The results are in excellent agreement with the quantum-mechanical data and, as previously discussed in Ref. [7], reflect the inherent limitations of standard DFT approaches to deal with magnetism [61]. The equation of state (EOS) reported in Fig. 2 shows how close GAP is with respect to the DFT curve even relatively far from the equilibrium volume. The maximum energy difference between the two EOS curves in the volume interval $[11.0:12.0] \text{ \AA}^3$ around the electronic equilibrium is $\approx 0.3 \text{ meV}$ [62], with a

TABLE III. Lattice parameter, bulk modulus, and elastic constants for α -iron at zero temperature. GAP results are compared to DFT (with and without quasiharmonic zero-point energy contributions) and to experimental data at 0 K.

	GAP		DFT		Expt.
	no ZPE	ZPE	no ZPE	ZPE	
a_0 (Å)	2.834	2.839	2.834 [7]	2.839 [7]	2.855 [64]
B_0 (GPa)	198.2	191.7	199.8 \pm 0.1 [7] 196.9 ^a	194.6 \pm 0.3 [7]	170.3 \pm 1 [65]
C_{11}	285.9	–	296.7 \pm 0.3 [7]	287.9 \pm 0.4 [7]	239.5 \pm 1 [65]
C_{12}	154.3	–	151.4 \pm 0.2 [7]	148.0 \pm 0.5 [7]	135.7 [65]
C_{44}	103.8	–	104.7 \pm 0.1 [7]	102.2 \pm 0.5 [7]	120.7 \pm 0.1 [65]

^aThis work.

measure Δ of the distance between the two curves calculated *à la* Cottenier [63] of 0.112 meV/atom. In the inset of Fig. 2, we also report for reference the GAP and DFT electronic bulk moduli $B(V) = V \frac{\partial^2 E(V)}{\partial V^2}$.

B. Bain path

The Bain path traverses the diffusionless transformations between bcc, body-centered tetragonal (bct), and fcc crystal structures, by varying c/a for the cell. It shows the relative stabilities of the bcc and fcc phases and the energy barrier for the transformation. Since the GAP training data only includes ferromagnetic bcc data, the Bain path is an interesting test of the performance well outside of the training data. We compute the Bain path using a two-atom bcc cell, so that $c/a = 1.0$ is the bcc configuration and $c/a = \sqrt{2}$ is fcc. At each point along the pathway, the structure is set to a fixed value of c/a and the cell volume and the position of the central atom are relaxed to the minimum energy structure for that value of c/a . The GAP calculated Bain path is shown in Fig. 3. GAP is able to estimate the error in its prediction, and this variance is also plotted, showing the greatest uncertainty in the prediction for the fcc structure. Each structure has also been recalculated

with DFT (using the GAP optimized volume), and GAP shows notably an excellent agreement with ferromagnetic DFT across the entire Bain path. Changes in the magnetic ground state complicate the path for iron; we find that an antiferromagnetic double layer magnetic state would stabilize the fcc structure at the GAP optimized volumes, and the true Bain path involves a number of complex magnetic states [66–68]. Including magnetic behavior in subsequent development of GAP for iron would be a fascinating challenge.

C. Phonons

The GAP phonon dispersions are shown along high symmetry paths in the first BZ at the DFT equilibrium volume V_0 (without zero-point energy contributions) and at an expanded value corresponding approximately to the equilibrium volume at 1000 K predicted by DFT quasiharmonic theory [7] (namely +3.0% V_0). For each volume, the GAP dynamical matrix is obtained with a frozen-phonon method using a supercell corresponding to a $8 \times 8 \times 8$ primitive cell and finite displacements of 0.01 Å; it is then Fourier interpolated on a denser $32 \times 32 \times 32$ mesh to give smoother frequency dispersions. Calculations are performed with the QUIP+GAP code [49]. Results are compared in Fig. 4 (top panels) to the DFT data

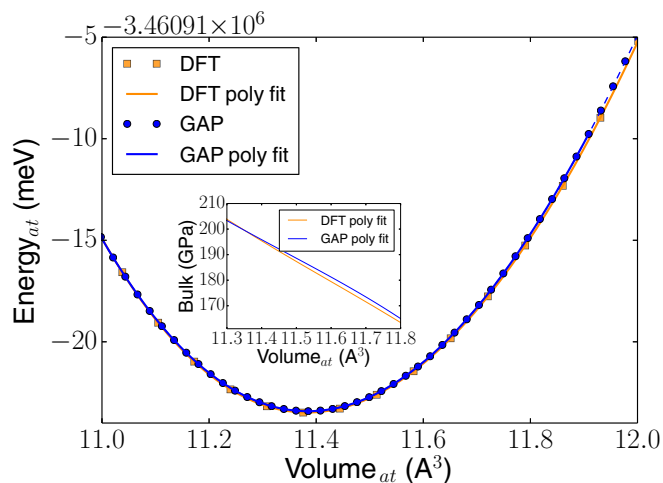


FIG. 2. Equation of state of the GAP potential (blue, circles) compared to DFT data (orange, squares) whose calculation details are consistent with those used for the database generation. In the inset, we report the GAP and DFT bulk moduli obtained analytically from a polynomial fitting of the corresponding total energy curves.

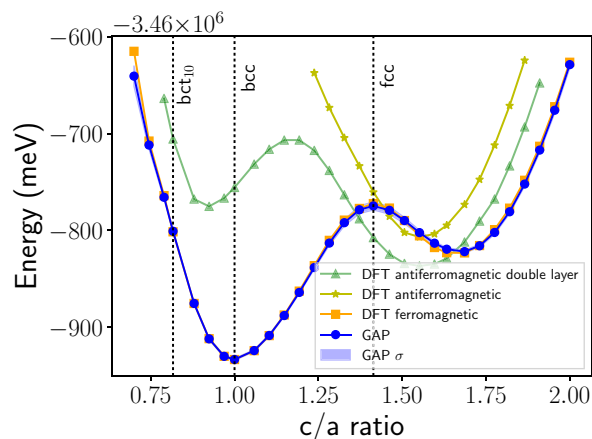


FIG. 3. Epitaxial Bain path calculated using GAP and optimizing the volume for each value of c/a . DFT energies are calculated using the same volume obtained in the GAP pathway. The GAP σ is the variance estimated by GAP due to extrapolation outside of the training data.

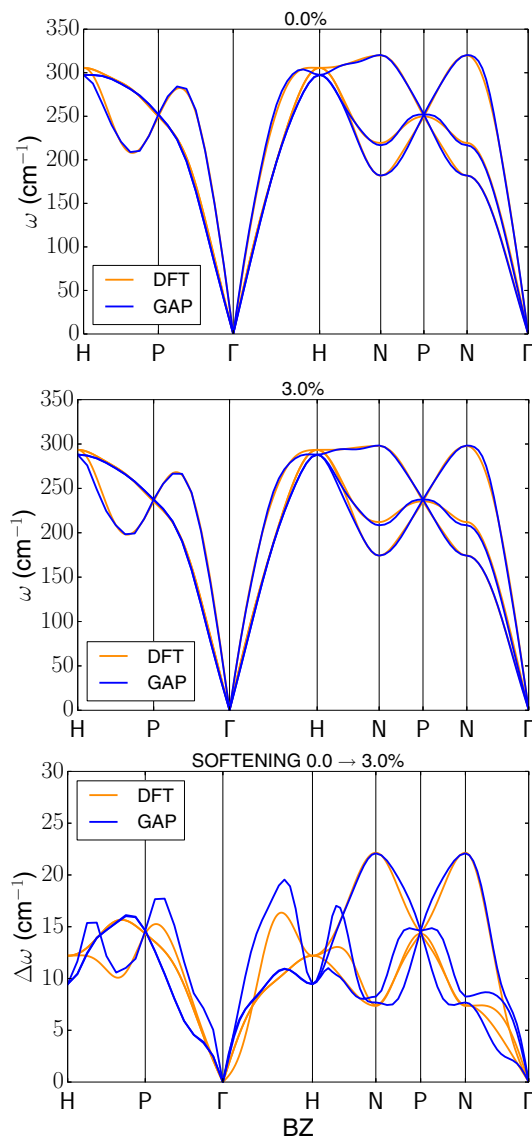


FIG. 4. (Top) From top to bottom we report the phonon dispersions computed at DFT equilibrium volume V_0 (without zero-point energy contributions) and at $+3\% V_0$ (approximately the QHA DFT equilibrium volume at 1000 K). GAP frequencies (orange solid lines) are obtained from frozen-phonon calculations on a supercell and are compared to DFT values (blue solid line) obtained from density-functional perturbation theory on a $4 \times 4 \times 4$ mesh. (Bottom) Softening of the phonon frequencies along the dispersion path due to the change in volume from $0.0\% \rightarrow 3.0\% V_0$. As above, the three blue lines are the three GAP modes while the three orange lines are the three DFT modes.

from Ref. [7]. In the bottom panel of Fig. 4, we also check the phonon softening between the two volumes since the softening of phonon frequencies as a function of volume is critical to the thermal expansion and ultimately for the thermodynamic properties of a material.

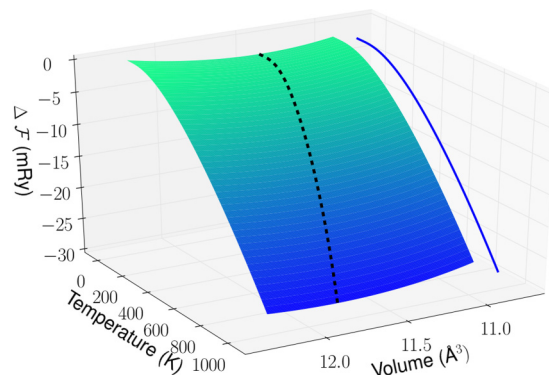


FIG. 5. Helmholtz free energy computed with GAP in the quasi-harmonic approximation. The black dashed curve is the locus of the points where the free energy is minimized at each temperature with respect to the volume. Its projection in the free energy-temperature plane is also reported.

D. Bulk thermodynamics

An important aspect, not always taken under consideration during the validation process of an interatomic potential, is the ability to reproduce finite-temperature properties. Iron has the bcc α phase displaying a ferromagnetic behavior and a Curie temperature of 1043 K. In fact, it has been shown that magnetic excitations come into play for the description of many thermodynamic quantities only above a large fraction of the Curie point [7,9,11,69]. As such, we can neglect them in first approximation and assume the thermal properties of the α phase as dominated by atomic vibrations. Given that our GAP model provides an excellent description of bulk vibrations, we then expect good finite-temperature performance.

We start our analysis of the bulk thermal properties making use of the quasiharmonic approach, which provides an accurate tool to access the low-temperature regime taking into account quantum statistical effects. By computing and integrating the phonon dispersions at 57 different volumes, from -3.6% up to $7.6\% V_0$ in steps of 0.2% of the electronic equilibrium volume, and using the same calculation details described in Sec. IV C, we obtain the Helmholtz free energy (see Fig. 5). From that, we calculate all the relevant thermodynamic quantities of interest in the quasiharmonic approximation. In parallel, in order to study the same quantities obtained from QHA in the high-temperature regime, where stronger anharmonicity comes into play and quantum statistical effects lose importance, we use an MD approach. We perform NPT runs at vanishing external pressure to find the equilibrium density at different temperatures from 200 to 1800 K in steps of 200 K. We use a $8 \times 8 \times 8$ supercell with 1024 atoms, a time step of 1 fs with temperature and pressure controlled by a Nose-Hoover chain thermostat [70], and a Parrinello-Rahman barostat [71] as implemented in the LAMMPS [72] package.

The first quantity that we analyze is the thermal expansion. In Fig. 6, we show the GAP QHA curve, which follows the proper quantum Bose-Einstein (BE) statistics, the GAP QHA modified to follow the classical Maxwell-Boltzmann (MB) statistics (zero point energy contribution is not included), and the GAP curve resulting from MD calculations. For comparison, we show the DFT QHA (BE) curve [7] plus

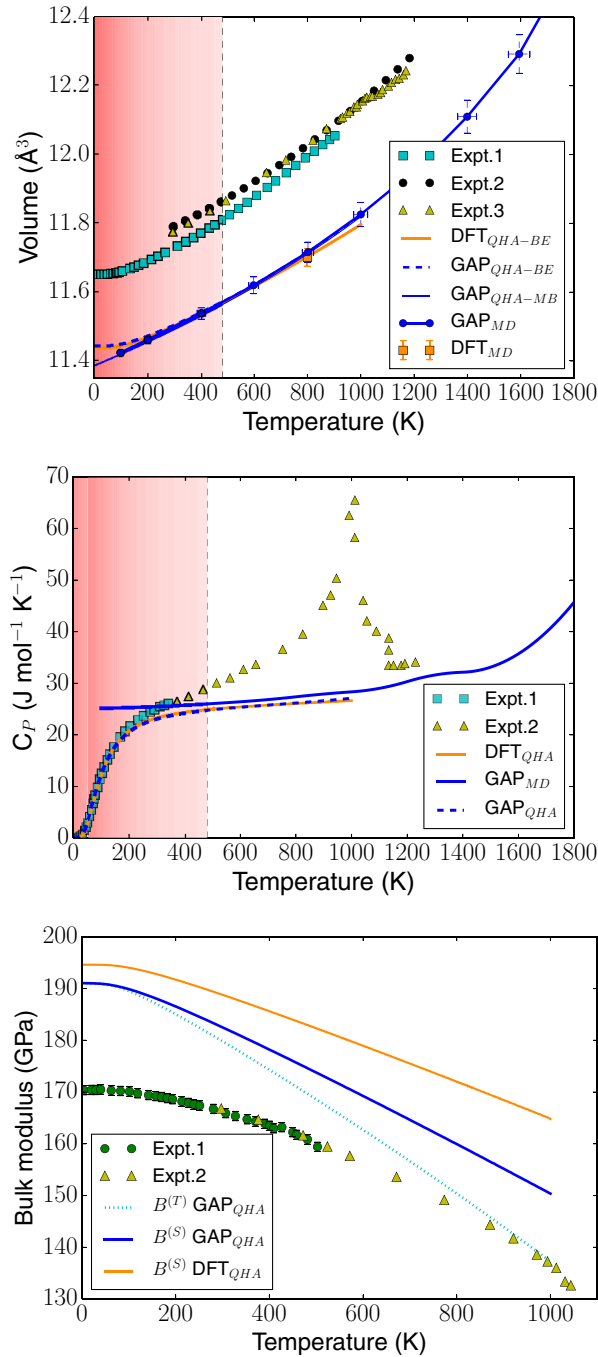


FIG. 6. (Top) Thermal expansion of GAP obtained from MD (blue solid line) and QHA (blue dashed line is QHA with Bose-Einstein statistics, blue thin line is QHA with Maxwell-Boltzmann statistics). Results are compared to DFT QHA data (orange solid line), to a single DFT MD point at 800 K, and to experiments (squares, triangles and circles data from Refs. [64,75,76]). Middle panel: heat capacity at constant pressure as a function of temperature. GAP QHA (dashed blue line) and GAP MD (solid blue line) are compared to DFT QHA (solid orange line) and to experimental data from Refs. [77,78] (squares and triangles, respectively). (Bottom) Adiabatic (blue solid line) and isothermal (azure dotted line) bulk modulus as a function of temperature from GAP. Comparison with DFT adiabatic data (orange solid line) from Ref. [7] and experiments from Refs. [65,79] (circles and triangles).

an estimate of the equilibrium volume at 800 K from DFT molecular dynamics. As a reference, we also report three sets of experimental data. It is immediately possible to note that the GAP QHA (BE) curve agrees remarkably well with the DFT QHA one up to 1000 K. The DFT and GAP results instead underestimate experiments. As exhaustively discussed in Ref. [7], this can be attributed to the DFT PBE functional which has been adopted for the database generation and for the DFT data used for comparison. Nonetheless, the experimental thermal trend is overall well reproduced. The GAP MD and the GAP QHA curves agree well up to 800 K, while they start to deviate above this temperature. The GAP MD curve also matches the DFT MD equilibrium volume at 800 K. This analysis seems to suggest that beyond quasiharmonic effects start to play a role only above 800 K. Interestingly, the MD results overlap with quasi-harmonic results modified to artificially reproduce a classical Maxwell-Boltzmann behavior at temperatures below 200 K. We finally notice that the bcc phase appears mechanically stable up to approximately the experimental melting point [73]. From the knowledge of the temperature-volume relation at equilibrium, we then calculate the temperature dependence of other relevant bulk thermodynamic quantities. The heat capacity at constant pressure results are reported in Fig. 6, including QHA and MD data. As for the thermal expansion, the heat capacities obtained with QHA and MD nicely converge at intermediate temperature. The experimental divergence at the Curie point is related to magnetic entropy [69,74]; as such, it is not captured by our DFT calculations and, consequently, by our model. Within the quasiharmonic framework, the heat capacity is used also to compute the adiabatic bulk modulus thermal behavior starting from the isothermal one as discussed in Ref. [7]. In Fig. 6, we show that GAP is capable to reproduce well the overall DFT thermal behavior, although slightly underestimating (in the direction of the experimental data) the absolute values. Since the bulk moduli are second partial derivatives of the Helmholtz free energy, these results reflect the ability of the model to accurately reproduce the details of the bulk quantum-mechanical PES.

E. Bulk point defects

Real crystals are far from being perfect, and contain defects that can be, e.g., pointlike or extended in space. Their study is fundamental for understanding the microscopic processes that govern the actual response of a macroscopic system under different external conditions. It is therefore important to test the capabilities of GAP in describing the energetics of some simple defects.

1. Atomic vacancies

We start from the monovacancy, which consists of a missing atom in an infinite lattice. This missing atom is assumed to be isolated, i.e., not interacting with any other defect in the surroundings. The energy of formation of a monovacancy, i.e., the cost of removing an atom from the perfect bulk, at the equilibrium volume is reported in Table IV. In addition, the dependence of the formation energy upon volume [80] is shown in Fig. 7. We also compute the energy profile or minimum energy path for a monovacancy migration to a first-, second-

TABLE IV. Formation and binding energies of defected configurations from GAP. Results are obtained with fully relaxed cells except for NEB calculations, which are performed at the equilibrium (electronic DFT) volume and compared to DFT (from this work and from the literature) and, when possible, to experimental data at 0 K.

	GAP	DFT (this work)	Other DFT calcs.	Expt.
E_f^v (eV)	2.26	2.22	2.15 [89], 2.07 [90], 2.10 [103]	1.6 [104], 2.0 [105]
E_{m1NN}^v	0.67	–	0.67 [86], 0.64 [87]	0.55 [106]
E_{m2NN}^v	2.75	–	–	–
E_{m3NN}^v	5.63	–	–	–
E_f^{1NNv}	4.41	4.24	4.02 [103]	–
E_f^{2NNv}	4.30	4.20	3.76 [87], 3.96 [103]	–
E_f^{3NNv}	4.55	4.45	–	–
E_f^{4NNv}	4.48	–	–	–
E_f^{5NNv}	4.47	–	–	–
E_b^{1NNv}	0.11	0.20	0.16 [18], 0.14 [87], 0.08 [103]	–
E_b^{2NNv}	0.22	0.24	0.23 [18], 0.28 [87], 0.15 [103], 0.2 [107]	–
E_b^{3NNv}	–0.03	–0.01	–0.015 [18], –0.02 [87]	–
E_b^{4NNv}	0.04	–	0.05 [18]	–
E_b^{5NNv}	0.05	–	0.06 [18]	–
$E_f^{[112]v}$	6.19	–	$E_f^{[112]v} < E_f^{[226]v}$ [83], 5.82 [84], 5.42 [87]	–
$E_f^{[226]v}$	6.38	–	$E_f^{[226]v} < E_f^{[223]v}$ [83], 6.13 [84]	–
$E_f^{[223]v}$	6.35	–	$E_f^{[223]v} < E_f^{[115]v}$ [83], 6.70 [84]	–
$E_f^{[115]v}$	6.47	–	6.15 [84]	–
$E_f^{[113]v}$	6.59	–	6.14 [84]	–
$E_f^{[333]v}$	6.85	–	–	–
$E_f^{[339]v}$	6.85	–	–	–
$E_b^{[112]v}$	0.60	–	0.66 [84], 0.37 [86]	–
$E_b^{[226]v}$	0.41	–	0.35 [84]	–
$E_b^{[223]v}$	0.44	–	–0.22 [84]	–
$E_b^{[115]v}$	0.32	–	0.33 [84]	–
$E_b^{[113]v}$	0.20	–	0.34 [84]	–
$E_b^{[333]v}$	–0.07	–	–	–
$E_b^{[339]v}$	–0.06	–	–	–
$E_{f(110)}^{SIA}$	4.21	4.37	3.77 [18], 3.93 [56], 4.02 [89], 3.64 [90], 3.94 [91]	4.7–5 [108]
$E_{f(111)}^{SIA}$	4.90	5.13	4.64 [56], 4.72 [89], 4.34 [90], 4.66 [91]	–
$E_{f(100)}^{SIA}$	5.47	5.48	4.80 [18], 5.05 [56], 5.13 [89], 4.64 [90], 5.04 [91]	–
$E_{f_{tet}}^{SIA}$	4.75	4.79	4.28 [18], 4.32 [56], 4.44 [89], 4.26 [90]	–
$E_{f_{oct}}^{SIA}$	5.53	5.58	4.97 [18], 5.21 [56], 5.29 [89], 4.94 [90]	–
$E_{m(110)}^{jump}$	0.31	–	0.34 [90]	0.30 [18]
E_{NPC}^{diSIA}	7.54	7.84	7.04 [19]	–
E_{110}^{diSIA}	8.36	8.95	7.15 [19], 6.56 [90]	–
E^{110} (J/m ²)	2.499	2.495	2.27 [92], 2.37 [93], 2.25 [109]	–
E^{100}	2.547	2.543	2.29 [92], 2.47 [93], 2.25 [109]	–
E^{211}	2.612	2.629	2.50 [93]	–
E^{111}	2.756	2.752	2.52 [92], 2.58 [93], 2.54 [109]	–

, and third-nearest-neighbor site through nudge-elastic band [81] (NEB) calculations. The energy profiles are reported in Fig. 8 and the corresponding migration energy barriers are summarized in Table IV. Results are closely consistent with DFT calculations.

Next, we consider divacancy defects, where two missing atoms are simultaneously present and interact with each other in the crystal. The formation energy and binding energy of first-, second-, third-, fourth-, and fifth-nearest-neighbor divacancies are reported in Fig. 9 and summarized in Table IV. In agreement with Refs. [18,82] the binding energy of the third-nearest neighbor is negative, thus suggesting the instability of such configuration compared to the condition of two

isolated monovacancies. As expected from DFT calculations, but contrarily to most of the semiempirical models available in the literature [18], the fifth-*nn* configuration is reported to be positive.

An analysis of selected trivacancy defects, identified here by means of the Beeler notation [60], is then carried out. As for divacancies, trivacancy calculations are performed using a $10 \times 10 \times 10$ conventional cubic supercell at zero pressure condition. The computed formation energies are summarized in Table IV, along with the corresponding binding energies also reported in Fig. 10. Results are in good agreement with available DFT data and suggest a ground-state [112] configuration [83] with a suppression of binding on the {111}

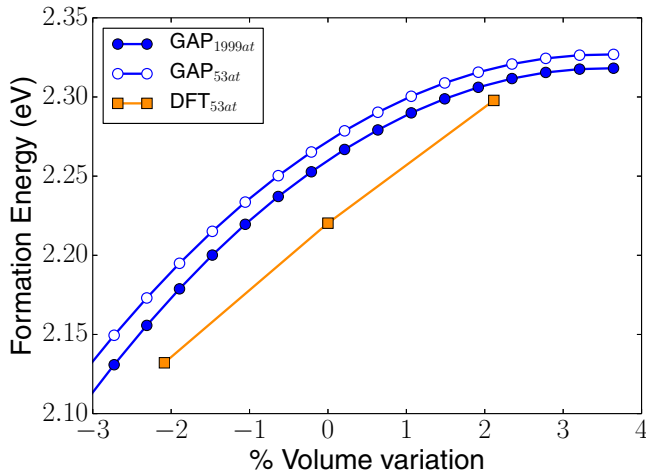


FIG. 7. Monovacancy formation energy as a function of percent-age volume variation with respect to the (electronic) equilibrium value. GAP results (blue full and empty circles) are obtained relaxing atomic positions in a $10 \times 10 \times 10$ and in a $3 \times 3 \times 3$ conventional cubic supercells, while DFT data are obtained relaxing only atomic positions of a $3 \times 3 \times 3$ cubic supercell and are reported as orange circles.

plane. The cost of formation of the [223] configuration is predicted lower than the [115] one in accordance with the PBE data of Ref. [83]. At the same time, at variance with DFT data, the ordering of formation of the [226] and [223] configurations is swapped. These findings, along with the fact that neither the [226] nor the [115] are included in the training database, suggest that some caution is needed when the potential is used as an extrapolation. Interestingly, at odds with results of Ref. [83], the authors of Ref. [84] report a positive formation energy difference between the [223] and

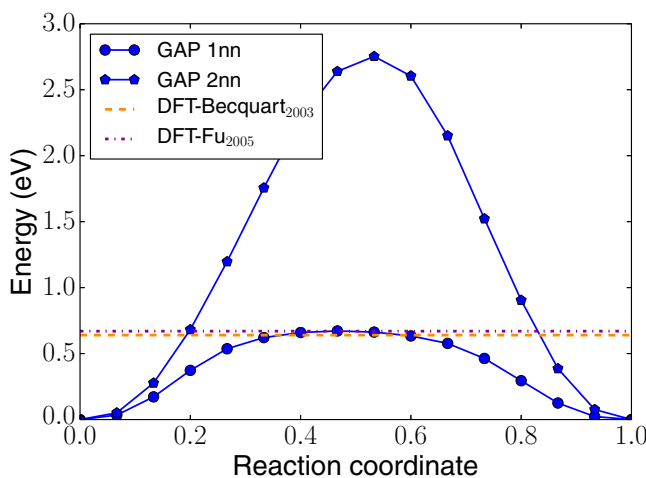


FIG. 8. Migration energy profiles of a monovacancy diffusing to a first- and second-nearest-neighbor site (blue circles and pentagons respectively) obtained from nudged-elastic-band calculations. The calculations are performed using 249 atoms in a $5 \times 5 \times 5$ cubic supercell. DFT monovacancy first-nearest-neighbor migration energy barriers are taken from Fu *et al.* [86] (dash-dot line) and Becquart *et al.* [87] (dashed line).

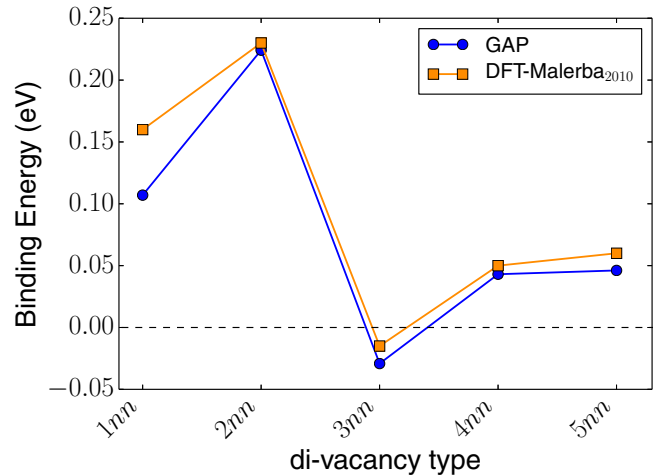


FIG. 9. Divacancy binding energy landscape. GAP results (blue circles connected by solid lines) are compared to DFT data from Ref. [18] (orange squares). Formation energy values are also reported in Table IV.

[115] configurations. Such discrepancy at the DFT level is here attributed to the different exchange-correlation functional used in the two calculations (PBE and PW91 [85], respectively), and it might also justify the qualitative difference in the [223] binding energy obtained by our GAP model, which has been trained on PBE data, and the corresponding binding energy obtained from PW91 calculations. Although we have included a few four and five vacancies configurations in the training database, an extensive analysis of tetra- and pentavacancies will be performed elsewhere.

2. Self-interstitials

Next, we consider self-interstitial atoms (SIAs), with particular interest for the crowdion₁₁₁, dumbbell₁₁₀, dumbbell₁₀₀, tetrahedral, and octahedral configurations. These are in fact the simple self-interstitial defects in bcc iron [18], which are considered the most relevant in the study of damage

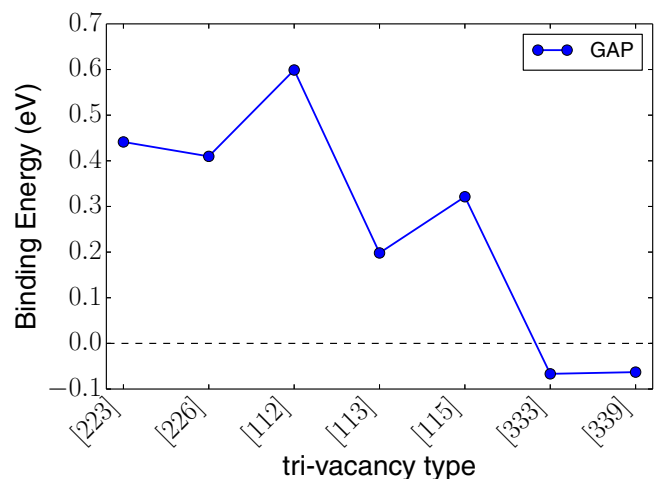


FIG. 10. Trivacancy binding energy landscape. Formation energy values are also reported in Table IV.

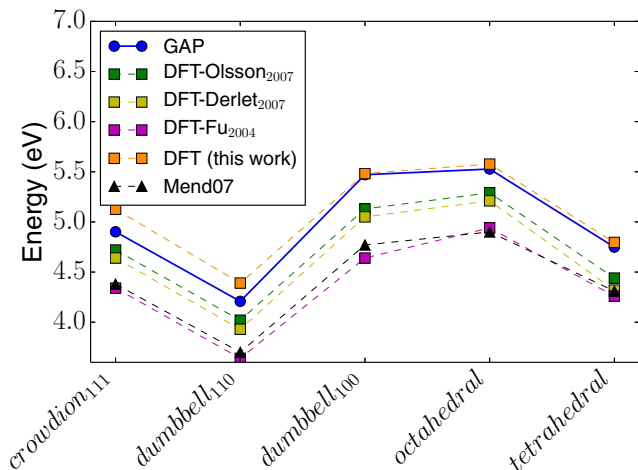


FIG. 11. Selected self-interstitial formation energies obtained from GAP. Results are compared to DFT data calculated in this work and others available in the literature, namely, Derlet *et al.* [56], Olsson *et al.* [89], and Fu *et al.* [90]. Results are also compared to a successful EAM model well known in the literature [18], the so-called Mendev07.

and aging of steel reactor vessels under strong irradiation. GAP formation and binding energies are reported in Table IV with formation energies also summarized in Fig. 11. As expected from our DFT calculations [88] and other DFT studies [56,89–91], we find that the most stable GAP interstitial is the dumbbell₁₁₀ configuration, followed by the tetrahedral, crowdion₁₁₁, dumbbell₁₀₀, and octahedral ones. The renormalization of the atomic distances of the atoms of the 111 string along the $\langle 111 \rangle$ direction are reported in Fig. 12 for reference. We have also computed the migration energy barrier

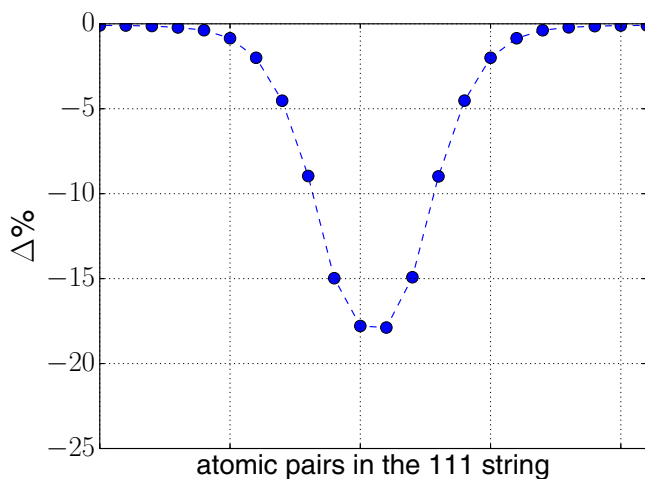


FIG. 12. Crowdion displacement field along the 111 string as predicted by GAP. $\Delta\% = (d - a_{111}^0)/a_{111}^0 \times 100$, with d being the distance between two consecutive atoms in the 111 string of the relaxed defective configuration, and $a_{111}^0 = \sqrt{3}/2 a_0$ the same quantity in a perfect lattice at the (electronic) equilibrium. The minimum of the curve indicates where the atomic pairs are closest. Here we use a $10 \times 10 \times 10$ cubic supercell containing 2001 atoms in total so that we have 21 atomic distances to consider along the $\langle 111 \rangle$ direction.

for a dumbbell₁₁₀ to jump to a first-nearest-neighbor site. The jump mechanism to the first-nearest neighbor is consistent with the one observed within the DFT framework [18]. The migration energy barrier E_m^{jump} of such mechanism is reported again in Table IV. At variance with most of the models available in the literature [18], the GAP model presented here is able to reproduce the relative ordering of binding energies of the nonparallel and $\langle 110 \rangle$ dumbbell di-interstitial configurations.

The GAP energetics of self-interstitial defective configurations is computed using a $10 \times 10 \times 10$ conventional cubic supercell at equilibrium volume with atomic relaxation. According to the authors of Ref. [89], the size of the cell used here is enough to guarantee consistency with calculations performed at zero pressure condition.

F. Free surfaces

Bulk-terminated surfaces can be regarded as a type of extended defect. The energy cost of creating a bulk terminated surface and its dependence on the crystallographic orientation influences the growth and equilibrium shape of the crystal during crystallization. Here we calculate the surface formation energy of four crystallographic surface orientations which are considered most relevant for bcc structures [92,93]. These are the low index surfaces $\{110\}$, $\{100\}$, and $\{112\}$, which are lowest in energy compared to other orientations, plus the $\{111\}$ surface orientation for the sake of completeness. The formation energy ordering obtained for GAP is reported in Table IV and agrees well both qualitatively and quantitatively with the DFT results.

G. Gamma surfaces and dislocations

We shall now examine the ability of the GAP potential to reproduce gamma surfaces, as introduced by Vitek [94]. Gamma surfaces, or generalized stacking faults, are two-dimensional functions describing the energy change due to a relative displacement of two halves of a crystal with respect to each other across a glide plane. Such surfaces are obtained by computing the energy associated to all possible relative shear displacement vectors spanning a given crystallographic plane. However, due to the crystal periodicity, the displacement vectors that need to be considered to fully characterize any gamma surface are bound by the lattice vectors of the crystallographic plane under consideration. Gamma surfaces provide a way for finding potential stacking faults in metals by looking at local minima in the computed energy landscape and their details have direct impact on the structure of screw dislocations core. Here we restrict our analysis to the $\{110\}$ and $\{112\}$ crystallographic planes which, due to their dense packing, are the most important slip planes in bcc metals. Gamma surface calculations are performed using slanted cells of 12 atoms with the long direction oriented perpendicular to the gamma surface. The crystal cell is distorted, without moving the atoms, in a grid of displacements in the $[1\bar{1}0]$ and $[001]$ directions for the $\{110\}$ gamma surface, and $[1\bar{1}\bar{1}]$ and $[0\bar{1}\bar{1}]$ directions for the $\{112\}$ gamma surface. Atoms are relaxed in the direction normal to the glide plane before evaluation of the total energy. GAP results are reported on the right column of Fig. 13 and

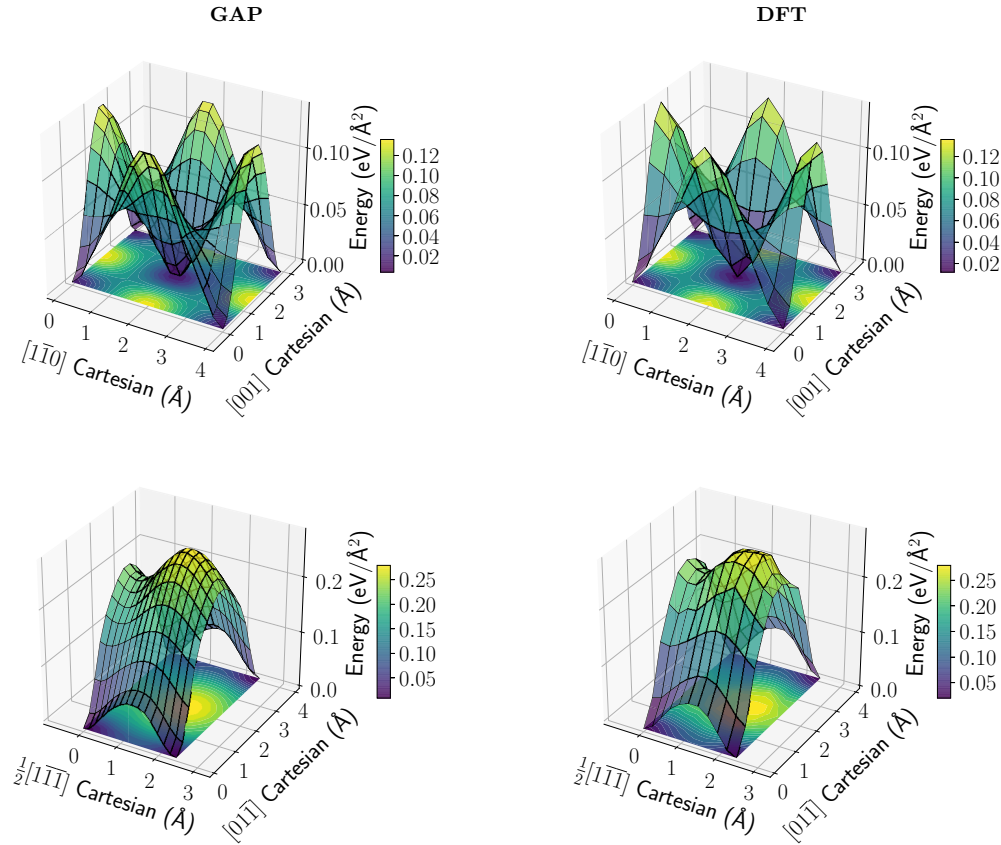


FIG. 13. Gamma surfaces computed in GAP (left column) and DFT (right column) for the $\{110\}$ (top row) and $\{211\}$ (bottom row) crystallographic orientation with atomic relaxation in the direction perpendicular to the surface.

appear in good agreement with DFT data reported on the left column of the same figure.

We proceed further with our validation process by assessing the Peierls energy barriers for a $1/2\langle 111 \rangle$ screw dislocation gliding along any of the equivalent $\langle 112 \rangle$ directions. As a first step, we determine the stable structure for the dislocation core predicted by the potential. For the simulation of dislocations with PBCs, quadrupolar arrangements of easy core $1/2\langle 111 \rangle$ are created by making a dislocation dipole in a slanted cell, which would be equivalent to a square arrangement of dislocations in a square cell [20,95]. The simulation cell lattice parameters used are

$$\vec{a} = N_x \vec{v}_{112}, \quad (5)$$

$$\vec{b} = \frac{N_x}{2} \vec{v}_{112} + N_y \vec{v}_{1\bar{1}0} + \frac{1}{2} \vec{v}_{111}, \quad (6)$$

$$\vec{c} = \vec{v}_{111}, \quad (7)$$

which is equivalent to half a cell of $N_x \vec{v}_{112} \times N_y \vec{v}_{1\bar{1}0}$, where \vec{v} are the directions in the bulk lattice, and the integer values of N_x and N_y are chosen to make the arrangement of dislocations as close to square as possible. Two cells are used in this study, one with 135 atoms (5×9), for which DFT calculations can also be performed for validation, where dislocations are separated by $\sim 17 \text{ \AA}$, and a larger cell containing 2330 atoms (21×37) where dislocations are separated by $\sim 70 \text{ \AA}$. Atoms are displaced in the z direction according to linear elastic theory

around the dislocation core positions. All atomic positions and the cell vectors are allowed to relax. The differential displacement map reported in Fig. 14 shows the screw components of the screw dislocation core structure (out of plane displacements [96]) computed with GAP. Results are in agreement with DFT [20,97,98] having a nondegenerate compact core structure with a D3 point-group symmetry. Separate plots of the in-plane edge components (magnified 20 times) show that GAP more closely matches the structure obtained with DFT than with the Mendelev potential.

The Peierls barrier is calculated by performing a NEB calculation with climbing images [81,99,100] between the initial configuration and with one dislocation moved by \vec{v}_{112} . The Peierls plot for the GAP potential shows a single saddle point in qualitative accordance with earlier DFT findings [20], whereas the Mendelev pathway has a double hump due to an incorrectly stabilized split-core structure [101]. The asymmetry in the barrier plot of Fig. 15 is due to moving only one of the dislocations in the cell so the final configuration deviates from an exact square quadrupole in the final arrangement. This finite-size effect vanishes for sufficiently large simulation boxes. The DFT reference energies for the Peierls barrier are generated by recalculating the structures obtained from the 135 atom GAP NEB with DFT (for computational efficiency reasons). The value of the Peierls barrier, 64 meV b^{-1} (where b is the Burgers vector), is in good agreement with our DFT calculations. The largest deviation from DFT calculated forces is for atoms in the dislocation core at the saddle point, and does

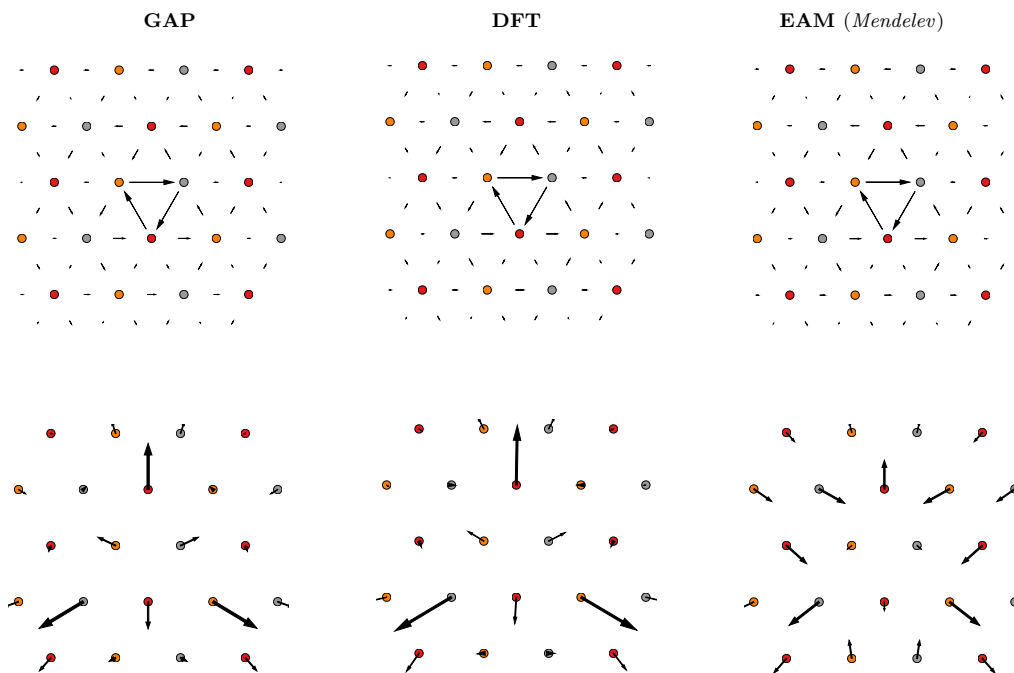


FIG. 14. Differential displacement maps of the screw dislocation core structure obtained with GAP (left column), DFT (center), and Mendeleev potential (right). The compact, nondegenerate core structure satisfying D3 point-group symmetry is consistent with earlier first-principles findings [20,97,98]. Circles of different colors represent atoms belonging to different parallel planes with the ABCABC stacking sequence of the $\langle 111 \rangle$ zone before introduction of the dislocations. The top row shows out-of-plane screw displacements and the bottom row shows the in-plane edge displacements (magnified 20 times).

not exceed $0.1 \text{ eV } \text{\AA}^{-1}$. Although the Peierls barrier seems high in comparison to barriers of $\approx 40 \text{ meV b}^{-1}$ found in the literature [20,102], the barrier itself is shown to vary by 10 to 20 meV b^{-1} for different DFT methods [101], and may also be sensitive to the method used to find the transition state, so we only make quantitative comparisons with our own calculations.

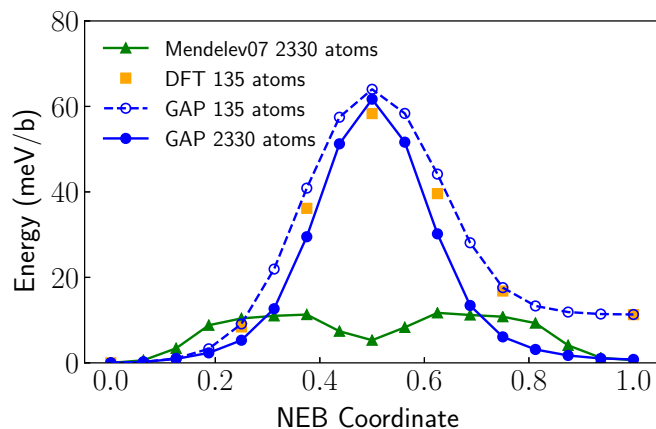


FIG. 15. Peierls energy barrier of a $1/2\langle 111 \rangle$ screw dislocation gliding in the $\langle 112 \rangle$ direction. GAP data are computed using two different cells containing 135 and 2330 atoms with two dislocations in a quadrupolar arrangement. DFT energies are computed in this work for the structures obtained from the 135 atoms GAP NEB pathway. The Mendeleev semiempirical EAM potential curve is shown for comparison using a 2330 atom cell.

V. REMARKS AND CONCLUSIONS

We have generated a Gaussian approximation potential for α -iron by training on DFT total energies, forces, and stresses for approximately 150k local atomic environments. The GAP model is presented and validated against DFT data not included in the training protocol, either computed in this work or taken from the literature. Results show that the new model is able to reproduce DFT energetics and thermodynamics with great accuracy, including energetics of point defects such as mono-, di-, and trivacancies and of self-interstitials and di-interstitials. Notably, the potential is able to reproduce the Bain bath, a positive 5-nn divacancy binding energy and the correct ordering of binding energies for the nonparallel and parallel $\langle 110 \rangle$ di-interstitials, rectifying some of the weaknesses displayed by the EAM interatomic potentials available in the literature [18]. Selected generalized stacking faults and the formation energy of selected free surfaces are also reproduced from a qualitative and quantitative point of view. The compact, nondegenerate core structure of the $1/2\langle 111 \rangle$ screw dislocation and the associated Peierls energy barrier are also consistent with DFT. In order to achieve such accuracy, we found it essential to use first-principles data with a high degree of convergence to the DFT Born-Oppenheimer PES, in particular, the k -point sampling needs to be high because supercells of different sizes cannot have a congruent Brillouin sampling, and the plane-wave cutoff is high enough that energies, forces, and virials are all converged.

We stress that, as pointed out in previous works, the model is built to interpolate between known atomic environments but does not extrapolate to completely new configurations. Caution

is therefore always suggested when dealing with such cases. In this work, we have tried to ensure transferability of the model by creating an extended training database that provides a good coverage of environments across the thermodynamic range of stability of the α -phase of iron. Such a database can be further extended in a modular way to include new environments which are relevant to a specific line of research. To this end, we have pointed out all the details of the data generation protocol needed to preserve the accuracy of the current database.

The DFT data used for the training are always performed in a collinear spin-polarized approximation starting from a ferromagnetic ordering. As such, the model can only reproduce reliable thermomechanical properties up to two-thirds of the Curie point, while the high-temperature paramagnetic behavior governed by magnetic disorder cannot be correctly captured. In order to study high-temperature bcc phases of iron, one needs to train the paramagnetic PES. In fact, however, accessing the paramagnetic PES with standard DFT calculations is a nontrivial task [11,110]. An alternative route to magnetism is to generalize the GAP formalism to treat magnetic degrees of freedom in a semiclassical way. This approach will possibly be a future direction of investigation.

The computational cost of GAP is higher than simple analytical models, at around 60 ms/atom/cpu-core, so at this

stage, the method is too expensive to tackle multimillion atom and/or nanosecond calculations using moderate computational resources. However, its linear scaling cost with respect to the number of processors, combined with its high accuracy, makes the methodology suitable to access intermediate time and size scales, which are not accessible by first principles. This paves the way, for example, to the use of interatomic potentials for studying thermodynamics of real materials with a reliability never achieved before.

The training database and the potential are freely available on the MATERIALS CLOUD ARCHIVE [111] and at www.libatoms.org. Software and data necessary for the reproduction of the results are freely available at www.libatoms.org.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support from the Swiss National Science Foundation (SNSF Project No. 200021-143636). This work was supported by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID *ch3*. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [Programme Grant No. EP/L014742/1]. This work used the ARCHER UK National Supercomputing Service (<http://www.archer.ac.uk>).

-
- [1] E. G. Moroni, G. Kresse, J. Hafner, and J. Furthmüller, *Phys. Rev. B* **56**, 15629 (1997).
- [2] D. Alfè, M. J. Gillan, and G. D. Price, *Nature (London)* **401**, 462 (1999).
- [3] A. Dal Corso and S. de Gironcoli, *Phys. Rev. B* **62**, 273 (2000).
- [4] A. Laio, S. Bernard, G. L. Chiarotti, S. Scandolo, and E. Tosatti, *Science* **287**, 1027 (2000).
- [5] A. J. Hatt, B. C. Melot, and S. Narasimhan, *Phys. Rev. B* **82**, 134418 (2010).
- [6] M. Pozzo, C. Davies, D. Gubbins, and D. Alfè, *Nature (London)* **485**, 355 (2012).
- [7] D. Dragoni, D. Ceresoli, and N. Marzari, *Phys. Rev. B* **91**, 104105 (2015).
- [8] X. Sha and R. E. Cohen, *Phys. Rev. B* **73**, 104303 (2006).
- [9] F. Körmann, A. Dick, B. Grabowski, B. Hallstedt, T. Hickel, and J. Neugebauer, *Phys. Rev. B* **78**, 033102 (2008).
- [10] F. Körmann, A. Dick, B. Grabowski, T. Hickel, and J. Neugebauer, *Phys. Rev. B* **85**, 125104 (2012).
- [11] F. Körmann, B. Grabowski, B. Dutta, T. Hickel, L. Mauger, B. Fultz, and J. Neugebauer, *Phys. Rev. Lett.* **113**, 165503 (2014).
- [12] M. S. Daw, S. M. Foiles, and M. I. Baskes, *Mater. Sci. Rep.* **9**, 251 (1993).
- [13] M. W. Finnis and J. E. Sinclair, *Philos. Mag. A* **50**, 45 (1984).
- [14] S. P. Chen, A. F. Voter, and D. J. Srolovitz, *Phys. Rev. Lett.* **57**, 1308 (1986).
- [15] F. Ercolessi, M. Parrinello, and E. Tosatti, *Surf. Sci.* **177**, 314 (1986).
- [16] M. I. Mendeleev, S. Han, D. J. Srolovitz, G. J. Ackland, D. Y. Sun, and M. Asta, *Philos. Mag.* **83**, 3977 (2003).
- [17] G. J. Ackland, M. I. Mendeleev, D. J. Srolovitz, S. Han, and A. V. Barashev, *J. Phys.: Condens. Matter* **16**, S2629 (2004).
- [18] L. Malerba, M. Marinica, N. Anento, C. Björkas, H. Nguyen, C. Domain, F. Djurabekova, P. Olsson, K. Nordlund, A. Serra, D. Terentyev, F. Willaime, and C. Becquart, *J. Nucl. Mater.* **406**, 19 (2010).
- [19] D. A. Terentyev, T. P. C. Klaver, P. Olsson, M.-C. Marinica, F. Willaime, C. Domain, and L. Malerba, *Phys. Rev. Lett.* **100**, 145503 (2008).
- [20] L. Ventelon and F. Willaime, *J. Comput.-Aided Mater. Des.* **14**, 85 (2007).
- [21] D. Dragoni, D. Ceresoli, and N. Marzari, [arXiv:1605.03334](https://arxiv.org/abs/1605.03334).
- [22] M. I. Baskes, J. S. Nelson, and A. F. Wright, *Phys. Rev. B* **40**, 6085 (1989).
- [23] M. Mrovec, D. Nguyen-Manh, C. Elsässer, and P. Gumbsch, *Phys. Rev. Lett.* **106**, 246402 (2011).
- [24] R. Drautz and D. G. Pettifor, *Phys. Rev. B* **84**, 214114 (2011).
- [25] M. E. Ford, R. Drautz, T. Hammerschmidt, and D. G. Pettifor, *Modell. Simul. Mater. Sci. Eng.* **22**, 034005 (2014).
- [26] S. Chiesa, P. M. Derlet, S. L. Dudarev, and H. V. Swygenhoven, *J. Phys.: Condens. Matter* **23**, 206001 (2011).
- [27] D. J. Hepburn and G. J. Ackland, *Phys. Rev. B* **78**, 165115 (2008).
- [28] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, *Phys. Rev. Lett.* **104**, 136403 (2010).
- [29] J. Behler and M. Parrinello, *Phys. Rev. Lett.* **98**, 146401 (2007).
- [30] A. Seko, A. Takahashi, and I. Tanaka, *Phys. Rev. B* **90**, 024101 (2014).
- [31] N. Artrith and A. Urban, *Comput. Mater. Sci.* **114**, 135 (2016).

- [32] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, *J. Phys. Chem. C* **121**, 511 (2017).
- [33] K. Miwa and H. Ohno, *Phys. Rev. B* **94**, 184109 (2016).
- [34] P. E. Dolgirev, I. A. Kruglov, and A. R. Oganov, *AIP Adv.* **6**, 085318 (2016).
- [35] A. V. Shapeev, *Multiscale Model. Simul.* **14**, 1153 (2016).
- [36] S. Faraji, S. A. Ghasemi, S. Rostami, R. Rasoulkhani, B. Schaefer, S. Goedecker, and M. Amsler, *Phys. Rev. B* **95**, 104105 (2017).
- [37] A. Glielmo, P. Sollich, and A. De Vita, *Phys. Rev. B* **95**, 214302 (2017).
- [38] C. Chen, Z. Deng, R. Tran, H. Tang, I.-H. Chu, and S. P. Ong, *Phys. Rev. Mater.* **1**, 043603 (2017).
- [39] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge, 2003).
- [40] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, MA, 2006).
- [41] A. P. Bartók, R. Kondor, and G. Csányi, *Phys. Rev. B* **87**, 184115 (2013).
- [42] W. J. Szlachta, A. P. Bartók, and G. Csányi, *Phys. Rev. B* **90**, 104108 (2014).
- [43] V. L. Deringer and G. Csányi, *Phys. Rev. B* **95**, 094203 (2017).
- [44] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, *Phys. Chem. Chem. Phys.* **18**, 13754 (2016).
- [45] N. Bernstein, A. Bartók, J. Kermode, and G. Csányi, A Gaussian Approximation Potential for Silicon, *APS Meeting Abstracts* (APS, 2017), <http://adsabs.harvard.edu/abs/2017APS..MARC25005B>.
- [46] A. P. Bartók and G. Csányi, *Int. J. Quantum Chem.* **115**, 1051 (2015).
- [47] J. Quiñonero-Candela and C. E. Rasmussen, *J. Mach. Learn. Res.* **6**, 1939 (2005).
- [48] M. W. Mahoney and P. Drineas, *Proc. Natl. Acad. Sci. USA* **106**, 697 (2009).
- [49] <http://www.github.com/libAtoms/QUIP>.
- [50] D. Dragoni, Ph.D. thesis, École Polytechnique Fédérale de Lausanne, 2016.
- [51] The simple measure of localization that we use here is the length of the perimeter connecting the vacancy sites of the perfect bidimensional lattice associated to a crystallographic plane.
- [52] H. J. Berendsen, J. v. Postma, W. F. van Gunsteren, A. DiNola, and J. Haak, *J. Chem. Phys.* **81**, 3684 (1984).
- [53] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch, *J. Phys.: Condens. Matter* **21**, 395502 (2009).
- [54] J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).
- [55] <http://www.qe-forge.org/gf/project/pslibrary>.
- [56] P. M. Derlet, D. Nguyen-Manh, and S. L. Dudarev, *Phys. Rev. B* **76**, 054107 (2007).
- [57] Ratio between cutoff on the electronic density and cutoff on the electronic wave function for the ultrasoft pseudopotential.
- [58] N. Marzari, D. Vanderbilt, A. De Vita, and M. C. Payne, *Phys. Rev. Lett.* **82**, 3296 (1999).
- [59] G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, and B. Kozinsky, *Comput. Mater. Sci.* **111**, 218 (2016).
- [60] J. R. Beeler and R. A. Johnson, *Phys. Rev.* **156**, 677 (1967).
- [61] A recent work based on a DFT+Gutzwiller approach (see Ref. [112]) has shown interesting improvements over standard DFT with respect to the agreement between experiments and theory for the description of the mechanical properties of α -iron. One might therefore try to use such approach to generate a database for GAP models with predictions closer to experiments. At the moment, however, its computational cost remains an important limitation.
- [62] This value is well within the DFT uncertainty and highlights the capabilities of our training procedure.
- [63] K. Lejaeghere, V. Van Speybroeck, G. Van Oost, and S. Cottenier, *Crit. Rev. Solid State Mater. Sci.* **39**, 1 (2014).
- [64] Z. S. Basinski, W. Hume-Rothery, and A. L. Sutton, *Proc. R. Soc. London, Ser. A* **229**, 459 (1955).
- [65] J. J. Adams, D. S. Agosta, R. Leisure, and H. Ledbetter, *J. Appl. Phys.* **100**, 113530 (2006).
- [66] D. Reith, R. Podloucky, M. Marsman, P. O. Bedolla-Velazquez, and P. Mohn, *Phys. Rev. B* **90**, 014432 (2014).
- [67] S. Schönecker, Ph.D. thesis, Technische Universität Dresden, 2011.
- [68] M. Friák, M. Šob, and V. Vitek, *Phys. Rev. B* **63**, 052405 (2001).
- [69] F. Körmann, A. Dick, T. Hickel, and J. Neugebauer, *Phys. Rev. B* **81**, 134425 (2010).
- [70] W. G. Hoover, *Phys. Rev. A* **31**, 1695 (1985).
- [71] M. Parrinello and A. Rahman, *J. Appl. Phys.* **52**, 7182 (1981).
- [72] S. Plimpton, *J. Comput. Phys.* **117**, 1 (1995).
- [73] By performing MD NPT simulations, we observe that the bcc crystal is destabilized in favor of a disordered phase between 1900 and 2000 K. However, as expected from the lack of liquid configurations in the training database, such disordered phase is not consistent with a realistic liquid.
- [74] M. Y. Lavrentiev, D. Nguyen-Manh, and S. L. Dudarev, *Phys. Rev. B* **81**, 184202 (2010).
- [75] N. Ridley and H. Stuart, *J. Phys. D: Appl. Phys.* **1**, 1291 (1968).
- [76] I. Seki and K. Nagata, *ISIJ Int.* **45**, 1789 (2005).
- [77] P. D. Desai, *J. Phys. Chem. Ref. Data* **15**, 967 (1986).
- [78] D. C. Wallace, P. H. Sidles, and G. C. Danielson, *J. Appl. Phys.* **31**, 168 (1960).
- [79] D. J. Dever, *J. Appl. Phys.* **43**, 3293 (1972).
- [80] The calculation of this quantity is performed at constant volume with relaxed atomic positions in a cubic cell containing 53/1999 atoms.
- [81] G. Henkelman and H. Jónsson, *J. Chem. Phys.* **113**, 9978 (2000).
- [82] F. Djurabekova, L. Malerba, R. Pasianot, P. Olsson, and K. Nordlund, *Philos. Mag.* **90**, 2585 (2010).
- [83] E. Hayward and C.-C. Fu, *Phys. Rev. B* **87**, 174103 (2013).
- [84] D. Kandaskalov, C. Mijoule, and D. Conntable, *J. Nucl. Mater.* **441**, 168 (2013).
- [85] Y. Wang and J. P. Perdew, *Phys. Rev. B* **44**, 13298 (1991).
- [86] C.-C. Fu, J. D. Torre, F. Willaime, J.-L. Bocquet, and A. Barbu, *Nat. Mater.* **4**, 68 (2005).
- [87] C. Becquart and C. Domain, *Nucl. Instr. Meth. Phys. Res. B* **202**, 44 (2003).
- [88] Performed at the constant equilibrium volume with atomic relaxation in a $3 \times 3 \times 3$ cubic cell.

- [89] P. Olsson, C. Domain, and J. Wallenius, *Phys. Rev. B* **75**, 014110 (2007).
- [90] C.-C. Fu, F. Willaime, and P. Ordejón, *Phys. Rev. Lett.* **92**, 175503 (2004).
- [91] C. Domain and C. S. Becquart, *Phys. Rev. B* **65**, 024103 (2001).
- [92] M. J. Spencer, A. Hung, I. K. Snook, and I. Yarovsky, *Surf. Sci.* **513**, 389 (2002).
- [93] P. Błoński and A. Kiejna, *Surf. Sci.* **601**, 123 (2007).
- [94] V. Vitek, *Philos. Mag.* **18**, 773 (1968).
- [95] M. Itakura, H. Kaburaki, and M. Yamaguchi, *Acta Mater.* **60**, 3698 (2012).
- [96] V. Vitek, *Philos. Mag.* **84**, 415 (2004).
- [97] S. L. Frederiksen and K. W. Jacobsen, *Philos. Mag.* **83**, 365 (2003).
- [98] C. Domain and G. Monnet, *Phys. Rev. Lett.* **95**, 215506 (2005).
- [99] S. R. Bahn and K. W. Jacobsen, *Comput. Sci. Eng.* **4**, 56 (2002).
- [100] E. L. Kolsbjerg, M. N. Groves, and B. Hammer, *J. Chem. Phys.* **145**, 094107 (2016).
- [101] L. Ventelon, F. Willaime, E. Clouet, and D. Rodney, *Acta Mater.* **61**, 3973 (2013).
- [102] L. Dezerald, L. Ventelon, E. Clouet, C. Denoual, D. Rodney, and F. Willaime, *Phys. Rev. B* **89**, 024104 (2014).
- [103] Y. Tateyama and T. Ohno, *Phys. Rev. B* **67**, 174105 (2003).
- [104] H. Matter, J. Winter, and W. Triftshäuser, *Appl. Phys.* **20**, 135 (1979).
- [105] L. De Schepper, D. Segers, L. Dorikens-Vanpraet, M. Dorikens, G. Knuyt, L. M. Stals, and P. Moser, *Phys. Rev. B* **27**, 5257 (1983).
- [106] A. Vehanen, P. Hautojärvi, J. Johansson, J. Yli-Kaupilla, and P. Moser, *Phys. Rev. B* **25**, 762 (1982).
- [107] S. Dudarev, *Annu. Rev. Mater. Res.* **43**, 35 (2013).
- [108] H. J. Wollenberger, in *Physical Metallurgy*, edited by R. W. Cahn and P. Haasen (North-Holland, 1996), Vol. 2.
- [109] P. Błoński and A. Kiejna, *Vacuum* **74**, 179 (2004).
- [110] B. Alling, F. Körmann, B. Grabowski, A. Glensk, I. A. Abrikosov, and J. Neugebauer, *Phys. Rev. B* **93**, 224411 (2016).
- [111] <http://archive.materialscloud.org/2017.0006/v2/>.
- [112] T. Schickling, J. Bünemann, F. Gebhard, and L. Boeri, *Phys. Rev. B* **93**, 205151 (2016).