Nuclear Instruments and Methods in Physics Research B 483 (2020) 15–21

Contents lists available at ScienceDirect

Nuclear Inst. and Methods in Physics Research, B

journal homepage: www.elsevier.com/locate/nimb

A DFT-driven multifidelity framework for constructing efficient energy models for atomic-scale simulations

Luca Messina [a,b,*], Alessio Quaglino [b], Alexandra Goryaeva [c], Mihai-Cosmin Marinica [c], Christophe Domain [d], Nicolas Castin [e], Giovanni Bonny [e], Rolf Krause [b]

[a] KTH Royal Institute of Technology, Nuclear Engineering, SE-106 91 Stockholm, Sweden
[b] Institute of Computational Science, Università della Svizzera Italiana, CH-6900 Lugano, Switzerland
[c] Université Paris-Saclay, CEA, Service de Recherches de Métallurgie Physique, F-91191 Gif-sur-Yvette, France
[d] Département Matériaux et Mécanique des Composants, EDF-R&D, Les Renardières, F-77250 Moret-sur-Loing, France
[e] SCK CEN, Nuclear Materials Science Institute, Boeretang 200, B-2400, Mol, Belgium

ARTICLE INFO

Keywords:
Machine learning
Multifidelity
Kinetic Monte Carlo
Atomistic simulations
Iron-copper alloys

ABSTRACT

The reliability of atomistic simulations depends on the quality of the underlying energy models providing the source of physical information, for instance for the calculation of migration barriers in atomistic Kinetic Monte Carlo simulations. Accurate (high-fidelity) methods are often available, but since they are usually computationally expensive, they must be replaced by less accurate (low-fidelity) models that introduce some degrees of approximation. Machine-learning techniques such as artificial neural networks can be employed to work around this limitation and extract the needed parameters from large databases of high-fidelity data. However, the latter are often computationally expensive to produce. This work introduces an alternative method based on the multifidelity approach. Correlations between high-fidelity and low-fidelity predictions are exploited to make an educated guess of the high-fidelity value based only on quick low-fidelity estimations, to be used for instance as an efficient and reliable source of physical data for atomistic simulations. With respect to neural networks, this approach requires less training data because of the lower amount of fitting parameters involved. The method is tested on the prediction of *ab initio* formation and migration energies of vacancy diffusion in iron-copper alloys, and compared with the neural networks trained on the same database.

1. Introduction

Numerical simulations play an important role in understanding and predicting the phenomena driving the microstructure evolution of solid materials. These processes cover such a wide range of length and time scales that a true multiscale approach is necessary. Typically, electronic-structure calculations based on Density Functional Theory (DFT) are used to parameterize atomic-scale simulations, which in turn open the way to higher scales thanks to various modeling techniques [1,2]. The reliability of these simulations crucially depends on the transfer of physical knowledge from the lower to the higher scale, which is often difficult to achieve given the intrinsic multifold complexity of materials at the electronic level. For this reason, several approximations are usually put in place. For instance, in atomistic kinetic Monte Carlo (AKMC) as well as in other methods it is necessary to compute the migration energy of given defects depending on the composition of the local atomic environment (LAE) [3]. These can be obtained to a high degree of accuracy with DFT and the nudged-elastic band (NEB) method [4,5], but due to the high computational cost, it is certainly impossible to cover even a small fraction of the possible combinations. More generally, the same issue arises when a so-called *energy model* is needed to predict the energy corresponding to a given atomic configuration, when the amount of possible configurations is so large that accurate calculations for all of them are out of reach.

This is a common parameterization problem that is found in many branches of science: even though the parameters can be obtained with accurate *high-fidelity* models, the associated computational cost is so high that their use is impractical. Therefore, they are replaced by *low-fidelity* models that introduce a certain degree of approximation, but can provide all needed parameters with little computational effort [6]. In atomistic simulations of microstructure evolution, DFT is the most typical high-fidelity choice, but for computational reasons it is paired

* Corresponding author at: CEA, DES, IRESNE, DEC, Cadarache F-13108 Saint-Paul-Lez-Durance, France.
E-mail address: luca.messina@cea.fr (L. Messina).

https://doi.org/10.1016/j.nimb.2020.09.011
Received 15 August 2018; Received in revised form 18 June 2020; Accepted 17 September 2020
Available online 12 October 2020
0168-583X/

with low-fidelity models that make use of a limited set of DFT data. The most common examples in AKMC simulations are pair-interaction models [7–9], cluster-expansion developments [10,11], and interatomic potentials [12,13]. Clearly, the main drawback is the loss of physical accuracy due to the involved approximations. These low-fidelity models are usually devised to correctly describe part of the physical properties, at the expense of others. Furthermore, it is not possible to employ many of them at once: for instance, for the same alloy it is common to have several interatomic potentials tackling different properties, but choosing one of them entails discarding all the others.

In this context, machine-learning (ML) algorithms can be very useful to extract data from the potentially large but for the most part inaccessible source of high-fidelity physical information, in a form that can be directly applied to the parameterization of higher-scale models [14]. It is indeed possible to create ML tools able to "learn" the physical properties from a dataset of high-fidelity examples (the *training set*), and provide accurate estimations for unknown cases. So far, this has been achieved mainly with artificial neural networks (ANN), trained on DFT data and used to construct interatomic potentials for molecular dynamics (MD) [15–17] or KMC simulations [18], or for the direct prediction of migration energies [19–21]. The latter works have proven that ANNs successfully achieve a full transfer of physical information from DFT to the higher scales, but they are also characterized by a few limitations. Firstly, many training examples are needed, so that using DFT as a source of training data is computationally demanding. In addition, the ANN is a sort of "black box" whose mathematical parameters are completely detached from the physical properties they are attempting to model. For this reason, the estimation of the target quantities can be very accurate, but there is limited room for understanding the physics behind it, and very little leveraging in case the simulation results are not correct.

This work proposes an alternative method to ANNs to achieve the same goal, i.e., extracting information from a high-fidelity database to ensure the best possible transfer of physical data across modeling scales. It relies on a *multifidelity* (MF) *approach* [6]: instead of interpolating in a large high-fidelity database, a correlation is sought between a limited high-fidelity dataset and the corresponding prediction by one or more low-fidelity models. With such a correlation, the MF model is able to provide quick "on-the-fly" estimations "mimicking" the response of the computationally expensive high-fidelity model. To the authors' knowledge, this method has never been applied to microstructure evolution simulations. With respect to ANNs, it is expected to reach comparable prediction accuracies with a lower amount of training samples, with significant savings of computational time. In addition, low-fidelity models can be combined together, rather than mutually exclude one another, and each of them contributes to improving the prediction accuracy. Since different models can target different physical properties, MF models are expected to be more flexible and "physics-aware" than ANNs.

The MF approach is tested here on a DFT database of vacancy migration barriers in FeCu alloys, that was used for ANN training in previous works [21,18]. The model is trained to predict the DFT energies and migration barriers of unknown atomic configurations and jump events, based on the estimation provided by some low-fidelity models [22,7]. If successful, this approach can pave the way towards efficient DFT-aware energy models to be used as sources of physical information in KMC simulations, as well as in a wide range of simulation methods requiring quick inexpensive estimations of formation or migration energies.

2. Method

2.1. Multifidelity framework

Multifidelity methods are a class of techniques that accelerate calculations by leveraging on the correlations between the output of accurate but computationally expensive models, and that of inexpensive but less accurate ones [6]. Mostly used in uncertainty quantification [23], they have been recently successfully applied to personalized medicine, in particular to cardiac electrophysiology [24] and cardiovascular modeling [25]. This work aims at testing a similar technique on the parameterization of atomistic microstructure-evolution simulations. In what follows, a summary of the general mathematical framework is provided. The reader is referred to [6] for a more detailed description.

In the MF context, a *model* is defined as a function $f : \mathscr{D} \rightarrow \mathscr{Y}$ that maps an input to an output, where $\mathscr{D} \subseteq \mathbb{R}^d$ is the input domain ($d \in \mathbb{N}$), and $\mathscr{Y} \subseteq \mathbb{R}$ the output domain. Let $z \in \mathscr{D}$ be the input and $y \in \mathscr{Y}$ the output. A *high-fidelity* model, denoted with $f^{(1)}$, yields an accurate approximation of the output of interest, whereas several *low-fidelity* models, denoted with $f^{(k)}$ ($k \in \mathbb{N}_{>1}$), provide less accurate estimations of the same output. Typically, since evaluating $y = f^{(1)}(z)$ is computationally expensive, it is unfeasible to sample y_i at many inputs of interest z_i. Therefore, $f^{(1)}$ is usually replaced with $f^{(k)}$ for one suitable choice of k that gives an acceptably low error. Standard strategies to generate low-fidelity models are simplified models, projection methods, and data-fit surrogates [6].

However, this approach has three main limitations. First, the low-fidelity replacement is often unable to meet the desired accuracy. Secondly, when generating a low-fidelity model $f^{(k)}$ by means of projection or data fit with a training set $(z_i, f^{(1)}(z_i))_{i=1,\dots,T}$, the amount of training samples needs to be large even for low-dimensional problems ($d \approx 15$). Finally, by selecting one low-fidelity model and discarding all the others, a large amount of information is lost. MF methods aim at overcoming such limitations by switching from model *selection* to model *fusion*, i.e., by combining all models to produce a better estimate of y. This can be done for instance via control variates [23] or Bayesian regressions [26]. This study is focused on the latter approach, in particular on its variant based on a Gaussian Process Regression (GPR) [27].

The key idea is to improve the low-fidelity estimates by relying on the statistical dependency between $f^{(k)}$ and $f^{(1)}$, instead of minimizing the errors $\left|f^{(k)}(z) - f^{(1)}(z)\right|$. Mathematically, this can be achieved by creating an output-to-output data-fit surrogate:

$$y(z) = \mathscr{G}\left[f^{(2)}(z), \dots, f^{(m+1)}(z)\right] + \varepsilon, \quad \varepsilon \sim \mathscr{N}(0, \sigma), \tag{1}$$

using a training set $[f^{(k)}(z_i), f^{(1)}(z_i)]$, with $i = 1, \dots, T$ and $k = 2, \dots, m+1$. In the GPR context, $\mathscr{G}$ is assumed to follow a GP distribution with a prescribed covariance kernel, whose parameters are fitted by maximizing a given likelihood. The underlying assumption is that the models $f^{(k)}$ contain partial information on the full model $f^{(1)}$, and that the missing information can be modeled as a normally-distributed systematic uncertainty term. Therefore, the availability of a large number of models contributes to better explain the total output variability, thus reducing the needed size of the training set. The accuracy of the fit can be thus improved by including more low-fidelity models, rather than adding more training data points. Moreover, the dimensionality and complexity of the model in Eq. (1) are typically lower than those of the input. Hence, a small training set is sufficient to achieve an accurate data fit, as opposed to the case of input-to-output surrogates (such as ANNs), which are in general more involved. Furthermore, the use of a Bayesian regression complements the estimate with confidence intervals, which can provide error indicators for refining the low-fidelity models, if necessary.

2.2. Application to vacancy migration in FeCu alloys

The multifidelity approach is tested on the modeling of single-vacancy diffusion in FeCu dilute alloys, applicable for instance to the parameterization of AKMC simulations. The high-fidelity model is a DFT database previously produced to train an ANN [21], as well as to create

fully ANN-based interatomic potentials [18]. The database consists of approximately 2000 NEB calculations [4,5] of vacancy migration in 249-atom supercells, half of them featuring a perfect solid solution with a random distribution of Cu atoms (up to 5 at.%), and the other half being snapshots of atomic configurations produced in previous AKMC simulations [22,21] and containing small Cu clusters. The 2000 cases were selected in order to obtain the most diverse database in terms of jumping atom type (Fe or Cu), local atomic environment (LAE) around the vacancy, and migration barrier values. The calculations were performed with the Vienna *ab initio* simulation package (VASP) [28–30] using the projector-augmented wave (PAW) pseudopotentials [31,32] in the Perdew-Burke-Erzernhof (PBE) approximation [33]. Further details on the *ab initio* calculations can be found in the previous ANN work [21].

The low-fidelity inputs are given by two independent energy models used in the past to investigate diffusion and precipitation of Cu in Fe alloys: an interatomic potential based on the embedded-atom method (EAM) [22], and a broken-bond (BB) pair-interaction model [7]. The latter yields the migration barrier of a given migration event as a function of the LAE of the vacancy and the jumping atom, limitedly to first- and second-nearest neighbors. On the other hand, the EAM potential is used here to provide the relaxed supercell energy in the initial and final configuration for each of the 2000 migration events, as well as to perform the corresponding NEB calculations following the same strategy and parameterization of previous works [19,20].

The multifidelity fitting is aimed at estimating the following properties:

1. The supercell formation energy E^{f}, computed from the supercell total energy E^{*} as:
$$E^{\mathrm{f}} = E^{*} - \left[E^{\mathrm{c}}_{\mathrm{Fe}} N_{\mathrm{Fe}} + E^{\mathrm{c}}_{\mathrm{Cu}} N_{\mathrm{Cu}}\right], \tag{2}$$
where $E^{\mathrm{c}}_{\mathrm{Fe}}$ and $E^{\mathrm{c}}_{\mathrm{Cu}}$ are arbitrary normalizing constants standing for the energy per atom computed in 250-atom pure bcc Fe and pure bcc Cu supercells with either DFT (-8.31 and -3.68 eV) or EAM (-4.12 and -3.49 eV).
2. The migration barriers $E^{\mathrm{mig}}_{\mathrm{i}\to\mathrm{f}}$ and $E^{\mathrm{mig}}_{\mathrm{f}\to\mathrm{i}}$, taken as the difference between the supercell energy at the saddle point and that in the initial or final state, respectively.
3. The energy difference between the final (f) and initial (i) state: $\Delta E = E_{\mathrm{f}} - E_{\mathrm{i}}$.
4. The saddle-point energy computed from the migration barrier as:
$$E^{\mathrm{sad}} = E^{\mathrm{mig}}_{\mathrm{i}\to\mathrm{f}} - \frac{\Delta E}{2}. \tag{3}$$

The rescaling in Eq. (2) is an arbitrary choice aimed at reducing the variability of the fitting target. Reformulating the migration barrier as per Eq. (3) allows for a net separation between the thermodynamic and kinetic contribution, and the consequent possibility of targeting the two aspects with different high-fidelity models, if needed [21,18]. In this specific application, such splitting is advisable because the PAW-PBE functionals are known to underestimate the solubility of Cu in Fe due to an inaccurate prediction of the solution energy [34]. Note that with the definition in Eq. (3), the saddle energy is identical for the forward and backward jumps.

For each of the target quantities (E^{f}, E^{mig}, ΔE, E^{sad}), a corresponding multifidelity model is trained by performing a Gaussian regression of the type in Eq. (1), with varying sets of low-fidelity models $f^{(k)}(\mathbf{z})$. The data are regressed with a multidimensional radial-basis function (RBF) kernel and a noise function [27].

2.3. Descriptors

In most MF applications, the low-fidelity outputs are directly correlated with the high-fidelity ones. However, the energy associated to an arrangement of atoms depends in a highly non-linear way on the position of each atom. The use of spatial descriptors is therefore essential to include information about atomic positions in the fitting. Here, a brief explanation is provided. The interested reader can find a more detailed presentation of the mathematical framework in a previous publication [35], and a more general discussion about descriptors in the work of Bartók *et al.* [36].

In this context, descriptors are symmetry-invariant mathematical representations of a crystal structure that replace the conventional Cartesian coordinates while leaving the physical properties unchanged. They allow the essential features of the LAE to be comprised into a space of lower dimensionality, known as the *descriptor space*. Each atom in the supercell is assigned a series of coefficients (one for each descriptor) based on the surrounding LAE. These coefficients establish a correlation between the atomic configuration and the corresponding physical properties (e.g., the supercell energy), allowing for some kind of regression to be performed.

This study relies exclusively on Gaussian regressions of the type in Eq. (1), and features the use of the spectral descriptor SO(4) bispectrum [36,37]. The latter was shown to be complete [38,39], i.e., able to describe LAEs uniquely, including all local symmetry operations such as translation, rotation, and reflection. Instead of using the atomic coordinates $\mathbf{r}$, the LAE around the a^{th} atom is described with a neighbor density function $\rho_a(\mathbf{r})$, corresponding to the bispectrum components of the four-dimensional (4-D) hyperspherical harmonics projected onto the $\mathscr{R}^3$-sphere $(\theta_0,\ \theta,\ \phi)$ [40]. The angular components are projected onto a spherical harmonic function defined by the polar angles θ and ϕ, and the radial component is converted into the third polar angle θ_0. The relation between polar and Cartesian coordinates is bijective.

The neighbor density function is expanded in 4-D hyperspherical harmonics $U_{jmm'}$ as follows:

$$\begin{aligned}\rho_a(\mathbf{r}) &= \sum_{b\,\in\,\mathscr{V}(a)} w_b\,\delta(\mathbf{r}-\mathbf{r}_{ab}) = \\ &= \sum_{j=0}^{\infty}\sum_{m=-j}^{j}\sum_{m'=-j}^{j} c^{a}_{jmm'}\,U_{jmm'}(\theta_0,\theta,\phi),\end{aligned} \tag{4}$$

where the sum runs over all neighbors $b \in \mathscr{V}(a)$ of the a^{th} atom within a cutoff distance R_{cut}, and w_b is an arbitrary weight associated with the chemical species of b (in this work: $R_{\mathrm{cut}} = 5\mathrm{a}_0, w_{\mathrm{Fe}} = 1$, and $w_{\mathrm{Cu}} = 2$). The power spectrum coefficients $c^{a}_{jmm'}$ are computed as the inner product between the density and the hyperspherical functions. Index j can take only positive integer or half-integer values ($j = 0, \frac{1}{2}, 1, \frac{3}{2}, \ldots$), and is limited for practical purposes to a maximal value $j_{\max} = 7/2$. The latter choice guarantees a satisfactory compromise between numerical precision and computational load.

Once the $c^{a}_{j;m'm}$ coefficients are known, the SO(4) bispectrum coefficients $B^{a}_{ll_1l_2}$ [36] associated to each atom a can be written as:

$$B^{a}_{ll_1l_2} = \sum_{m',m=-l}^{l}\ \sum_{m'_1,m_1=-l_1}^{l_1}\ \sum_{m'_2,m_2=-l_2}^{l_2} c^{a*}_{j;m'm}\,C^{ll_1l_2}_{m'm'_1m'_2}\,C^{ll_1l_2}_{mm_1m_2}\,c^{a}_{l_1;m'_1m_1}\,c^{a}_{l_2;m'_2m_2} \tag{5}$$

where $C^{ll_1l_2}_{mm_1m_2}$ are the Clebsch-Gordan (CG) coefficients. From the total set of 512 coefficients for $j_{\max} = 7/2$, the CG selection rules narrow them down to 30 coefficients only [37,41]. However, Kakarala [39] has shown that this choice is still overcomplete, and the coefficient set can be further restricted to the diagonal components ($l_1 = l_2$) only, yielding 26 independent coefficients ($B^{a}_1, B^{a}_2, \ldots, B^{a}_{26}$) for each atom [41–43]. This strategy, originally suggested by Varshalovich *et al.* [40], is implemented in the MiLaDy package [35,44] used here to compute the

descriptor coefficients.

Finally, for each component i the corresponding coefficients of all atoms are summed to obtain a global coefficient $D_i = \sum_a B_i^a$ (with $i = 1, \ldots, 26$). The latter is used to seek a correlation with the supercell energy, and can be thus regarded as a low-fidelity energy estimation that contains information about mutual atomic interactions. This procedure hence yields 26 low-fidelity guesses that are added to the low-fidelity models (EAM and BB) mentioned in the previous section. For the targeted quantities obtained as energy differences or sums (E^{mig}, ΔE, E^{sad}), the same transformations are applied to each D_i. For instance, the high-fidelity ΔE is correlated to $D_i^{\mathrm{f}} - D_i^{\mathrm{i}}$, and by analogy for the other quantities. This is to avoid uncertainty propagation that would occur when computing ΔE from the formation energies.

Building additional low-fidelity models with descriptors requires the availability of some low-fidelity estimations of the atomic positions. Interatomic potentials such as the EAM model do provide this information, as they allow for relaxation calculations, whereas rigid-lattice schemes such as the BB model do not. However, positions corresponding to the rigid lattice can be taken as a rough low-fidelity guess of the atomic coordinates. This is tested in Fit E, among the fitting sessions listed in Table 1.

2.4. Fitting sessions

Several fitting sessions are performed, first with the energy guesses from the EAM potential and the BB model (E_{EAM}, E_{BB}), and then including the descriptor coefficients computed with the rigid-lattice or the EAM-relaxed atomic coordinates, respectively $D_{i=1,\ldots,n}(x_{\mathrm{rigid}})$ and $D_{i=1,\ldots,n}(x_{\mathrm{EAM}})$. The sessions are summarized in Table 1. Sessions A and B are aimed at verifying whether the descriptor coefficients are effectively correlated with the supercell energy. The training data points are picked in a complete random fashion among the 2000 NEB cases, or twice as many for E^{f} and E^{mig}, while the remaining ones are used for validation.

3. Results

3.1. Descriptor performance

If the descriptor coefficients correlate well with the target quantities, and in particular with the supercell formation energy, they can provide valuable additional low-fidelity estimations. This correlation is checked by performing a multi-dimensional Gaussian regression on the 26 descriptor coefficients computed on the relaxed atomic positions, using a varying amount of training data points (N_{train}). The accuracy of the fitting is then validated on the remaining data. The validation results are shown in Table 1 for all target quantities with $N_{\mathrm{train}} = 300$, and in Fig. 1 for the supercell formation energy with $N_{\mathrm{train}} = 100$. In Fit A (B), the descriptor coefficients are computed based on the EAM-(DFT-)relaxed atomic positions, and are correlated with the EAM (DFT) energy.

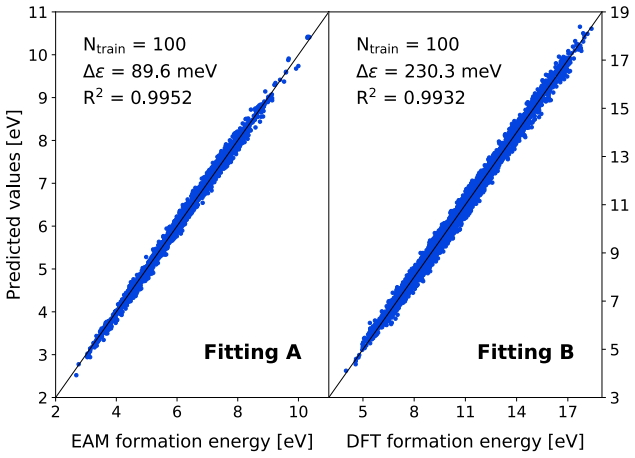

Fig. 1. Validation results of the fitting session A (B), where the descriptor coefficients computed with the EAM-relaxed (DFT-relaxed) atomic positions are used to predict the EAM (DFT) supercell formation energy, based on a training set of 100 data points. $\Delta\varepsilon$ is the mean validation error.

Fig. 1 shows that the predicted formation energies correlate well with the validation data in both fitting sessions, and this correlation is visible already at low N_{train}. A weaker correlation is found for the other target quantities in Table 1. This confirms that the descriptor coefficients can be used as source of low-fidelity data, which improves the MF fitting shown in the next section. However, the mean validation errors are higher than those obtained with ANNs. The ANN-based rigid-lattice potential [18] trained on 1300 data points reached an accuracy on the formation energy of 0.191 meV/atom, or 48 meV per supercell, while in Fit B the mean error with $N_{\mathrm{train}} = 1300$ (not shown) is 156 meV, and just slightly higher (195 meV) with $N_{\mathrm{train}} = 300$. This accuracy has been reached with a limited amount of training data, and might be reduced even further with additional refinements in the descriptor calculation procedure. The prediction of migration barriers is more accurate (55 meV with $N_{\mathrm{train}} = 1300$, and 67 meV with $N_{\mathrm{train}} = 300$), but still larger than the ANN performance (23.5 meV) [21]. In Fit A, the accuracy is considerably better thanks to the intrinsic simplicity of the EAM model with respect to DFT, so that it is easier for descriptors to predict its outcome.

Table 1
Validation results of the fitting sessions with Gaussian regression performed in this work, with a training set of 300 samples, and different low-fidelity sources. E_{EAM} and E_{BB} denote the energy computed with the EAM potential [22] or the broken-bond model [7], respectively, while $D_i(x)$ marks the 26 descriptor coefficients based on either the rigid positions (x_{rigid}) or the EAM-relaxed positions (x_{EAM}). $\Delta\varepsilon$ is the mean validation error. Session D was not performed on E^{f} because the BB model does not predict the formation energy.

Fit	High-fidelity	Low-fidelity	Formation energy (E^{f})		Migration barriers (E^{mig})		Energy difference (ΔE)		Saddle energy (E^{sad})	
			$\Delta\varepsilon$ [meV]	R^2	$\Delta\varepsilon$ [meV]	R^2	$\Delta\varepsilon$ [meV]	R^2	$\Delta\varepsilon$ [meV]	R^2
Descriptor performance										
A	E_{EAM}	$D_i(x_{\mathrm{EAM}})$	57.6	0.9980	34.3	0.9312	48.1	0.8951	19.1	0.9704
B	E_{DFT}	$D_i(x_{\mathrm{DFT}})$	194.8	0.9951	67.4	0.8902	113.8	0.9031	43.4	0.7066
Multifidelity fitting										
C	E_{DFT}	E_{EAM}	941.6	0.8675	97.3	0.5610	102.7	0.8231	62.2	0.4416
D	E_{DFT}	$E_{\mathrm{EAM}}, E_{\mathrm{BB}}$	–	–	72.5	0.7624	74.1	0.9065	53.5	0.5864
E	E_{DFT}	$D_i(x_{\mathrm{rigid}})$	146.5	0.9972	56.0	0.9249	111.1	0.9087	42.0	0.7242
F	E_{DFT}	$D_i(x_{\mathrm{EAM}})$	132.3	0.9974	59.6	0.8387	62.0	0.9348	40.6	0.7636
G	E_{DFT}	$E_{\mathrm{EAM}}, E_{\mathrm{BB}}, D_i(x_{\mathrm{rigid}}), D_i(x_{\mathrm{EAM}})$	110.4	0.9982	51.6	0.8777	58.0	0.9433	37.2	0.8008

3.2. Multifidelity fitting

The MF framework described in Section 2 is applied to the prediction of the DFT target quantities with an increasing number of low-fidelity models: first with energy guesses only (Fit C and D), then with the descriptor coefficients based on rigid-lattice positions (Fit E) and the EAM-relaxed positions (Fit F). Finally, all low-fidelity inputs are used in the last fitting (Fit G). Table 1 shows the validation results with $N_{\text{train}} = 300$, while Fig. 2 visually depicts those obtained in Fit G. The variation of the mean validation error with increasing number of training data points is shown in Fig. 3.

Without descriptors, the performance of the MF model trained only on energy estimations (Fit C and D) is very poor, with a ≈ 1 eV mean error on the formation energy, and low R-squared values on the other quantities. The addition of the broken-bond model, which targets saddle-point interactions and is therefore expected to be more accurate on migration energies than the EAM model, improves the correlation and reduces the mean error, proving that the MF approach can successfully take advantage of several models at once. However, Fig. 3 shows that increasing N_{train} does not lead to any gain of accuracy, and the error on the migration barriers is never lower than 0.07 eV even with $N_{\text{train}} = 1300$. The results without descriptors are thus fully unsatisfactory, and cannot be improved by adding more training examples, because the model is missing any information about the influence of atomic positions on the formation and migration energies.

Including the descriptor coefficients in the low-fidelity input improves drastically the formation energy prediction, and moderately the migration barriers and the other quantities, with much better R-squared values. This confirms once more the strength of the MF approach, where the contributions of several low-fidelity models targeting different properties can be joined together to provide more accurate predictions. Fig. 3 shows indeed that the accuracy can be increased by either increasing the amount of high-fidelity training data, as was the case for ANNs, or by including (or devising) additional low-fidelity models. Furthermore, Fig. 3 shows that a fair level of accuracy is already reached with 200–300 training data points, and increasing the training set does not lead to substantial improvements. This means that potentially, if a sufficient amount of low-fidelity models is available, the computational cost to produce high-fidelity data can be considerably reduced to at least 1/4 and up to 1/6 of the amount required by ANNs.

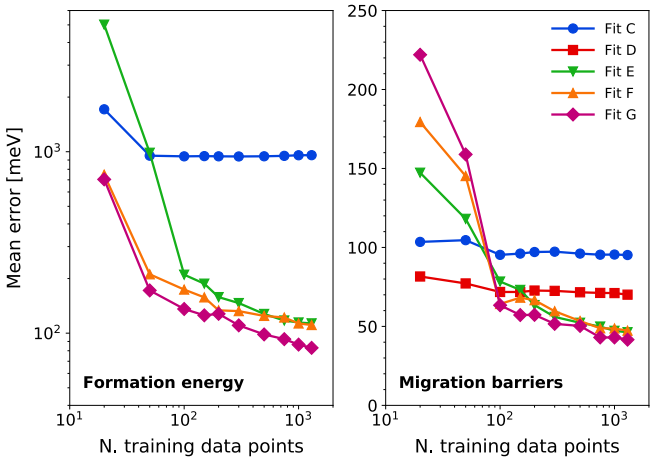

Fig. 3. Mean validation error $\Delta\varepsilon$ as a function of the number of training data points, for the prediction of the formation energy (left) and the migration barriers (right).

4. Discussion

The results of this work prove that the MF framework allows for a higher degree of flexibility (by choosing or developing the needed low-fidelity models), and a likely substantial reduction of the computational requirements with respect to ANNs. If compared to the ANN accuracy with the same amount of training data ($N_{\text{train}} = 1300$), the MF accuracy is not as good (83 vs 48 meV on the formation energy, and 41.6 vs 23.5 meV on the migration barriers). However, the MF predictions are already quite accurate with much fewer data points ($N_{\text{train}} = 300$). The ANN performance with as many data points has not been tested in the corresponding previous works [21,18], but is likely to be rather poor because of the high number of ANN parameters to be fitted. Since the MF accuracy could be improved by adding more low fidelity models, it is expected that, for systems for which several interatomic potentials exist, the MF performance should be closer to the ANN one. It is worth mentioning that the comparison at $N_{\text{train}} = 1300$ between MF and ANN refers to the amount of DFT NEB calculations needed in the training dataset, although in practice each NEB calculations in the ANN provided 12 data points, corresponding to the 6 symmetric replica of the forward and backward jump configuration [21].

Once fitted on the high-fidelity training data, the MF model can be used as energy model in atomic-scale simulations as follows. Each time the simulation requires as input the energy of a given never-seen configuration, for instance after an AKMC step, the MF model provides an estimate of this energy, in place of a computationally unfeasible on-the-fly high-fidelity (DFT) calculation. The only input parameters needed for this estimate are the atomic positions of that configuration,

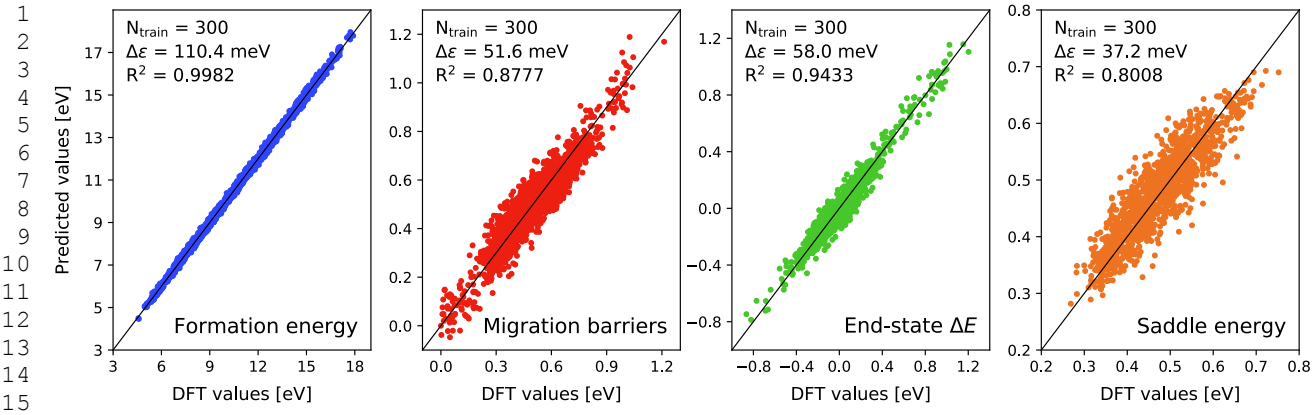

Fig. 2. Validation results of the fitting session G, based on a training set of 300 data points, to predict the DFT target quantities (formation energy, migration barriers, energy difference, and saddle energy) with all available low-fidelity models: the EAM energy [22], the broken-bond model energy [7], and the descriptor coefficients obtained with the rigid-lattice and the EAM-relaxed atomic positions. $\Delta\varepsilon$ is the mean validation error.

used for the on-the-fly computation of descriptors and low-fidelity (EAM, BB) energies. The MF model, trained to reproduce the correlation between the atomic positions and the DFT energies to a certain degree of accuracy, relies on low-fidelity models to improve its guesses. The impact on the computational efficiency of the atomic-scale simulation depends on the computational overhead introduced by low-fidelity and descriptor calculations. Specifically, the time required for a single descriptor calculation with the MiLaDy code [35], comparable to that of an EAM energy evaluation (about a few milliseconds [42]), guarantees the applicability of the MF method to KMC simulations. The MF approach can be applied as well, with no computational issues, to analytical models requiring the knowledge of configuration energies and jump frequencies, for instance for the mean-field calculation of transport coefficients by the self-consistent mean-field (SCMF) method [45] or the Green-function approach [46]. In particular, the coupling with the SCMF-based KineCluE code [47] will be the target of future work.

Thanks to the general character of the approach, the MF method is certainly suitable for developing lattice-free potentials, for instance by extending the fitting to interatomic forces. It is also worth noting that in the specific case of FeCu alloys, the thermodynamic properties stemming out of the EAM potential should be more accurate than the DFT PAW-PBE functionals, but thanks to the splitting of the migration energy according to Eq. (3), it could be possible to use ΔE_{EAM} as a high-fidelity reference for the thermodynamic part, and $E_{\mathrm{DFT}}^{\mathrm{sad}}$ for the kinetic part.

5. Conclusions

This work presented the multifidelity (MF) method as an innovative machine-learning approach for constructing energy models that can predict the static energy of a given arrangement of atoms, as well as the migration barriers of given jump events. The aim is to provide the relevant parameterization to atomic-scale simulations or analytical models, while ensuring the best possible transfer of physical properties from *ab initio* calculations. MF makes use of several approximate (low-fidelity) models to learn predicting the outcome of an accurate (high-fidelity) model without the need for performing the actual time-consuming calculation. The method, tested on the prediction of DFT formation energies and migration barriers in dilute FeCu supercells, proved to be successful in describing the correlation between the high-fidelity and the low-fidelity values. Although the performance in terms of prediction accuracy is not as good as artificial neural networks (ANN), the MF mean prediction errors are already reasonably low with much smaller training datasets than ANNs. This shows that potentially the amount of high-fidelity data needed for training can be considerably reduced, with important savings of computational power. Furthermore, it has been shown that the more low-fidelity models, the more accurate the results. MF methods are indeed capable of combining together the information given by several models, which ensures a high degree of flexibility, while keeping the needed amount of high-fidelity calculations to its minimum. In this context, the use of descriptors has been necessary to describe the influence of the local atomic environment on the energetics. In conclusion, this work has provided an example of applicability of MF methods to multiscale materials modeling and microstructure evolution simulations. They can contribute to improving the reliability of atomic-scale simulations by maximizing the amount of accurate DFT properties that can be transferred, with little approximation, to the higher modeling scales.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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 under grant agreement No 633053. A. Goryaeva acknowledges the financial support by the Cross-Disciplinary Program on Numerical Simulations of CEA, the French Alternative Energies and Atomic Energy Commission. The high-performance computational resources have been provided by Électricité de France (R&D). SCK CEN acknowledges financial support from FOD for fusion R&D.

References

[1] C.S. Becquart, C. Domain, Modeling microstructure and irradiation effects, Metall. Mater. Trans. A 42 (4) (2010) 852–870.
[2] J. Knaster, A. Moeslang, T. Muroga, Materials research for fusion, Nat. Phys. (2016) 424–434.
[3] C.S. Becquart, C. Domain, Introducing chemistry in atomistic kinetic Monte Carlo simulations of Fe alloys under irradiation, Phys. Status Solidi B 247 (1) (2010) 9–22.
[4] G. Henkelman, B.P. Uberuaga, H. Jónsson, A climbing image nudged elastic band method for finding saddle points and minimum energy paths, J. Chem. Phys. 113 (22) (2000) 9901–9904.
[5] G. Henkelman, H. Jónsson, Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points, J. Chem. Phys. 113 (22) (2000) 9978–9985.
[6] B. Peherstorfer, K. Willcox, M. Gunzburger, Survey of multifidelity methods in uncertainty propagation, inference, and optimization, SIAM Rev. 60 (3) (2018) 550–591.
[7] F. Soisson, C.-C. Fu, Cu-precipitation kinetics in α-Fe from atomistic simulations: vacancy-trapping effects and Cu-cluster mobility, Phys. Rev. B 76 (21) (2007), 214102 .
[8] R. Ngayam-Happy, C.S. Becquart, C. Domain, First principle-based AKMC modelling of the formation and medium-term evolution of point defect and solute-rich clusters in a neutron irradiated complex Fe–CuMnNiSiP alloy representative of reactor pressure vessel steels, J. Nucl. Mater. 440 (1–3) (2013) 143–152.
[9] C.-H. Huang, L. Gharaee, Y. Zhao, P. Erhart, J. Marian, Mechanism of nucleation and incipient growth of Re clusters in irradiated W-Re alloys from kinetic Monte Carlo simulations, Phys. Rev. B 96 (9) (2017) 1514–1518.
[10] M.Y. Lavrentiev, D. Nguyen-Manh, S.L. Dudarev, Cluster expansion models for Fe–Cr alloys, the prototype materials for a fusion power plant, Comput. Mater. Sci. 49 (S) (2010) S199–S203.
[11] J.S. Wrobel, D. Nguyen-Manh, K.J. Kurzydłowski, S.L. Dudarev, A first-principles model for anomalous segregation in dilute ternary tungsten-rhenium-vacancy alloys, J. Phys.-Condens. Matter 29 (14) (2017) 145403–145416.
[12] C. Pareige, M. Roussel, S. Novy, V. Kuksenko, P. Olsson, C. Domain, P. Pareige, Kinetic study of phase transformation in a highly concentrated Fe–Cr alloy: Monte Carlo simulation versus experiments, Acta Mater. 59 (6) (2011) 2404–2411.
[13] G. Bonny, N. Castin, J. Bullens, A. Bakaev, T.P.C. Klaver, D. Terentyev, On the mobility of vacancy clusters in reduced activation steels: an atomistic study in the Fe–Cr–W model alloy, J. Phys.-Condens. Matter 25 (2013), 315401 .
[14] S. Curtarolo, G.L.W. Hart, M.B. Nardelli, N. Mingo, S. Sanvito, O. Levy, The high-throughput highway to computational materials design, Nat. Mater. 12 (3) (2013) 191–201.
[15] N. Artrith, J. Behler, High-dimensional neural network potentials for metal surfaces: a prototype study for copper, Phys. Rev. B 85 (4) (2012), 045439.
[16] N. Artrith, B. Hiller, J. Behler, Neural network potentials for metals and oxides – first applications to copper clusters at zinc oxide, Phys. Status Solidi B 250 (6) (2013) 1191–1203.
[17] J. Behler, Representing potential energy surfaces by high-dimensional neural network potentials, J. Phys.-Condens. Matter 26 (18) (2014), 183001 .
[18] N. Castin, L. Messina, C. Domain, R.C. Pasianot, P. Olsson, Improved atomistic Monte Carlo models based on ab-initio-trained neural networks: application to FeCu and FeCr alloys, Phys. Rev. B 95 (21) (2017), 214117 .
[19] N. Castin, M.I. Pascuet, L. Malerba, Modeling the first stages of Cu precipitation in α-Fe using a hybrid atomistic kinetic Monte Carlo approach, J. Chem. Phys. 135 (6) (2011), 064502 .
[20] N. Castin, M.I. Pascuet, L. Malerba, Mobility and stability of large vacancy and vacancy-copper clusters in iron: an atomistic kinetic Monte Carlo study, J. Nucl. Mater. 429 (1–3) (2012) 315–324.
[21] L. Messina, N. Castin, C. Domain, P. Olsson, Introducing ab initio based neural networks for transition-rate prediction in kinetic Monte Carlo simulations, Phys. Rev. B 95 (6) (2017), 064112 .
[22] R.C. Pasianot, L. Malerba, Interatomic potentials consistent with thermodynamics: the Fe–Cu system, J. Nucl. Mater. 360 (2) (2007) 118–127.
[23] B. Peherstorfer, K. Willcox, M. Gunzburger, Optimal model management for multifidelity monte carlo estimation, SIAM J. Sci. Comput. 38 (5) (2016) A3163–A3194.
[24] A. Quaglino, S. Pezzuto, P.-S. Koutsourelakis, A. Auricchio, R. Krause, Fast uncertainty quantification of activation sequences in patient-specific cardiac

electrophysiology meeting clinical time constraints, Int. J. Numer. Meth. Biomed. Eng. 34 (7) (2018), e2985 .
[25] J. Biehler, W. Wall, The impact of personalized probabilistic wall thickness models on peak wall stress in abdominal aortic aneurysms, Int. J. Numer. Meth. Biomed. Eng. 34 (2) (2018), e2922 .
[26] P.-S. Koutsourelakis, Accurate uncertainty quantification using inaccurate computational models, SIAM J. Sci. Comput. 31 (5) (2009) 3274–3300.
[27] C. Rasmussen, C. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.
[28] G. Kresse, J. Hafner, Ab initio molecular-dynamics for liquid-metals, Phys. Rev. B 47 (1) (1993) 558–561.
[29] G. Kresse, J. Hafner, Ab-initio molecular-dynamics simulation of the liquid-metal amorphous-semiconductor transition in germanium, Phys. Rev. B 49 (20) (1994) 14251–14269.
[30] G. Kresse, J. Hafner, Norm-conserving and ultrasoft pseudopotentials for first-row and transition-elements, J. Phys.-Condens. Matter 6 (40) (1994) 8245–8257.
[31] P.E. Blöchl, Projector augmented-wave method, Phys. Rev. B 50 (24) (1994) 17953–17979.
[32] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B 59 (3) (1999) 1758–1775.
[33] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (18) (1996) 3865–3868.
[34] P. Olsson, T.P.C. Klaver, C. Domain, Ab initio study of solute transition-metal interactions with point defects in bcc Fe, Phys. Rev. B 81 (5) (2010), 054102 .
[35] A.M. Goryaeva, J.-B. Maillet, M.-C. Marinica, Towards better efficiency of interatomic linear machine learning potentials, Comput. Mater. Sci. 166 (2019) 200–209.
[36] A.P. Bartók, R. Kondor, G. Csányi, On representing chemical environments, Phys. Rev. B 87 (2013), 184115 .
[37] A.P. Bartók, Gaussian approximation potential: an interatomic potential derived from first principles quantum mechanics (Ph.D. thesis), University of Cambridge, 2009.
[38] R. Kondor, A novel set of rotationally and translationally invariant features for images based on the non-commutative bispectrum, eprint arXiv:cs/0701127 (2007).
[39] R. Kakarala, The bispectrum as a source of phase-sensitive invariants for fourier descriptors: a group-theoretic approach (Ph.D. thesis), Irvine University, 1992.
[40] D.A. Varshalovich, A.N. Moskalev, V.K. Khersonskii, Quantum Theory of Angular Momentum, Singapore World Scientific, 1988.
[41] A. Thompson, L. Swiler, C. Trott, S. Foiles, G. Tucker, Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials, J. Comput. Phys. 285 (2015) 316–330.
[42] M.A. Wood, A.P. Thompson, Quantum-accurate molecular dynamics potential for tungsten, eprint arXiv:1702.07042 (2017).
[43] M.A. Wood, A.P. Thompson, Extending the accuracy of the snap interatomic potential form, J. Chem. Phys. 148 (24) (2018), 241721 .
[44] A.M. Goryaeva, W. Unn-Toc, M.C. Marinica, MiLaDy - Machine Learning Dynamics, CEA, Saclay, 2015-2018.
[45] M. Nastar, A mean field theory for diffusion in a dilute multi-component alloy: a new model for the effect of solutes on self-diffusion, Philos. Mag. 85 (32) (2005) 3767–3794.
[46] D.R. Trinkle, Automatic numerical evaluation of vacancy-mediated transport for arbitrary crystals: onsager coefficients in the dilute limit using a Green function approach, Philos. Mag. 97 (28) (2017) 2514–2563.
[47] T. Schuler, L. Messina, M. Nastar, KineCluE: a kinetic cluster expansion code to compute transport coefficients beyond the dilute limit, Comput. Mater. Sci. 172 (2020), 109191.