Machine learning guided prediction of solute segregation at coherent and semi-coherent metal/oxide interfaces

Lu, Yizhou
Choudhury, Samrat
Uberuaga, Blas P.

To be published in: Computational Materials Science

DOI to publisher's version: 10.1016/j.commatsci.2025.114480

Computational Materials Science 264 (2026) 114480

Contents lists available at ScienceDirect

Computational Materials Science

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

Machine learning guided prediction of solute segregation at coherent and semi-coherent metal/oxide interfaces

Yizhou Lu [a], Blas Pedro Uberuaga [b], Samrat Choudhury [a,*]

[a] Department of Mechanical Engineering, University of Mississippi, University, MS 38677, USA
[b] Materials Science and Technology Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

ARTICLE INFO

Keywords:
Metal/oxide interfaces
Solute segregation
Density functional theory
Machine learning

ABSTRACT

Investigation of semi-coherent metal/oxide interfaces with misfit dislocations using density functional theory (DFT) is computationally intensive to the point of being prohibitive, as it involves several hundreds to many thousands of atoms. In this study, we examined the solute segregation behavior at the Fe/Y_2O_3 interface—a model interface for cladding applications in nuclear fission reactors—using a combination of DFT calculations and machine learning (ML) approaches. Both coherent and semi-coherent interfaces were considered. ML models were trained on DFT-calculated segregation energies to identify the key chemical, geometric and strain energy related features that govern solute segregation behavior at coherent Fe/Y_2O_3 interfaces. Furthermore, it was found that ML models when trained on DFT calculated segregation energy of elements at a coherent interface, comprising of about a hundred-atom supercell, can predict the segregation energy of elements at a semi-coherent Fe/Y_2O_3 interface (with multiple hundreds of atoms) at a fraction of computational cost (1/35th), with an accuracy comparable to DFT calculations.

1. Introduction

Materials systems with metal/oxide interfaces are used in many advanced applications ranging from nanoelectronics [1,2], catalysis [3–5], aerospace [6] to nuclear reactors [7,8]. These applications leverage the distinct atomic structure and chemistry at the interface; both these factors depend on the coherency at the metal/oxide interface. For the case in which oxide particles are embedded within a metallic matrix, the metal/oxide interface may transition from coherent to semi-coherent as the oxide particle grows [9]. While the atoms are in registry across a coherent interface, semi-coherent interfaces comprise of regions at the interface with atom-by-atom matching across the interface (coherent regions) which are separated by a region of least coincidence (misfit dislocations) [10].

Solutes are often added to the metallic matrix to improve various properties, such as adhesion of the metal/oxide interface [11] and corrosion resistance [12]. These solutes may segregate to metal/oxide interfaces thus altering the atomic structure, chemistry and properties of the interface. The presence of misfit dislocations is known to affect the solute segregation behavior at semi-coherent metal/oxide hetero-interfaces. For example, it was previously reported [13] that the local oxygen environment varies significantly between the misfit dislocations and the coherent terraces at a semi-coherent metal/oxide interface. Such non-uniform interfacial oxygen content governs the spatially varying segregation tendency of solutes to a metal/oxide interface. Similarly, the combination of the in-plane variation of the strain field and the variable chemistry surrounding a misfit dislocation can lead to drastically different (up to 50 times) concentration of solutes between the terrace and the misfit region in a metal/ceramic interface [14]. Hence, an ability to predict solute segregation tendencies to metal/oxide interfaces is critical in designing novel metal/oxide interface with optimum properties.

Traditional computational tools such as molecular dynamics (MD) have widely been used for investigating solute segregation tendency to other interfaces [15]. However, the application of MD to investigate solute segregation behavior to a metal/oxide interface is limited due to the lack of suitable interatomic potentials. Another widely used method, density functional theory (DFT), an ab initio approach grounded in quantum mechanics, is limited by system size due to the scaling of the computational expenses with the cube of the number of electrons in the system. Except for a few studies, semi-coherent interfaces have not been extensively considered in DFT calculations, as they require hundreds to

* Corresponding author.
E-mail addresses: luyizhou727@gmail.com (Y. Lu), blas@lanl.gov (B.P. Uberuaga), schoudhu@olemiss.edu (S. Choudhury).

https://doi.org/10.1016/j.commatsci.2025.114480
Received 19 September 2025; Received in revised form 27 November 2025; Accepted 29 December 2025
Available online 12 January 2026

thousands of atoms to describe misfit dislocation structures [10,13,16].

Recent reports have demonstrated that modern machine learning (ML) tools in combination with DFT calculations can be used to investigate structure-chemistry-property linkage in a wide range of hetero-interfaces including metal/oxide interface. For example, metal/oxide heterojunction artificial synapses for visual sensory were studied through experiments and artificial neural networks (ANNs) [17]. In our previous published work [18], we have shown that ML model trained on DFT calculated segregation energy for one semi-coherent interface (with 345 atoms) with one orientation relationship (OR) could accurately predict the solute segregation energy on a larger interface (with 578 atoms) with a different OR, at a computational cost 1/45 of similar DFT calculations with an accuracy comparable to DFT calculations. However, it was observed that to train a ML model with the DFT-calculated segregation energy of 28 solutes at a single interfacial site on a metal/oxide semi-coherent interface with 345 atoms required 241.66 (hrs) $\times$ 144 (cpus) $= 3.48 \times 10^4$ cpu-hours, which is highly computationally expensive, requiring high-performance computing facilities for data generation. A question arises whether it is possible to train the ML model at a considerably less computational cost such that the training data can be generated using a workstation. In this regard a possible alternative would be to generate data using smaller supercells such as a fully coherent interface. In this work we have demonstrated that an ML model can be trained with segregation data from a coherent metal/oxide interface (~100 atoms) to predict the segregation energy of at a semi-coherent interface with high fidelity at a fraction of the computational cost. We have used solute segregation behavior within the Fe/Y_2O_3 system, a surrogate metal/oxide interface observed in Nanostructured Ferritic Alloys (NFAs). NFAs are promising cladding materials for fission reactors due to their exceptional tolerance to harsh operating conditions [7].

2. Methodology

2.1. Electronic structure calculation

2.1.1. Fe/Y_2O_3 interface supercells

The solute segregation behavior of different solutes was determined using the electronic structure calculation. To perform the calculations, one Fe/Y_2O_3 semi-coherent interface supercell and two Fe/Y_2O_3 coherent interface with different strain conditions (coherent interface A and B) were created. All the three interfaces share the same experimentally observed OR: $[001]_{Fe}||[100]_{Y2O3}$ and $(010)_{Fe}||(011)_{Y2O3}$ [19]. The semi-coherent interface has a strain condition at the Y_2O_3 slab: −1.1 % in *x*-direction, and − 4.67 % in *y*-direction [13]. For coherent interface A, the strain at the Y_2O_3 slab is in line with the semi-coherent interface, while the Fe slab is strained to −8.50 % in the *x*-direction and 24.73 % in the *y*-direction to match with the oxide. To compare the ML models trained by coherent interfaces with different strain conditions, coherent interface B is similar to coherent interface A except the strain condition differs: −9.50 % in the *x*-direction and 18.73 % in the *y*-direction for the Fe slab; −2.18 % in the *x*-direction and − 9.25 % in the *y*-direction for the Y_2O_3 slab.

The relaxed structures for the semi-coherent interface and coherent interface A are presented in Fig. 1. For the semi-coherent interface (Fig. 1(a)), the Fe slab has 165 atoms with 55 Fe atoms in each layer while the Y_2O_3 slab contains 180 atoms, with each layer consisting of 24 Y atoms and 36 O atoms. At the semi-coherent interface, a misfit dislocation is observed along the *y*-direction, as 6 columns of Fe atoms correspond to 5 columns of O atoms. A misfit dislocation is also present along the *x*-direction where 11 columns of Fe atoms align with 12 columns of O atoms. In coherent interface A (Fig. 1(b)) and B (not shown), the Fe slab is composed of 48 atoms, arranged in layers containing 16 Fe atoms each. The Y_2O_3 slab includes 60 atoms, with each layer comprising 8 Y atoms and 12 O atoms. Atomic position of the topmost Fe layer and bottom most layer of yttria in all the interface structures were fixed to mimic bulk-like conditions.

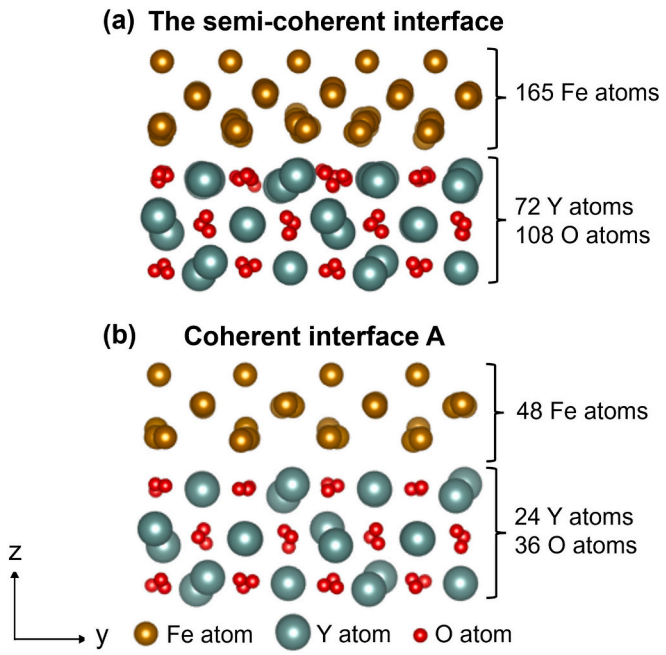

Fig. 1. Relaxed structures of Fe/Y_2O_3 interfaces, both with the OR $[001]_{Fe}||[100]_{Y2O3}$ and $(010)_{Fe}||(011)_{Y2O3}$: (a) the semi-coherent interface; (b) coherent interface A. In (a) and (b), the *x*-direction is perpendicular to the *y*- and *z*- directions.

2.1.2. Computational details for DFT calculations

All the electronic structure calculations were performed by DFT as implemented in the Vienna ab-initio Simulation Package (VASP) [13]. The projector augmented wave (PAW) method was utilized with a plane wave cut-off of 450 eV [20,21]. The Perdew–Burke–Ernzerhof (PBE) parameterization of the generalized gradient approximation (GGA) for the exchange–correlation potential was applied to all calculations with spin polarization [22,23]. Monkhorst-Pack scheme with a $1 \times 1 \times 1$ k-point mesh was used for the Brillouin-zone sampling. In all calculations, structures were relaxed until the maximum force on each atom reached 0.05 eV/Å or less.

2.1.3. Solute segregation details

The segregation behavior was evaluated by the solute segregation energy (E^{Seg}) [13]:

$$E^{Seg} = \left(E^{Interface}_{with\ solute} - E^{Interface}_{without\ solute}\right) - \left(E^{Bulk}_{with\ solute} - E^{Bulk}_{without\ solute}\right), \quad (1)$$

where $E^{Interface}_{with\ solute}$ denotes energy of interface with the solute; $E^{Interface}_{without\ solute}$ is defined as the energy of interface without the solute; $E^{Bulk}_{with\ solute}$ is energy of the bulk Fe with the solute; and $E^{Bulk}_{without\ solute}$ represents energy of bulk Fe without solute. The bulk Fe is composed of 128 Fe atoms under stress free condition. According to Eq. (1), a positive or negative E^{Seg} represent the solute segregation tendency into bulk iron, and the tendency of the solute to segregate to the interface, respectively.

Calculations of E^{Seg} were performed for 28 substitutional solutes at atomic sites within the interfacial Fe layer: Ag, Al, Ba, Cd, Co, Cu, Hf, Ir, K, Mg, Mo, Na, Nb, Ni, Os, Pd, Pt, Rb, Rh, Sc, Sr, Ta, Ti, V, W, Y, Zn, Zr. 10 solute segregation sites were selected randomly within the interfacial Fe layer of the semi-coherent interface, where 5 sites were randomly chosen each from the misfit and coherent region of the semi-coherent interface. Hence a total of 280 (28 solutes $\times$ 10 sites) segregation energies were performed using DFT calculations. At coherent interface A and B, the solute segregation was performed at all four chemically

unique interfacial iron sites (see Supp. Fig. 1 in supplementary material). It is notable that compared to fluorite structure yttria has structural vacancies leading to more than one chemically unique interfacial iron site. To create different interatomic distances that are present in the semi-coherent interface for ML model training, the Fe slab in coherent interfaces was translated along the *x*- and *y*-directions. During translation the strain on the iron and yttria layer remain unchanged. Details of these translated structures are presented in the Supplementary Information (see Supp. Fig. 2 and Supp. Tables 1-2 in supplementary material). A total of six structures (five translated structures plus the original untranslated structure) were chosen for each coherent interface A and B. Then E^{Seg} for all 28 solutes were calculated for each of these variants of the two coherent structures. A total of 672 (= 28 solutes × 4 sites × 6 displacements) segregation energies were generated each for coherent interface A and B.

2.2. ML investigations

2.2.1. Input features

The parameters that affect the E^{Seg} at the Fe/Y_2O_3 interface, also referred to features in ML, are summarized in Table 1. These 18 features are organized into three categories based on literature and domain knowledge: geometry, chemistry, and strain energy.

Geometrical features describe the local environment around the segregation site, particularly the distances between the solute site (i) (see a) in Table 1) and its neighboring atoms (see b)-d) in Table 1). Previous studies have demonstrated that E^{Seg} at Fe/Y_2O_3 interfaces is strongly influenced by the local oxygen environment, particularly the oxide forming tendency of the solute [10,13]. Consequently, the distance between the segregation site (i) and the nearest O atom (O), referred to as i–O, is considered as a key geometrical descriptor. In addition, competition may arise among the solute, the nearest Fe atom (Fe), and the nearest Y atom (Y) for the nearest O atom. To capture these effects, other geometrical features such as interatomic distances–i–Fe, i–Y, Fe–Y, Fe–O, and Y–O–are also included in our ML-model as they may affect E^{Seg}.

Chemical features pertain to the chemical properties of the solutes, as differences in chemistry between the solute and neighboring atoms can impact the solute segregation behavior at the Fe/Y_2O_3 interface [10,13]. Strain energy features account for effects of atomic misfit strain (between the solute and Fe) on solute segregation, expressed in terms of local strain energy [10,13]. The chemical and strain energy features are defined with details in the supplemental information.

2.2.2. ML techniques

This section outlines the working principles of the ML tools used in this work.

Table 1
Input features for ML. Values of the chemistry and strain energy features were obtained from literature, while geometric features were measured from the relaxed DFT structures.

Geometry	Chemistry	Strain Energy
i[a)]-Fe[b)] (Å)	Ionic radius (Å) [13]	Bulk modulus (GPa) [24]
i[a)]-Y[c)] (Å)	Crystal structure [13]	
i[a)]-O[d)] (Å)	Valency [13]	
Fe[b)]-O[d)] (Å)	Electronegativity [13]	
Fe[b)]-Y[c)] (Å)	Oxide forming tendency (eV/atom) [13]	
Y[c)]- O[d)] (Å)	Bulk cohesive energy (eV/atom) [25]	
	Electron affinity (eV/atom) [26]	
	First Ionization potential (eV/atom) [27]	
	Dipole polarizability (a.u.) [28]	
	Electronic configuration [29]	
	Magnetic moment (BM) [30]	

[a)] segregation site;
[b)] nearest Fe atom of segregation site;
[c)] nearest Y atom of segregation site;
[d)] nearest O atom of segregation site.

2.2.2.1. Data scaling. The data obtained from the DFT calculations include various units and ranges, which could hinder the ML model from accurately learning the relationship between features and target i.e. segregation energy. Thus, data scaling is applied to reduce the impact of feature units and ranges on model training. Then the training set was scaled using standardization in this study.

The formula of the standard scaler is expressed as [31]:

$$x_{new} = \frac{x - \mu}{\sigma}, \tag{2}$$

where x_{new} is the data after scaling; x denotes original data; μ represents mean value within the feature; and σ is defined as standard deviation within the feature.

2.2.2.2. Random forest. In this study, random forest [32–34] is used as a feature importance analysis tool on the dataset collected from coherent interface A to investigate the relative importance of each feature in determining E^{Seg}. Unlike other commonly used feature importance analysis tools [35–37], random forest is more versatile in capturing nonlinear relationships between features, analyzing them numerically, and addressing the issue of overfitting.

Random forest consists of multiple decision trees, with each decision tree being a non-parametric supervised learning method suitable for both categorical and numerical targets. It is a tree-like model constructed by recursively splitting data based on decisions made under various conditions. In this research, the optimal attribute for splitting the data at the nodes of each decision tree is determined by mean squared error (MSE) [32–34]:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2, \tag{3}$$

where n represents the total number of samples; $\hat{y}_i$ is the ith predicted target; and y_i represents the ith true target. More details of random forest are covered elsewhere [38].

In this study, the hyperparameters of the random forest model were optimized with the following configuration setting: 600 trees, the max depth of 60 for each tree, 4 features to apply at a root or decision node for the best split, minimum 1 sample at a leaf node and minimum 2 samples to split a root or decision node.

2.2.2.3. Kernel ridge regression. To predict E^{Seg} at the Fe/Y_2O_3 coherent and semi-coherent interfaces for a given solute and atomic site, kernel ridge regression (KRR) is applied to model training in this study. By combining Ridge regression with a kernel function, KRR explores the nonlinear relationship between features and the numerical target in a higher-dimensional space compared to the original space [39]. There are several kernel functions available in KRR, including the polynomial kernel, radial basis function (RBF) kernel, sigmoid kernel, and others. Numerous engineering studies have employed KRR for numerical predictions, which demonstrates its effectiveness in training regression models [39–41]. In addition to its popularity in the engineering field, the closed-form solution and high computational efficiency of KRR make it an ideal choice for this work. In KRR, the loss function to improve the model is expressed as below:

$$\mathrm{Loss} = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 + \lambda \boldsymbol{w}^T \boldsymbol{K} \boldsymbol{w} \tag{4}$$

where $\hat{y}_i$ is the ith predicted target; y_i represents the ith true target; λ is a hyperparameter that determines the regularization strength, $\boldsymbol{w}$ is the model weight and $\boldsymbol{K}$ is the kernel function. The first term in Eq. (4) is to evaluate the errors between the true target and the prediction from KRR, while the second term calculates a penalty of large model weight values

which may cause the overfitting issue. More details of KRR can be found elsewhere [42]. Grid search was conducted to tune the hyperparameters of the KRR models in this study, with the values presented in a)-d) from Table 2.

2.2.2.4. Evaluation metrics. In this study, the performance of each model was evaluated using the coefficient of determination (R^2) [43], mean absolute error (MAE), and root mean square error (RMSE) [44]. R^2 is calculated by the difference between the true target and the predicted target and the difference between the true target and the mean true target, evaluating how well the data fits the model. MAE represents the average of the absolute differences between the true target and the predicted target, providing a measure of the average residuals. RMSE quantifies the square root of the average squared differences between the true target and the predicted target, capturing the variance of the residuals.

R^2, MAE and RMSE are calculated by the following equations:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}, \tag{5}$$

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|, \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}, \tag{7}$$

where n is total number of samples; $\bar{y}_i$ represents the mean value of the target; $\hat{y}_i$ is the predicted target; and y_i represents the true target.

All the ML techniques utilized in this work were implemented in the Scikit-learn library.

3. Results

3.1. Feature importance analysis

The segregation energy of solutes at a given site is governed by the competition between the different features listed in Table 1. However, it's difficult to systematically understand the relative role of a large number of features on E^{Seg} using the traditional computational approaches (e.g., DFT and MD). In this work, random forest, a supervised learning ML technique, is applied to extract the relationship between the features and E^{Seg}. The relative importance for each feature on E^{Seg} of coherent interface A is illustrated in Fig. 2. The horizontal-axis is the feature importance value while the vertical-axis lists the features according to the feature importance. Higher feature importance implies a stronger co-relation between the feature and the E^{Seg}. From Fig. 2, bulk modulus, electronegativity, dipole polarizability, ionic radius, first ionization potential and bulk cohesive energy show much more profound effects on E^{Seg} compared to the remaining 12 features. Bulk modulus is the most important out of all the 18 features, which implies the strain energy arising from the atomic misfit between the solute and iron atom has a far-reaching influence on the E^{Seg}.

Among the features related to the chemistry of the solutes, electronegativity of the solute exhibits the strongest link to E^{Seg}. This can be explained based on Hume Rothey rules, which states that, when the electronegativity difference between a solute and a solvent is close to zero, the solute will have maximum solubility in bulk Fe. However, as the difference in electronegativity increases between the solute and Fe, the tendency of the solute to segregate to the interface is expected to increase, as observed in our random forest analysis. It is notable that solutes with less electronegativity (e.g.: Mg) than iron exhibits higher oxide forming tendency compared to iron. It can be concluded that tendency of a solute to form oxide with nearest O atom within the interfacial yttria layer and its relative oxidation tendency compared to neighboring Fe atom, plays a vital role in determining E^{Seg}. This explains feature i-Fe, O–Fe and i-O scores higher in feature importance (Fig. 2), compared to other three geometrical features. It should be pointed out the values of all the geometrical features are unique at all the four coherent sites as presented in Supp. Fig. 1, leading to some variation in E^{Seg} among the four sites. It is clear from Fig. 2 that the top 9 features are all related to the local strain energy or chemistry of the solute. Hence, to describe E^{Seg} at coherent interface A, strain energy and chemistry of solute occupy more crucial positions over the local atomic environment surrounding the segregation site.

Table 2
Hyperparameters used in the KRR models.

KRR Model	Alpha[a)]	Gamma[b)]	Degree[c)]	Kernel[d)]
Fig. 3(a)	0.3	0.1	3	polynomial
Fig. 3(b)	0.8			
Fig. 4(a)	0.4			
Fig. 4(b)	0.7			

a) regularization strength;
b) coefficient of the polynomial kernel function;
c) degree of the polynomial kernel function;
d) mathematical functions used in KRR models.

3.2. ML models for prediction of solute segregation behavior

3.2.1. Coherent interface A and coherent interface B

KRR was utilized to develop a quantitative model for the prediction of E^{Seg}. A dataset comprising 672 samples from coherent interface A was applied to train the KRR model. The dataset was randomly split between training and test set in a ratio of 80:20. Model training was performed using the training set, while the trained model was evaluated by the test set. The model performance for coherent interface A is shown in Fig. 3 (a), where the horizontal and vertical axis are the DFT-calculated and predicted E^{Seg}, respectively. In this scatter plot, all the samples lie close to the diagonal parity line, where DFT calculated E^{Seg} equals the KRR predicted E^{Seg}. The distribution of the samples suggests the potential of the model to predict the E^{Seg} at coherent interface A.

In Fig. 3(a) solutes with highly negative segregation energy E^{Seg} (−6 eV ~ −2 eV) exhibit higher errors compared to solutes with lower E^{Seg} (−2 eV ~ 2 eV). An analysis of all the features demonstrates that solutes with highly negative E^{Seg} possess much smaller electronegativity (K, Rb, Ba, Na, Sr < 1) compared to Fe atoms (1.83). Hence, these solutes present a strong tendency to segregate to the coherent interface A instead of remaining in bulk Fe. The errors caused in these ML predicted E^{Seg} may originate from distribution of data within the dataset. The dataset for small electronegativity (<1) solutes corresponds to only 17.85 % percent of the total dataset. Thus a lack of data points for solutes with small electronegativity may lead to higher error in ML predictions. However, the overall model performance can also be evaluated by R^2 (0.9720), RMSE (0.2738 eV) and MAE (0.1879 eV). R^2 value close to unity demonstrates strong model performance in predicting segregation energy at the coherent interface A.

Solute segregation energy at a coherent metal/oxide interface depends on the misfit strain between the metal and the oxide. To test the ability of our ML tools to predict the solute segregation energy for a range of interfacial strain conditions we created coherent interface B between Fe and Y_2O_3 with the same OR at coherent interface A with different strain conditions as presented earlier.

KRR model was trained on the segregation energy dataset of coherent interface B and the model performance is presented in Fig. 3 (b). Similar to coherent interface A, the dataset was randomly split to the training set (80 %) and test set (20 %). In coherent interface B, R^2 of 0.9718 describes strong model performance, similar to the model for

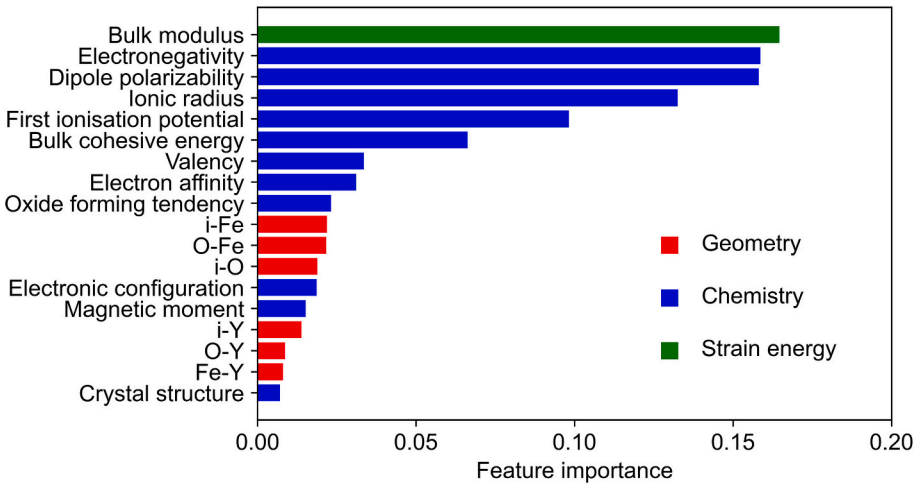

Fig. 2. Feature importance analysis of using random forest (data source: coherent interface A).

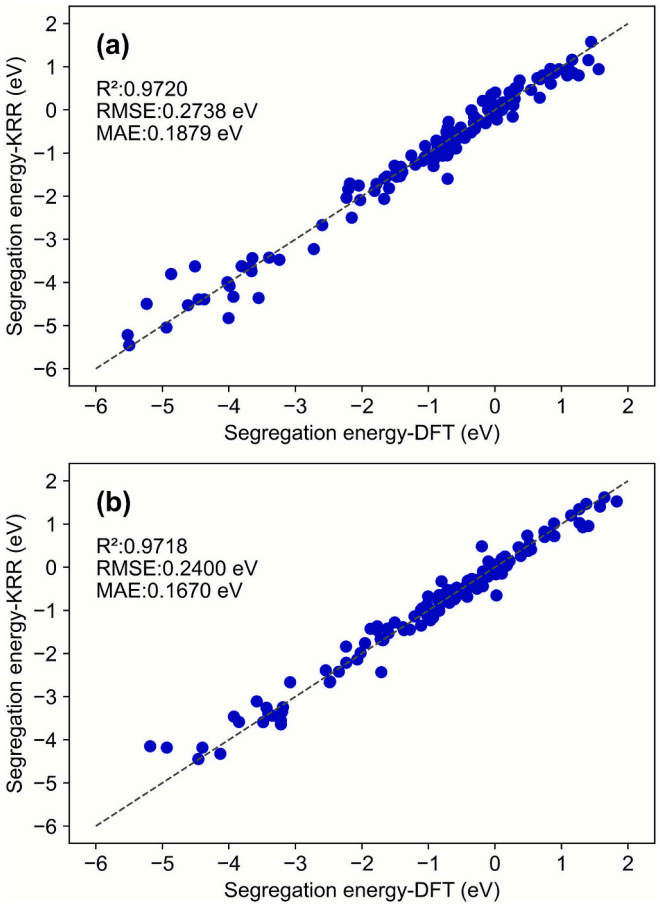

Fig. 3. KRR model performance; predicted E^{Seg} vs. computational E^{Seg} for: (a) coherent interface A; (b) coherent interface B.

coherent interface A, which verifies the potential of ML tool to predict the E^{Seg} of Fe/Y_2O_3 coherent interfaces under different strain conditions. For the data ranging from −6 eV to −2 eV, in contrast to coherent interface A, the data points from coherent interface B are distributed more closely along the parity line. Considering the strain conditions of the two coherent interfaces, the largest strains were imposed in the *y*-direction on the Fe side (24.74 % in coherent interface A compared to 18.73 % in coherent interface B). One possible reason is that the higher strain in coherent interface A elevates the distortion of the local atomic arrangement, which requires more geometric features for ML model training. As a result, the model performance for coherent interface A with higher strain presents higher errors than coherent interface B. A possible way to improve the prediction for coherent interface A is to include geometric features beyond the first nearest neighbor including second or third nearest neighbors.

3.2.2. *Semi-coherent interface*

Calculation ofE^{Seg} at a semi-coherent metal/oxide interface is highly computationally expensive due to the large number of atoms needed to describe the misfit dislocation structure. To design a more efficient method to investigate the solute segregation at semi-coherent Fe/Y_2O_3 interfaces, we utilized the E^{Seg} at coherent interface A or B to train ML models. Later the trained models were used to predict E^{Seg} at the semi-coherent interface. The predictions of E^{Seg} for the semi-coherent interface using KRR model trained on E^{Seg} data for coherent interface A is presented in Fig. 4(a). Data points from misfit and coherent regions of the semi-coherent interface are plotted in red and green colors, respectively. It can be observed that the scattered symbols lie close to the parity line indicating that the model trained on segregation energy data obtained for coherent interface A can predict the segregation tendency for the semi-coherent interface. Additionally, the model's R^2 value of 0.9577—a strong indicator of model fitting, with 1 representing a perfect fit—further supports our conclusion.

In contrast to the model performance (R^2 of 0.9720, RMSE of 0.2738 eV and MAE of 0.1879 eV) in Fig. 3(a), the scores for models in Fig. 4(a) are slightly lower (R^2 of 0.9577, RMSE of 0.3276 eV and MAE of 0.2361 eV). The gap between the two model performances is linked to the characteristics of the coherent and semi-coherent interfaces. Unlike coherent interface A, the semi-coherent interface is composed of coherent terraces separated by misfit dislocations. The atomic environment, especially the local Fe/O ratio at the misfit region (2 columns of Fe sharing electrons with one column of oxygen), is different compared to the coherent interfaces leading to additional ML-predicted error ofE^{Seg} for the semi-coherent interface. Furthermore, in Fig. 4(a), the samples from misfit region (red) deviate more from the parity line than the samples from coherent region (green). Hence, the KRR model trained by coherent interface A performs more accurately on predicting E^{Seg} at coherent region over the misfit region of the semi-coherent interface.

Next, we tested the ability of our ML models to predict the segregation energy of the semi-coherent interface with a model trained on segregation energy data from coherent interface B and the results are presented in Fig. 4(b). Compared to Fig. 4(a), a more apparent deviation of samples from the parity line and worse metrics (R^2 of 0.9075, RMSE of

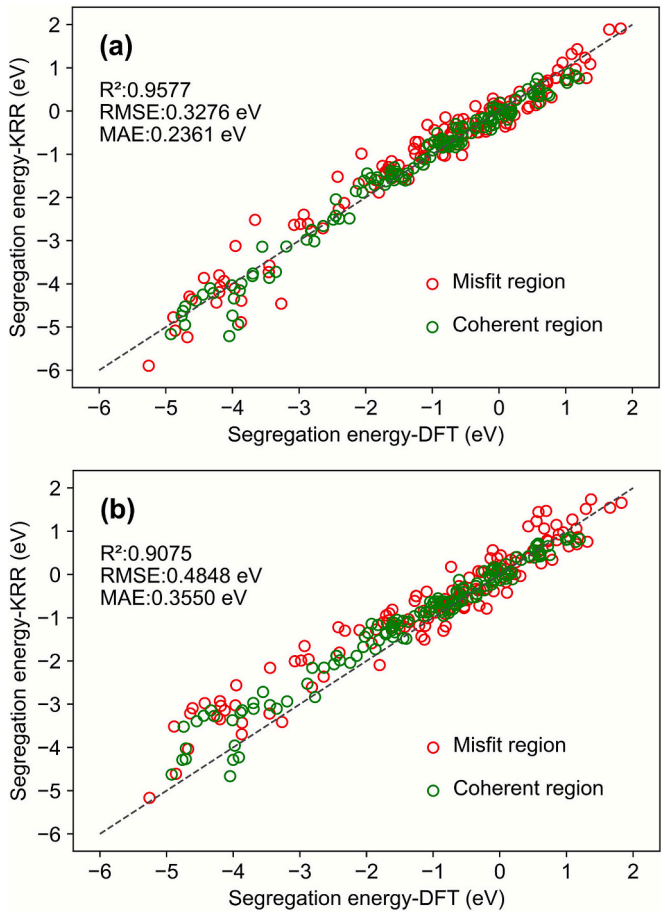

Fig. 4. KRR model performance of predicting E^{Seg} at the semi-coherent interface; predicted E^{Seg} vs. computational E^{Seg} for: (a) coherent interface A as the training set; (b) coherent interface B as the training set.

0.4848 eV and MAE of 0.3550 eV) is observed in Fig. 4(b) which can be explained based on the strain conditions of the semi-coherent interface, coherent interface A and B. Similar strain conditions were applied to Y_2O_3(−1.1 % in *x*-axis, and - 4.67 % in *y*-axis) in both the semi-coherent interface and coherent interface A, while in coherent interface B Y_2O_3 is strained by −2.18 % in the *x*-direction, −9.25 % in the *y*-direction. Therefore, the similar Y_2O_3 strain conditions between the semi-coherent interface and coherent interface A help predict the E^{Seg} at semi-coherent interface A. Furthermore, in Fig. 4(b), it is observed that the samples with low E^{Seg} values (ranging from −6 eV to −2 eV) are shifted slightly upward possibly because these solutes exhibit much larger ionic radii (e.g., K, Rb, Ba, Na, Sr, with ionic radii >1.15 Å) compared to Fe atoms (0.75 Å).

The accuracy in predicting E^{Seg} at the semi-coherent interface using a model trained on E^{Seg} at coherent interface A (Fig. 4(a)) is further tested by comparing the DFT and ML-predicted non-uniform distribution of segregation energy at the semi-coherent interface. For a given solute in Fig. 5(a), E^{Seg}_{max} and E^{Seg}_{min} represent the maximum and the minimum segregation energy among the ten previously chosen interfacial iron atomic sites at the semi-coherent interface, respectively. The bar chart presents the energy difference (ΔE^{Seg}) between maximum and minimum E^{Seg} (E^{Seg}_{max} and E^{Seg}_{min}) obtained from DFT calculations and ML-predictions, respectively. Thus, the blue bars (ΔE^{Seg} from DFT) represent the range (or in-plane variation) of segregation energy for a given solute obtained from DFT calculations at the semi-coherent interface. ΔE^{Seg} from ML (red bars) when compared to ΔE^{Seg} from DFT can serve as a metric to evaluate the sensitivity of ML models to capture in-plane variation of the segregation tendency of solutes at the semi-coherent interface. That is, this metric provides insight into how well the model trained only on data from the coherent interface can capture variations in segregation energy resulting from local variations in atomic structure induced by the presence of the misfit dislocations in semi-coherent interface. The solutes in Fig. 5(a) are arranged by decreasing electronegativity from left to right. For most solutes the ML model predicted ΔE^{Seg} compares well with the DFT predicted ΔE^{Seg}, implying that the ML model can effectively capture the in-plane variation of segregation tendencies in the semi-coherent interface even though the models were trained with segregation energy data from the smaller fully coherent interface A. However, eight (Na, Hf, K, Ta, Y, Sr, Rb and Zr) out of 28 solutes display higher discrepancies (>0.6 eV) between the ΔE^{Seg} from DFT and ML models. All eight of these solutes also have a low electronegativity and a strong oxide forming tendency (ranging from −13.12 eV/atom to −6.94 eV/atom) compared to Fe (−2.64 eV/atom). One potential reason for the discrepancy in ΔE^{Seg} between DFT and ML is due to the limitation of our ML model. For a given solute segregation site at the interface, only the distance to the first nearest neighboring O, Fe and Y atoms was considered. For simplicity, in our model distances to other O atoms in the neighborhood of the solute segregation site were ignored. This may limit the ability of our model to capture the unique local geometrical effects at the semi-coherent interface on E^{Seg}, which is particularly important for solutes with lower electronegativity (or solute with higher tendency to form oxide than iron).

Fig. 5(b) tests the ability of the ML model to accurately predict the region (coherent vs. misfit) where the maximum or minimum segregation energies at the semi-coherent interface were obtained from DFT calculations. This is also an alternative way to evaluate the ability of ML model to predict non-uniform segregation tendency of a given solute across the interface. Each quadrant shows the location of the maximum and minimum segregation energy obtained from DFT calculations. For example, Na is in the first quadrant, which implies that from DFT calculations E^{Seg}_{min} and E^{Seg}_{max} are both located at the coherent region of the semi-coherent interface. Later, the color of the dots was used to indicate if the ML model was able to predict the location of the E^{Seg}_{min} and E^{Seg}_{max} obtained from DFT calculations. For example, the green dots represent solutes for which ML model accurately predicted the region for both E^{Seg}_{min} and E^{Seg}_{max} obtained from DFT calculations. It should be pointed out that for the solutes with green dots not only the ML-model correctly predicts the region (coherent vs. misfit) of E^{Seg}_{min} and E^{Seg}_{max}, our ML-model also accurately predicts the particular atomic site with each region for which each of the extreme segregation energy was calculated from DFT. The cyan and the red dots represent solutes for which ML model failed to predict the region for E^{Seg}_{max} and E^{Seg}_{min}, respectively. It is clear from Fig. 5(b) that the ML model accurately predicted the region for both E^{Seg}_{min} and E^{Seg}_{max} for 23 out of 28 solutes. However, a mismatch is observed in the cases of Ba (E^{Seg}_{min}), Co (E^{Seg}_{max}), Cu (E^{Seg}_{max}), K (E^{Seg}_{max}) and Na (E^{Seg}_{max}). The absolute difference in ML predicted E^{Seg} at the misfit or coherent region for the five cases are Ba (0.28 eV), Co (0.05 eV), Cu (0.25 eV), K (0.70 eV) and Na (0.97 eV). Although Co exhibits a discrepancy, the ML predicted E^{Seg} for the misfit and coherent regions show a small difference of 0.05 eV, which might be caused by the limitation of the ML model. In the cases of Ba, Cu, K, and Na, the discrepancy is found to originate from a single site within the misfit region. Examination revealed that the distance of this Fe interfacial site to the nearest Y atom (2.58 Å), which is feature i-Y in Table 1, is beyond the range of i-Y distances in the training set of sites in coherent interface A (2.82–3.6 Å), which increases the uncertainty in predicting the segregation energy for this site. Inclusion of other i-Y geometric features beyond the first nearest neighbors may help improve the prediction for the remaining four elements.

Although in Fig. 5(a) several solutes show large differences in ΔE^{Seg} between DFT and the ML model, the percentage of cases in Fig. 5(b)

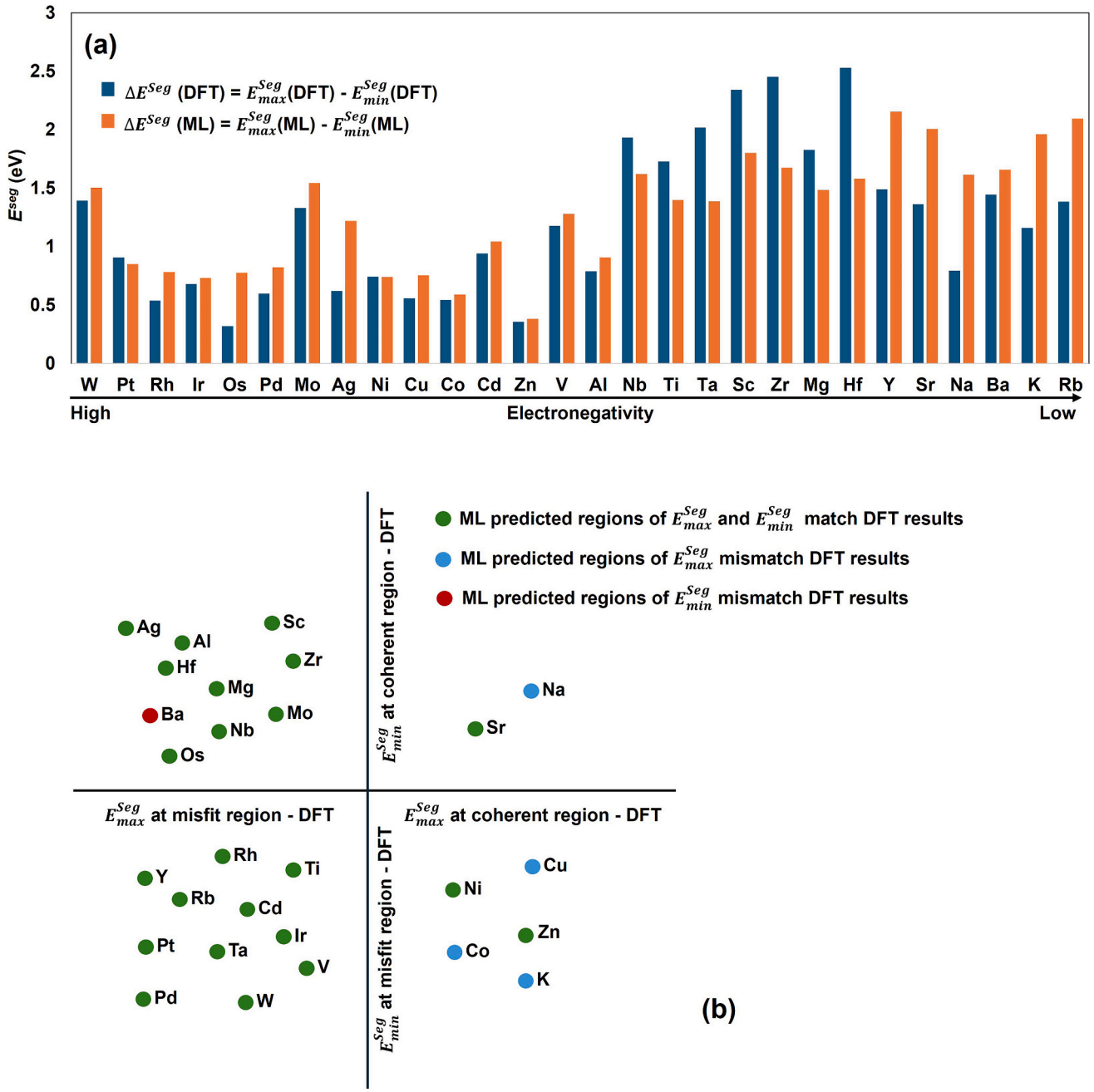

Fig. 5. Analysis of computational and predicted E^{Seg} at the semi-coherent interface: (a) energy difference between E^{Seg}_{max} and E^{Seg}_{min} among the 10 sites for each solute; (b) test of the regions (coherent vs. misfit) where E^{Seg}_{max} and E^{Seg}_{min} are located. The predicted E^{Seg} for semi-coherent interface was obtained by KRR model when trained on E^{Seg} data obtained for coherent interface A.

where the location of Fe-sites with extreme (both max and min) E^{Seg} was correctly identified by the ML model was greater than 91 % (51 out of 56 extreme sites (two extreme energies for each of the 28 solutes)). This implies that the ML model trained on segregation energy data at coherent interface A can predict the non-uniform segregation tendency within the interfacial Fe-layer with complex dislocation configurations within the semi-coherent interface. Combining the observations from Figs. 4(a) and 5, we can see that the ML model trained on segregation energy data at coherent interface A provides an excellent approach to predict ΔE^{Seg} at the semi-coherent interfaces with DFT level accuracy at a significantly reduced computational cost compared to direct DFT calculations of E^{Seg} at the semi-coherent interface. For example, the average computational cost for DFT calculations of the segregation energy of 28 solutes at one interfacial atomic site for the semi-coherent interface (with 345 atoms) versus the coherent interface A (with 108 atoms) are 34,799 cpu-hours and 994.3 cpu-hours, respectively. This implies that the computational cost of generating DFT-calculated solute segregation energies at coherent interface A followed our ML models to predict the segregation energy of solutes at the semi-coherent interface is 1/35th of the computational cost to calculate the segregation energy directly from DFT. Moreover, the DFT calculations of coherent interfaces (with 108 atoms) can be performed using a workstation and do not require high-end computational clusters, typically needed to perform DFT calculations of semi-coherent interfaces with several hundred atoms. This indicates that a combination of DFT calculations and ML tools provides an excellent approach for investigating the properties of

semi-coherent interfaces using traditional computational resources where direct DFT calculations of semi-coherent interfaces are not possible.

4. Conclusion

In this work we have investigated the segregation tendency of substitutional solutes to coherent and semi-coherent Fe/Y_2O_3 interfaces using a hybrid method of DFT and ML. Feature importance analysis showed that features related to elastic strain due to atomic misfit as well as chemical features of the solutes play the most vital role in governing the solute segregation energy (E^{Seg}) at coherent Fe/Y_2O_3 interface, similar to semi-coherent interfaces [18]. Moreover, the DFT-calculated solute segregation energy at coherent interfaces was used to train and optimize the ML model to predict E^{Seg} at a semi-coherent interface with high fidelity, including non-uniform distribution of segregation energies between the misfit and coherent region of the semi-coherent interface.

It is notable that as the size of oxide particles grows within a metallic matrix the metal/oxide interface may transition from coherent to a semi-coherent interface [9]. This work shows that using a combination of DFT calculations and ML approaches it is possible to predict properties at a metal/oxide interface for a wide range of particle sizes when the model is trained using properties for a small surrogate coherent supercell. Furthermore, it has been observed experimentally that, as an oxide particle within a metallic matrix grows, spherical particle with coherent metal/oxide interfaces transition to faceted structures characterized by semi-coherent interfaces with each facet having a different OR with the matrix [45]. Our previous published work revealed that it is not necessary to calculate the E^{Seg} for each facet individually [18]. Instead, once E^{Seg} is calculated for one semi-coherent interface with a given OR, ML can be applied to predict E^{Seg} for the other facets with other ORs, thus providing a significant reduction in computational cost to investigate segregation tendency at particle-matrix interface. Combining these two works, there is a considerable potential that DFT-calculated properties at a single coherent metal/oxide interface in combination with ML would provide accurate predictions of properties for a wide range of particle size, coherency, misfit strain and OR, thus providing a comprehensive understanding of properties at a range of metal/oxide interfaces at a fraction of computational cost.

Finally, our ML-based approach exhibits the potential to quantitatively predict properties for a wider range of interfaces, including metal/metal, oxide/oxide, and broader metal/ceramic interfaces, provided appropriate features are included in the ML models. For example, chemical features such as electronegativity play vital roles in governing properties at metal/oxide interfaces as significant charge transfer often occurs across the metal and oxide [46]. However, such chemical features may play a smaller role in dictating properties at immiscible metal/metal interfaces such as the Cu/V interface [47]. Overall, a hybrid DFT and ML approach shows excellent promise for investigating properties at a broader range of heterointerfaces which is otherwise computationally expensive if not prohibitive using traditional DFT calculations.

CRediT authorship contribution statement

Yizhou Lu: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Blas Pedro Uberuaga:** Writing – review & editing, Visualization, Validation, Methodology, Formal analysis, Conceptualization. **Samrat Choudhury:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Funding sources

Dr. Samrat Choudhury was financially supported by the National Science Foundation Award number: 2150816.

Declaration of competing interest

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

Acknowledgements

This work was supported through the INL Laboratory Directed Research & Development (LDRD) Program and the Nuclear Science User Facilities under Contract No. DE-AC07-05ID14517. Dr. Blas Pedro Uberuaga was supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001). This research made use of the resources at the High-Performance Computing Center at Idaho National Laboratory, which is supported by the office of Nuclear Energy of the U.S. Department of Energy under contract No. DE-AC07-05ID14517.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.commatsci.2025.114480.

Data availability

Data will be made available on request.

References

[1] G. Pacchioni, L. Skuja, D.L. Griscom, Defects in SiO2 and related dielectrics: science and technology, Springer Science & Business Media 2 (2012).
[2] C. Cagli, et al., Resistive-switching crossbar memory based on Ni–NiO core–shell nanowires, Small 7 (20) (2011) 2899–2905.
[3] H.-J. Freund, S. Shaikhutdinov, N. Nilius, Model studies on heterogeneous catalysts at the atomic scale, Top. Catal. 57 (2014) 822–832.
[4] X. Guo, et al., Ferrous centers confined on core–shell nanostructures for low-temperature CO oxidation, J. Am. Chem. Soc. 134 (30) (2012) 12350–12353.
[5] S. Surnev, A. Fortunelli, F.P. Netzer, Structure–property relationship and chemical aspects of oxide–metal hybrid nanostructures, Chem. Rev. 113 (6) (2013) 4314–4372.
[6] M.F. Montemor, Functional and smart coatings for corrosion protection: a review of recent advances, Surf. Coat. Technol. 258 (2014) 17–37.
[7] G.R. Odette, M.J. Alinger, B.D. Wirth, Recent developments in irradiation-resistant steels, Annu. Rev. Mater. Res. 38 (1) (2008) 471–503.
[8] G.R. Odette, D.T. Hoelzer, Irradiation-tolerant nanostructured ferritic alloys: transforming helium from a liability to an asset, JOM 62 (9) (2010) 84–92.
[9] F. Ernst, Metal-oxide interfaces, Mater. Sci. Eng. R. Rep. 14 (3) (1995) 97–156.
[10] S. Choudhury, D. Morgan, B.P. Uberuaga, Massive interfacial reconstruction at misfit dislocations in metal/oxide interfaces, Sci. Rep. 4 (1) (2014) 6533.
[11] X.G. Wang, J.R. Smith, A. Evans, Fundamental influence of C on adhesion of the Al2O3/Al interface, Phys. Rev. Lett. 89 (28) (2002), 286102.
[12] P. Hou, Segregation phenomena at thermally grown Al2O3/alloy interfaces, Annu. Rev. Mater. Res. 38 (1) (2008) 275–298.
[13] S. Choudhury, J.A. Aguiar, M.J. Fluss, L.L. Hsiung, A. Misra, B.P. Uberuaga, Non-uniform solute segregation at semi-coherent metal/oxide interfaces, Sci. Rep. 5 (1) (2015) 13086.
[14] D. Isheim, D.N. Seidman, Nanoscale studies of segregation at coherent heterophase interfaces in α-Fe based systems, Surface and Interface Analysis: An International Journal Devoted To The Development And Application Of Techniques For The Analysis Of Surfaces, Interfaces And Thin Films 36 (5–6) (2004) 569–574.
[15] N. Kaur, C. Deng, O.A. Ojo, Effect of solute segregation on diffusion induced grain boundary migration studied by molecular dynamics simulations, Comput. Mater. Sci. 179 (2020) 109685.
[16] V. Shutthanandan, S. Choudhury, S. Manandhar, T.C. Kaspar, C. Wang, A. Devaraj, B.D. Wirth, S. Thevuthasan, R.G. Hoagland, P.P. Dholabhai, B.P. Uberuaga, Radiation tolerant interfaces: influence of local stoichiometry at the misfit

dislocation on radiation damage resistance of metal/oxide interfaces, Adv. Mater. Interfaces 4 (14) (2017) 1700037.
[17] Q. Liu, L. Yin, C. Zhao, Z. Wu, J. Wang, X. Yu, Z. Wang, W. Wei, Y. Liu, I. Z. Mitrovic, L. Yang, All-in-one metal-oxide heterojunction artificial synapses for visual sensory and neuromorphic computing systems, Nano Energy 97 (2022) 107171.
[18] Y. Lu, B.P. Uberuaga, S. Choudhury, Prediction of solute segregatiom at metal/oxide interfaces using machine learning approaches, Molecules 30 (16) (2025) 3344.
[19] B.J. Inkson, P.L. Threadgill, Y2O3 morphology in an oxide dispersion strengthened FeAl alloy prepared by mechanical alloying, MRS Online Proceedings Library (OPL) 460 (1996) 767.
[20] G. Kresse, Ab initio molecular dynamics for liquid metals, J. Non-Cryst. Solids 192-193 (1995) 222–229.
[21] G. Kresse, J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B Condens. Matter 54 (16) (1996) 11169–11186.
[22] P.E. Blöchl, Projector augmented-wave method, Phys. Rev. B Condens. Matter 50 (24) (1994) 17953–17979.
[23] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (18) (1996) 3865.
[24] D. McLean. Grain boundaries in metals, Oxford university press, 1957.
[25] V.I. Razumovskiy, A. Lozovoi, I. Razumovskii, First-principles-aided design of a new Ni-base superalloy: influence of transition metal alloying elements on grain boundary and bulk cohesion, Acta Mater. 82 (2015) 369–377.
[26] A. Cottrell, Unified theory of effects of segregated interstitials on grain boundary cohesion, Mater. Sci. Technol. 6 (9) (1990) 806–810.
[27] D. Hudson, G.D. Smith, Initial observation of grain boundary solute segregation in a zirconium alloy (ZIRLO) by three-dimensional atom probe, Scr. Mater. 61 (4) (2009) 411–414.
[28] J. Petrazoller, J. Guénolé, S. Berbenni, T. Richeton, On the effect of elastic anisotropy and polarizability on solute segregation at low-angle grain boundaries, Comput. Mater. Sci. 249 (2025) 113642.
[29] R. Magri, A. Zunger, Effects of interfacial atomic segregation and intermixing on the electronic properties of InAs/GaSb superlattices, Phys. Rev. B 65 (16) (2002) 165302.
[30] K. Ito, H. Sawada, First-principles analysis of the grain boundary segregation of transition metal alloying elements in γFe, Comput. Mater. Sci. 210 (2022) 111050.
[31] P.J.M. Ali, R.H. Faraj, E. Koya, P.J.M. Ali, R.H. Faraj, Data normalization and standardization: a technical report, Mach, Learn. Tech. Rep. 1 (1) (2014) 1–6.
[32] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[33] T. Hastie, R. Tibshirani, J. Friedman. The elements of statistical learning: data mining, inference, and prediction, Springer, 2009.
[34] L. Breiman, Classification and Regression Trees, Routledge, 2017.
[35] R.A. Dobie, M.J. Wilson, A comparison of t test, F test, and coherence methods of detecting steady-state auditory-evoked potentials, distortion-product otoacoustic emissions, or other sinusoids, J. Acoust. Soc. Am. 100 (4) (1996) 2236–2246.
[36] N. Veyrat-Charvillon, F.-X. Standaert, Mutual information analysis: How, when and why?, in: International Workshop on Cryptographic Hardware and Embedded Systems Springer, 2009.
[37] M.U. Ali, A. Kabiru, C. Haruna, Multi-objective feature selection using non-dominated sorting mechanisms and bi-directional elimination for heart disease classification, Lautech J. Computing and Informatics 4 (2) (2024) 1–14.
[38] Y. Lu, J. Treadway, P. Ghimire, Y. Han, S. Choudhury, Prediction of electrohydrodynamic printing behavior using machine learning approaches, Int. J. Adv. Manuf. Technol. 136 (10) (2025) 4439–4454.
[39] X. Wu, P. Zhao, Predicting nuclear masses with the kernel ridge regression, Physical Rev. C 101 (5) (2020) 051301.
[40] L. Tang, Z.-H. Zhang, Nuclear charge radius predictions by kernel ridge regression with odd–even effects, Nucl. Sci. Tech. 35 (2) (2024) 19.
[41] D. Moreno-Salinas, R. Moreno, A. Pereira, J. Aranda, J.M. de la Cruz, Modelling of a surface marine vehicle with kernel ridge regression confidence machine, Appl. Soft Comput. 76 (2019) 237–250.
[42] K. Chitnis, Y. Lu, B. Rhoads, L.R.J. Chakka, S. Choudhury, M. Maniruzzaman, Optimization of print parameters for batch and continuous manufacturing of three-dimensional (3D) printed dosage forms using artificial intelligence and machine learning, Drug Deliv. Transl. Res. (2025) 1–16.
[43] J. Miles, R-squared, adjusted R-squared, Encyclopedia of Statistics in Behavioral Science 4 (2005) 1655–1657.
[44] A. Vidianto, A. Sindunata, N. Yudistira, Air Pollution Particulate Matter (PM2.5) Forecasting Using Long Short Term Memory Model. New York, NY, USA: ACM, 2025.
[45] M. Klimiankou, R. Lindau, A. Möslang, HRTEM study of yttrium oxide particles in ODS steels for fusion reactor application, J. Cryst. Growth 249 (1–2) (2003) 381–387.
[46] L. Giordano, G. Pacchioni, Charge transfers at metal/oxide interfaces: a DFT study of formation of K δ+ and Au δ− species on MgO/Ag (100) ultra-thin films from deposition of neutral atoms, Phys. Chem. Chem. Phys. 8 (28) (2006) 3335–3341.
[47] N. Li, M. Demkowicz, N. Mara, Y. Wang, A. Misra, Hardening due to interfacial he bubbles in nanolayered composites, Mater. Res. Lett. 4 (2) (2016) 75–82.