Correlation analysis of materials properties by machine learning: Illustrated with stacking fault energy from first-principles calculations in dilute fcc-based alloys

Xiaoyu Chong,1,2 Shun-Li Shang,2,\* Adam M. Krajewski,2 John D. Shimanek,2 Weihang Du,2 Yi Wang,2 Jing Feng,1,\* Dongwon Shin,3 Allison M. Beese,2 and Zi-Kui Liu2

1Faculty of Material Science and Engineering, Kunming University of Science and Technology, Kunming 650093, China

2Department of Materials Science and Engineering, The Pennsylvania State University, University Park, PA, 16802, USA

3Materials Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA

\*Emails: sus26@psu.edu (S. L. Shang) and jingfeng@kust.edu.cn (J. Feng)

**ABSTRACT**

Advances in machine learning (ML), especially in the cooperation between ML predictions, density functional theory (DFT) based first-principles calculations, and experimental verification are emerging as a key part of a new paradigm to understand fundamentals, verify, analyze, and predict data, and design and discover materials. Taking stacking fault energy (γSFE) as an example, we perform a correlation analysis of γSFE in dilute Al-, Ni-, and Pt-based alloys by descriptors and ML algorithms. These γSFE values were predicted by DFT-based alias shear deformation approach, and up to 49 elemental descriptors and 21 regression algorithms were examined. The present work indicates that (i) the variation of γSFE affected by alloying elements can be quantified through 14 elemental attributes based on their statistical significances to decrease the mean absolute error (MAE) in ML predictions, and in particular, the number of p valence electrons, a descriptor second only to the covalent radius in importance to model performance, is unexpected; (ii) the alloys with elements close to Ni and Co in the periodic table possess higher γSFE values; (iii) the top four outliers of DFT predictions of γSFE are for the alloys of Al23La, Pt23Au, Ni23Co, and Al23Be based on the analyses of statistical differences between DFT and ML predictions; and (iv) the best ML model to predict γSFE is produced by Gaussian process regression with an average MAE < 8 mJ/m2. Beyond detailed analysis of the Al-, Ni-, and Pt-based alloys, we also predict the γSFE values using the present ML models in other fcc-based dilute alloys (i.e., Cu, Ag, Au, Rh, Pd, and Ir) with the expected MAE < 17 mJ/m2 and observe similar effects of alloying elements on γSFE as those in Pt23X or Ni23X.

# 1. Introduction

Advances in machine learning (ML) – especially those pertaining to finding insights or building predictive models from existing data and situations [1–5] – are making considerable impacts in materials research and discovery. Cooperation between ML predictions, first-principles calculations based on density functional theory (DFT), and experimental verification is emerging as a key paradigm to understand fundamentals, generate and analyze data, and design and discover materials. This paradigm is closely related to the data-centric approach depicted in the U.S. Materials Genome Initiative and the Materials 4.0 [6]. The process of exploring materials by machine learning has already been showcased in the design and discovery of catalysts [7,8], metallic glasses [9], lithium superionic conductors [10], and high-entropy alloys [11,12]; accurate predictions of crystal structures [13], band gaps [14,15], enthalpies of formation [16,17], and phase diagrams [18]; molecular dynamics simulations using ML-trained force fields [19]; on-the‐fly data analyses of high‐throughput experiments [20]; and gaining insights into creep properties of high-temperature alloys [21].

Machine learning approaches usually require large datasets for effective learning; however, in materials science we are often limited to thousands, hundreds, or even tens of high-quality data points for a target property [1,3,22]. Limited datasets may be handled well by regression algorithms in terms of supervised learning [1] to establish a mapping between the easily accessible representations (attributes) of a material and the target property [10]. It is evident that the more suitable the representations of a material, the more accurately an algorithm can map them to the property of interest [3]. Representations of a material are also referred to as “descriptors” (adopted in the present work) as well as “features”, “fingerprints”, “profiles”, or “predictors” [1,23]. Relevant to the present work, stacking fault energy (γSFE), which is not easily measured, is selected as a target property, aiming to demonstrate materials properties explored by supervised machine learning by combining target property, descriptors, and ML algorithms. In the present work, a set of descriptors have been analyzed to achieve a comprehensive understanding of γSFE based on the statistical significance of these descriptors to decrease the mean absolute error (MAE) in ML predictions; and suitable ML models have been produced by using an optimized combination of descriptors and ML algorithms.

The property of interest here, stacking fault energy, is a measure of energy penalty due to the interruption of stacking sequence in a crystal structure induced by phenomena such as shear deformation or crystal growth [24]. For example, the planes of the face-centered cubic (fcc) lattice may have a faulted stacking sequence of …*ABC|BCABC*…; where *A*, *B*, and *C* represent different planes and “|” stands for the stacking fault. The missing *A* plane at the stacking fault can be viewed as a slide of ( is the lattice parameter of the fcc lattice) of the plane along the direction [24–26]. The associated stacking fault energy, γSFE, is a fundamental material property, which can be used to understand and model many material properties and phenomena related to dislocations, plastic deformation, crystal growth, and phase transitions [24–28]. For instance, a lower γSFE corresponds to a larger distance between dislocation partials, a higher twin propensity [25], and a retardation of cross-slip and climb, hence a reduction of steady-state creep rate [26]. The γSFE values can also be used to understand and predict the increase of yield-strength with increasing temperature in, e.g., the Ni3Al alloys [27]. Despite advances in measuring stacking fault energy experimentally using such techniques as the weak-beam method of electron microscopy [29], an accurate value of γSFE is difficult to obtain for many reasons, such as, the small separation distance between partial dislocations in electron microscope images [30]. For example, Table 1 shows that experimental γSFE values are in the range of 110 ~ 280 mJ/m2 for fcc Al [31], 24 ~ 169 mJ/m2 for fcc Cu [31], 90 ~ 450 mJ/m2 for fcc Ni [31], and 45 ~ 322 mJ/m2 for fcc Pt [29,32,33]. On the other hand, DFT-based first-principles calculations have been extensively employed to predict γSFE. For instance, the γSFE values affected by dilute alloying elements were predicted in Al alloys [34], Mg alloys [35], and Ni alloys [25,26] in terms of the alias shear deformations [25,26,34,35]. On the basis of empirical knowledge (instead of machine learning analysis), these works indicated that (i) the γSFE values are closely related to atomic volumes of Al23X, Ni23X, and Mg95X as well as the position of alloying element X in the periodic table [26,34,35] and (ii) stacking fault energy can be understood qualitatively in terms of the (re)distribution and the shape of charge density [26,35]. These results suggest the elemental attributes of atomic number, volume, and valence electrons of alloying element X to correlate with stacking fault energy.

Recently considerable efforts have been performed on characteristic energies in terms of machine learning. For example, Cheng et al. [36] studied vacancy formation energy and its connection to bonding behavior in phase-change material GeTe by using an artificial neural network. Wang and Xiong [37] investigated the influence of alloying elements on γSFE of austenitic steels using an integrated thermodynamic modeling and machine learning approach. Wang et al. [38] predicted the generalized stacking fault energies and associated Peierls stresses in refractory metals (Mo, Nb, Ta, and W) based on machine learning based interatomic potentials. Vilalta et al. [39] examined the relationship between yield stress and the γSFE landscape in high entropy alloys by means of a machine learning approach. Arora and Aidhy [40] indicated that γSFE in concentrated multi-elemental alloys can be predicted using a machine learning based framework. However, the machine learning based study of γSFE affected by dilute alloying elements is not available in the literature.

The present work aims to develop insights into fundamental factors underlying stacking fault energy in terms of elemental attributes (i.e., the descriptors). The process includes first-principles calculations of γSFE values in Al-, Ni-, and Pt-based dilute alloys, an analysis of statistical significance regarding elemental descriptors via iterations between sequential descriptor selections and ML verifications, and the construction of predictive ML models. In the Sec. 2, we show the computational methodologies involving stacking fault energy predicted by first-principles calculations and various descriptors and ML models. In the following section (Sec. 3), we first present and discuss the computed stacking fault energies in 9 fcc elements (Ag, Al, Au, Cu, Ir, Ni, Pd, Pt, and Rh) and 85 dilute alloys of Al23X, Ni23X, and Pt23X; next we present the correlation analyses of γSFE through statistical significance of elemental descriptors; then we present and discuss the predictive ML models; and finally we present the outlier of DFT predictions in Al23X, Ni23X, and Pt23X, together with the ML-predicted γSFE values in fcc-based dilute alloys of Cu23X, Ag23X, Au23X, Rh23X, Pd23X, and Ir23X. It is worth mentioning that it is beyond the scope of the present work to build a predictive ML model to calculate stacking fault energy as a function of composition. However, the presently suggested descriptors are one of the fundamentals to build this ML model in addition to collecting and analyzing the values of stacking fault energy in the present work and especially in the literature, for example [35,41–50].

# 2. Methods

## 2.1. Stacking fault energy from first-principles calculations

Accurate stacking fault energy can be predicted by the following three methods: (i) the energy difference between the initial (perfect) structure and the stacking fault structure [51], (ii) the slab deformation of two adjacent planes [35,52], and (iii) the alias shear deformation of bulk materials [24,26,27,35,53]. Both the slab deformation approach and the alias shear approach can predict the general stacking fault energy as a response to shear deformation, including both the stable and the unstable stacking fault energies. In addition, the alias shear deformation can predict both ideal shear strength and stacking fault energy. Notably, the number of atoms employed in the supercell for the alias shear approach is usually only half of that required for the other two methods [24,26,35]. The alias shear deformation is hence selected in the present work to predict the stacking fault energy. In alias shear, only atoms in one plane are involved in shear deformation while the other atoms initially remain in their original positions; see the schematic diagram of alias shear in Figure 1a. Relaxations of the supercell are allowed after alias shear deformation, making shear displacement propagate through the supercell due to the interaction between the atoms during shear processes [24,27,35,53]. There are two schemes typically used to relax shear deformation. The first one is “pure shear”, indicating that all relaxations, including cell shape, cell volume, and atomic positions are allowed except for the fixed shear angle during the alias shear [24,27,35,53]. It has been shown that the alias shear with pure shear relaxations (i.e., the pure alias shear used in the present work) represents a deformation closer to actual shear processes [53]. The second relaxation scheme is “simple shear” (i.e., the simple alias shear used in the present work), indicating that only atomic positions are allowed to be relaxed.

For the facility of alias shear deformation along the close-packed direction in the close-packed plane in fcc, the conventional fcc lattice is represented by a 6-atom orthorhombic cell with its lattice vectors ***a*orth**, ***b*orth**, and ***c*orth** parallel to the , , and directions of the conventional fcc lattice, respectively. The lengths of ***a*orth**, ***b*orth**, and ***c*orth** are , , and , respectively. More details about this 6-atom orthorhombic cell can be found in the literature [24,26]. The alias shear deformation, along in the orthorhombic cell (i.e., along in the conventional fcc cell), can be expressed by [24,27,35,53],

|  |  |
| --- | --- |
| and | Eq. 1 |

where and **R** are the deformed and the undeformed matrices of lattice vectors in the orthorhombic cell, respectively. **T** is thedeformation matrix and *ε* the shear magnitude corresponding to the engineering shear strain, i.e., the ratio of shear displacement with respect to the height of the supercell [24,27,35,53]. Specifically, can be used to generate a stacking fault for ***c*orth** equal to .

In the present work, two kinds of orthorhombic supercells are employed: (i) the 6-atom orthorhombic cell used for fcc elements and (ii) the 24-atom (1×2×2) supercell with respect to the 6-atom orthorhombic cell for both the fcc elements and the Al23X, Ni23X, and Pt23X alloys, where X represents the dilute alloying element within the stacking fault plane to introduce the strongest effect on the value of stacking fault energy [46]. Table 2, as well as the supplemental csv (comma-separated value) files, gives a complete list of alloying elements (X’s) and Figure 1b shows the 24-atom orthorhombic supercell of Pt23X. Note that the mole fraction of X in the stacking fault plane is 0.25, which is a key concentration to regulate γSFE instead of the global concentration in dilute alloys [26,35]. Our previous tests using 24-atom up to 72-atom supercells indicated that the presently selected composition is still in the dilute limit, at least for most of the alloying elements [26,35].

## 2.2. DFT-based first-principles calculations

All DFT-based first-principles calculations in the present work were performed using the Vienna *Ab initio* Simulation Package (VASP) [54] for pure fcc elements and the Pt23X alloys. The γSFE values of Al23X and Ni23X were predicted previously [26,34] using the same approach adopted in the present work. The ion-electron interaction in the present work was described by the projector augmented wave method [55] and the exchange-correlation functional was depicted by the generalized gradient approximation (GGA-PBE) [56]. The recommended electronic configurations by VASP were adopted in the present work, for example, ten electrons (5d96s1) used as the valence for Pt and twelve (4s24p65s24d2) for Zr. Note that the semi-core s states are also treated as valence electrons for Zr, which is hence marked as Zr\_sv; see the similar notations for all other alloying elements used in Pt23X in Table 2. In the VASP calculations, the *k*-point meshes of 9×8×3 were used for the 24-atom supercell (or 12×21×8 for the 6-atom supercell). During relaxations of supercells, the precision was set to “PREC = Accurate”, and a cutoff energy of 520 eV was employed for the final calculations of total energies by the tetrahedron method incorporating a Blöchl correction [57]. The electronic degrees of freedom were converged to at least 10-6 eV/atom for all calculations.

After alias shear deformation, the cell shape was fixed but atomic positions were allowed to relax to lower the total energy, i.e., the aforementioned simple alias shear. For fcc elements, the pure alias shear deformations were also performed. These calculations were carried out by an external optimizer, GADGET, developed by Bučko *et al.* [58] to control VASP calculations of stresses and forces acting on each atom. All relaxed stress components were less than 0.15 GPa except for the one corresponding to the fixed shear angle (see the *ε* value in Eq. 1) and the forces acting on atoms were less than 0.03 eV*/*Å. More details about pure alias shear deformation can be found in our previous works [24,27,35].

For each structure with or without stacking fault, VASP calculations were performed on about eight volumes within the volume range of -10% < (*V* - *V*0)/*V*0 < +10%, where *V*0 is the equilibrium volume. A four-parameter Birch-Murnaghan equation of state (EOS) [59] was used to determine the lowest equilibrium energy based on the resulting volume versus energy data points. In addition, spin polarization was considered in DFT-based calculations for structures containing magnetic elements of Cr, Mn, Fe, Co, and Ni.

## 2.3. Supervised machine learning

In addition to the targeted properties, “descriptors” and “algorithm” are two key factors for supervised machine learning. Relevant to the present γSFE in Al23X, Ni23X, and Pt23X with the same fcc structure, the descriptors related to crystal structure and alloying concentration can be ignored, and we explore only the elemental attributes/representations for the sake of simplicity. By considering the elemental descriptors introduced by Seko et al. [60], Ward et al. [17], and our selections, the examined descriptors in the present work contain quantities related to:

* Periodic table: atomic number, period, group, Mendeleev number, and mass;
* Electronic structure: the filled and unfilled s, p, d, f, and total valence electrons;
* Atomic size: covalent radius, van der Waals atomic radius, atomic volume, and nonlocal pseudopotential radii for the s orbital and for the p orbital;
* Physical properties: the first-, the second-, and the third-ionization energies, thermal conductivity, electrical conductivity, electron density, electron affinity, electronegativity, number of spectral lines of the elements, and the maximum range of electrons in solid elements;
* Thermochemical properties: heat of fusion, heat of sublimation, heat of vaporization, cohesive energy, heat capacity, standard entropy, melting temperature, boiling temperature, Debye temperature, vacancy formation energy, and vacancy activity energy;
* Elastic properties: bulk modulus, shear modulus, and Young’s modulus;
* Host elements: Al, Ni, and Pt; and
* Space groups of pure elements at ambient conditions.

All of these 49 elemental descriptors are explained in the supplemental Table S1. Their values are given in the supplemental csv file based on the collections by Wolfram Mathematica [61,62]; data used in Miedema model [63,64], PubChem [65], Wikipedia [66], and Azom materials [67]; and data presented by Zunger [68], Samsonov [69], Dinsdale [70], Kittel [71], Shang et al. [72], and Anderson et al. [73]. It is worth mentioning that the characteristic energies of vacancy formation energy and vacancy activity energy for pure elements in fcc lattice, although they were determined by DFT-based calculations [72], are related to properties such as cohesive energy, the products of *G*0*V*0 and *B*0*V*0 (where *G*0 is the shear modulus, *B0* the bulk modulus, and *V*0 the equilibrium volume), and melting temperature [72]. These characteristic energies are also selected as candidates to correlate with stacking fault energy.

In practice, we adopt the relative values of these descriptors ( or ReDes),

|  |  |
| --- | --- |
|  | Eq. 2 |

where “Des” represents the value of descriptor. The reference state is the descriptor value of the host element, , except for the three descriptors representing the host elements Al, Ni, and Pt; see the values in the supplemental csv file. Our tests indicate that using the values could yield a better ML model than using the values directly. In addition to the descriptors, we examine a total of 21 automated regression algorithms in two of the mathematic software packages, i.e., the MATLAB (version R2019b) and the Wolfram Mathematica (WM, version 12.0). Table 4 summarizes these regression algorithms and their key hyperparameters, including, the linear regression (LR), various tree models, the support vector machine (SVM), the nearest neighbors (NN), the random forest (RF), the neural network (NNet), the Gaussian process regression (GPR), and the optimizable GPR (OptGPR). In the present work, special attention is paid to the GPR [74,75], which is a Bayesian nonparametric regression to generate distributions over functions. It is found that GPR works well on small datasets and has the ability to provide quantified uncertainty on the predictions as a function of the descriptors [74,76]. It should be mentioned that achieving a suitable ML model depends on the optimization of hyperparameters (i.e., the internal variables) in each algorithm. This process is performed in an automated fashion and the results depend on the employed software. As detailed in Results section, it is found that the GPR in WM is the worst performing algorithm for our study, but the GPR in MATLAB is the best one to study the present property of interest, i.e., γSFE.

We also employ the ASCENDS code [77] to perform correlation analysis by quantifying the relationship between the input descriptors and the variation of γSFE, using the Pearson’s correlation coefficient (PCC) [78] and the maximal information coefficient (MIC) [79]. Note that the PCC [78] represents a linear correlation between the descriptors and the γSFE. PCC has a value between 1 and -1, where the value 1 represents a total positive linear correlation, -1 a total negative linear correlation, and 0 no linear correlation[77]. The absolute value of PCC, i.e., |PCC|, is adopted in the present work. In principle, , where R2 is the familiar coefficient of determination to measure the goodness-of-fit for linear regression. In addition to |PCC|, the MIC [77,79] is also used to measure the strength of the linear or non-linear relationship between the descriptors and the γSFE. The MIC roughly equals to R2, and MIC has been used to measure the dependence of large datasets [79]. The MIC value lies between 0 and 1: the closer the coefficient to 1, the stronger the correlation between the two variables.

It should be remarked that both the PCC and the MIC measure the individual relationship between each descriptor and the targeted property. In order to examine the relationship between a subset of descriptors and the targeted property, we adopt iterations between sequential descriptor selection (SDS) and ML verifications of the ranked descriptors by SDS to gain insight into the statistical significance of the descriptors, see details listed in Table 3.

# 3. Results and Discussion

## 3.1. DFT-based results of stacking fault energy (γSFE)

To examine the capability to predict γSFE by DFT-based alias shear deformation, Table 1 summarizes the presently computed γSFE values for 9 fcc elements (Ag, Al, Au, Cu, Ir, Ni, Pd, Pt, and Rh) in comparison with experimental data selected by de Campos [29], Hirth and Lothe [32], Smallman et al. [33], Lee et al. [80], and Zhang et al. [31], as well as calculated results from the DFT-based climbing-image nudged elastic band method [81]. It is found that relaxation schemes and supercell sizes influence significantly the γSFE value (> 10% in some cases) [35], while the pure alias shear could result in a reliable γSFE due partially to its close resemblance to the actual shear process [53]. For test purposes, the present calculations were performed in three cases: (i and ii) pure alias shear in terms of 6-atom supercells and 24-atom supercells, and (iii) simple alias shear with relaxations of atomic positions only using 24-atom supercells (RX24). Table 1 shows that the calculated γSFE values by pure shears ( and in terms of the 6- and 12-atom supercells, respectively) match well with each other (< 5 mJ/m2, except for Al where the difference is 9 mJ/m2). The recommended γSFE values in the present work, labeled as , of fcc elements are the average from the reliable and results; see the values in Table 1. These values agree with the calculated values from RX24 (< 5 mJ/m2, except for Pt with a difference of 11 mJ/m2) as well as from Jin et al. [81] (< 16 mJ/m2 with the worst ones being for Al, Ir, and Pt). These values also agree with some of the selected experimental data such as 128 vs. 130 mJ/m2 for Al and 131 vs. 125 mJ/m2 for Ni (see Table 1). This indicates the capability of DFT-based alias shear deformation to accurately predict γSFE. Based on the present DFT calculations, we believe the following experimental results have much lower reliabilities, i.e., 450 mJ/m2 for Ni [33], 45 mJ/m2 and 95 mJ/m2 for Pt [29,33], and 750 mJ/m2 and 330 mJ/m2 for Rh [32,33]; see Table 1.

Figure 2 illustrates the present γSFE values of Pt23X from DFT-based calculations as a function of equilibrium volume of Pt23X and atomic number of X along different rows in the periodic table. These γSFE values are also listed in the supplemental csv file. Figure 2a shows that the γSFE values of Pt23X decrease with increasing volume. Figure 2b shows that the γSFE values of Pt23X increase and then decrease with increasing atomic number of X along different rows, where the maximum γSFE values are observed in the alloys of Pt23Ni, Pt23Rh, and Pt23Ir for the 3d, 4d, and 5d alloying elements, respectively. The similar trends were also observed in our previous calculations for Al23X [34] and Ni23X [26]; indicating the critical role of atomic volume and atomic number of alloying elements to regulate γSFE.

For the facility of the machine learning study, stacking fault energy, γSFE, is expressed as follows,

|  |  |
| --- | --- |
|  | Eq. 3 |

where is the stacking fault energy of host element with its value given in Table 1, and the relative stacking fault energy affected by alloying element X. The supplemental csv file lists of the and values for Al23X, Ni23X, and Pt23X with 85 high-quality data points. The absolute values of are in the range of zero to hundreds mJ/m2. They (see also Table 2) indicate that the VIII group alloying elements Co (for Al23X) and Ni (for both Ni23X and Pt23X) increase greatly the values, while the farther the alloying elements are from Co and Ni in the periodic table, the smaller the values will be.

## 3.2. Correlation analysis of γSFE via statistical significance of descriptors

A simple way to understand (or γSFE) is using the correlation between the values of each elemental descriptor and the values. Figure S1 in the supplemental material shows correlation analyses of the values using the absolute value of the Pearson’s correlation coefficient, |PCC|, as well as the maximal information coefficient, MIC. The values of |PCC| and MIC are also shown in the supplemental Table S1. As expected, atomic size shows the largest impact to correlate ΔγSFE. The |PCC| values (as well the MIC values) show that the relative covalent radius (|PCC| = 0.790) is the most correlated descriptor to predict stacking fault energy affected by alloying elements; see also the plot in Figure 3. Except for the top descriptor of covalent radius, the descriptor rankings based on the |PCC| and MIC values are different to a large extent. For example, the second top descriptor is atomic volume (V0\_Miedema) and Mendeleev number (M\_Num2) based on |PCC| and MIC, respectively; see the supplemental Table S1. We hence examine the ranked descriptors by |PCC| and MIC using one of the best ML algorithms, GPR4 (see details below). Supplemental Figure S2 shows an irregular decrease or increase of the average MAE values in prediction of when adding the ranked descriptors one-by-one. This indicates that it is inappropriate to select descriptors for construction of ML models based on the scores by |PCC| or MIC.

It is self-evident that an optimal (sub)set of descriptors should result in the lowest MAE in the prediction of a targeted property, i.e., any additional or fewer descriptors would cause the MAE to increase. To this end, iterations between sequential descriptor selection (SDS) and ML verification of the ranked descriptors by SDS have been performed to quantify the statistical significance of the descriptors, see details in Table 3. Each iteration results in one generation of the ranked descriptors. Here, the statistical significance is defined as the frequency (probability) of occurrence for a given descriptor in the SDS process. It is found that the first generation of the ranked descriptors cannot achieve the optimal subset of descriptors (see the supplemental Figure S2). After a few iterations as detailed in Table 3, the final generation of the ranked descriptors is listed in the supplemental Table S1. Figure 4 illustrates the MAE values in ML predictions by sequentially adding descriptors according to this new final ranking; showing a monotonic decreasing and then a monotonic increasing of the MAE values. This V-type curve has a lowest MAE value corresponding to the optimal subset of descriptors, i.e., the top 14 descriptors of ReDes14 as highlighted by blue and especially by green in Figure 4 (see also the supplemental Table S1 and the sheet of “ReDes14” in the supplemental csv file).

In addition to the top descriptor of covalent radius, Figure 4 shows that the numbers of p, d, and total valence electrons have an unexpected but significant impact on the values. The top 8 descriptors (colored by green in Figure 4) also include the group of elements in the periodic table, heat of sublimation, shear modulus, and vacancy formation energy of pure elements. Note that these top 8 descriptors were determined directly by the SDS method (i.e., Step 1 in Table 3) and verified by the ML predictions (i.e., Step 2 in Table 3). In addition to the top 8 descriptors, the optimal subset of descriptors (i.e., the ReDes14) includes 6 less significant descriptors, colored by blue in Figure 4. They are melting temperature, electron density, electronegativity, standard entropy at 298 K of pure elements, and the period and number of pure elements in the periodic table.

These 14 descriptors in ReDes14 highlight the relationship between stacking fault energy and both well-known and less-known fundamental properties. For example, it is believed that shear modulus is closely related to the determination of stacking fault energy from experiments [33],

|  |  |
| --- | --- |
|  | Eq. 4 |

where *G* is the shear modulus for the material of study, *b* the Burgers vector, and *r* the curvature radius of the dislocation nodes [29]. However, among the properties of *G*, *b*, and *r*, the Burgers vector (related to atomic size) is expected to be more important to determine since the top descriptor is covalent radius (see Figure 3).

The descriptors in ReDes14 related to electronic structure of materials include the numbers of p, d, and total valence electrons as well as the positions of pure elements in the periodic table (the group, period, and number). In fact, stacking fault energy as a result of alias shear deformation can be understood qualitatively using the density, shape, distribution, and re-distribution of valence charge density [26,35,82,83]. For example, non-spherical distribution of valence charge density hinders shear deformation, resulting in larger stacking fault energy [82]; anisotropic valence charge density leads to elastic anisotropy, as shown for fcc Al [83]; and re-distribution of (magnetic) charge density in Ni23X and Mg95X due to alloying element X corresponds to the variation of stacking fault energy [26,35].

Two of the three descriptors used in the Miedema model [63,64] to predict thermochemical properties are also important to correlate : electronegativity and electron density, with the third, atomic volume, being less important. Note that covalent radius is the top descriptor in the present ReDes14, making the other descriptors related to atomic size redundant and excluded as a result. For example, atomic volume, Van der Waals atomic radius, and nonlocal pseudopotential radius, do not show up in ReDes14. In addition, the electronegativity employed in the Miedema model [63] is more significant in the present work than the other electronegativities such as the one of Pauling scale (see the supplemental Table S1).

Statistical significance of the descriptors in ReDes14 indicates that is influenced mainly by covalent radius (instead of other descriptors related to atomic size). The other key contributions to regulate include the descriptors related to (i) electronic structure such as the numbers of p, d, and total valence electrons; (ii) shear modulus; (iii) properties to predict thermochemical properties such as electronegativity and electron density employed in the Miedema model [63,64]; and (iv) heat of sublimation and its closely related vacancy formation energy and melting temperature. In particular, the present work indicates that the combined effect of descriptors is different from the effect of individual descriptors; and only one descriptor in the same category (e.g., covalent radius in the category of atomic size) should be included in the optimal subset of descriptors.

### 3.3. Statistically best ML algorithms

Figure 5 shows the average MAE values to predict using various ML algorithms in MATLAB and WM (see Table 4) in terms of the optimal subset/combination of descriptors in ReDes14 and the cross validation method. Due to the low volume of data available, the random selection of test data is often unrepresentative of the whole set, and therefore each ML model has been tested multiple times to get the average MAE values, i.e., 1000 times for algorithms in MATLAB with 5-fold cross validation and 50 times for algorithms in WM with 10-fold cross validation (for test purpose). Note that the reported NNet model performance corresponds to validation set error from 10 neural network models automatically designed and trained based on a random training set selection from the dataset. The automated design parameters have been set to a performance goal of "Quality" (prioritizing accuracy, rather than training speed and memory use) and a time goal of 10 standardized central processing unit (CPU) seconds (about 20 seconds on our desktop computer). Each resulting model has been different due to a different starting point (training data sample). A common example was a relatively small neural network with 40 thousand double-precision trainable parameters (0.3MB) distributed over about 4-8 layers.

Figure 5 indicates that the GPR algorithms (i.e., GPR2 and GPR4) in MATLAB give the lowest MAE values (< 8 mJ/m2) followed by the SVM algorithms (i.e., SVM2 and SVM3 with MAE < 10 mJ/m2). It is also found that the optimizable algorithms as implemented in MATLAB cannot reduce the MAE values further based on our tests, especially for the Tree and SVM algorithms. Regarding the algorithms in WM, Figure 5 shows that the neural network (WM\_NNet) gives a low MAE value (< 9 mJ/m2), followed by the WM\_GBT and WM\_LR algorithms (MAE > 16 mJ/m2). The ML algorithms of WM\_NNet and especially the Gaussian process regressions in MATLAB (GPR2, GPR4, and OptGPR) are the present ML models to predict . Figure 5 shows also the Gaussian process algorithm is the worst one in WM, but the Gaussian process algorithm is the best one in MATLAB. This conclusion indicates that the hyperparameters (i.e., the internal variables) in each algorithm and the schemes to optimize them have a significant influence on the final MAE values.

As an example, Figure 6 illustrates the effect of k-fold value, i.e., how many times random subset of data used is chosen and used for cross validation, on the values of average R2, average MAE, average MaxAE (the maximum absolute error), and their standard derivations and range of uncertainties (i.e., from the maximum to the minimum value). Note that the minimum k-fold value is 2, corresponding to using half of the data points for training and half for predictions. In addition, the calculations were performed 1000 times using the GPR2 algorithm and the descriptors in ReDes14. With increasing the k-fold value (i.e., data points for training), Figure 6b and c show that both the average MAE and MaxAE values (12 and 45 mJ/m2, respectively) as well as their standard derivations and uncertainties decrease to less than 8 mJ/m2 and 30 mJ/m2, respectively, when k-fold > 8, i.e., 12.5% or less data used for cross validation. Figure 6a shows that the average R2 value increases from 0.88 (k-fold = 2) to 0.95 and above (k-fold > 8). Note that using only the top one descriptor of covalent radius, the obtained R2 = 0.62 (see Figure 3); indicating the combined descriptors of ReDes14 can predict well stacking fault energy.

Table 5 summarizes the average MAE values together with their standard deviations when k-fold = 5, 10, or 30 for the ML algorithms of GPR2, GPR4, OptGPR, and WM\_NNet, showing the average MAE values when k-fold ≥ 10 are less than 8 mJ/m2 for the GPR models and less than 9 mJ/m2 for the WM\_NNet model. These MAE values (8 or 9 mJ/m2) are 8% and less than the changes of the values (up to 270 mJ/m2 for Al23X, 107 mJ/m2 for Ni23X, and 137 mJ/m2 for Pt23X; see the supplemental csv file) and are 7% and less than the stacking fault energies of the host elements of Al (128 mJ/m2), Ni (131 mJ/m2), and Pt (272 mJ/m2); see Table 1.

*3.4. Statistical outlier of DFT predictions of γSFE in Al23X, Ni23X, and Pt23X*

Here, we define that the worst outlier of DFT predictions of γSFE possesses the maximum absolute error (MaxAE) between the DFT-predicted γSFE using the alias shear deformation and the ML-predicted γSFE using the cross-validation method. Based on the MaxAE values, we can identify and search the possible outliers of DFT predictions, which do not follow the broader trends found by ML-based correlation analysis as shown in Section 3.2.  These DFT outliers may be caused by unstable structures used in the calculations or relaxations to a local instead of a global minimum. Very likely these DFT calculations are less reliable, and it is necessary to re-examine these calculations and/or discard these DFT results. For these cases, we conclude that either the ML predictions are more accurate and reliable than the DFT calculations or both the ML predictions and DFT calculations are questionable.

Figure 7 illustrates the top 20 worst outliers of DFT predictions of γSFE (or ) ranked by the occurrence probability ofthe worst outliers. Here, we perform a total of 87,000 ML predictions using the algorithms of GPR2, GPR4, and OptGPR with k-fold values from 2 to 30, and training of 1000 times for each ML algorithm and each k-fold value (i.e., 3 ML models × 29 k-fold values × 1000 trainings = 87,000), in order to examine the first to the fourth worst outliers of DFT vs. ML results in terms of the top four MaxAE values in each ML prediction (total number of worst outliers = 4 × 87,000 = 348,000). Figure 7 shows that the worst outlier of DFT calculations is for Al23La with a worst probability *p*worst = 0.25 (note that the total *p*worst = 1), which is defined as the number of worst outliers of DFT vs. ML results for Al23La (86,279 times)over the total number of worst outliers of DFT vs. ML results (348,000 times). The following worst outliers of DFT predictions with *p*worst > 0.1 are for Pt23Au, Ni23Co, and Al23Be; see Figure 7. In addition, the alloys with 0.1 > *p*worst > 0.01 include Ni23Zn, Pt23Y, Al23Zn, Al23Co, Al23Sr, Al23Mn, Pt23Sc, and Al23Pb. The rest 73 DFT calculations which are shown or not shown in Figure 7 can be claimed as the best DFT calculations with *p*worst < 0.01. In addition, Figure 7 indicates that the stacking fault energy values predicted by DFT calculations are more accurate for Ni23X, followed by Al23X and Pt23X, according to the probabilities of worst outliers of DFT predictions with *p*worst > 0.01.

### 3.5. Machine learning results in Al23X, Ni23X, Pt23X, and beyond

Table 2 shows both the DFT- and the ML-predicted values for the alloying elements of interest. Here, the ML-predictions include 13 new alloying elements in Al23X, 4 in Ni23X, and 2 in Pt23X. The reported values are the average based on the four ML algorithms of GPR2, GPR4, OptGPR, and WM\_NNet (see Figure 5). The complete ML predictions of are given in the supplemental csv file. The average MAE values of these predictions (less than 9 mJ/m2) are reported in Table 5. In addition, Figure 6 shows the complete analyses of MAE and MaxAE for GPR2, one of the best algorithms for these data. Table 2 also presents the values of standard derivation (SD) for the ML-predicted from GPR2, GPR4, OptGPR, and WM\_NNet. The larger the standard derivation between runs, the lower the quality of ML predictions will be, as high values might be an indication of overfitting to the training data. It is shown that Al23Os and Al23Ir might have the worst ML predictions (SD values > 10 mJ/m2), followed by Al23Re, Ni23Ca, and Ni23Au (SD values > 5 mJ/m2). For the other dilute alloys in Table 2, the ML-predicted values for new alloys have smaller standard derivation (SD values < 5 mJ/m2).

The present focus is correlation analysis of stacking fault energy in dilute Al-, Ni-, and Pt-based alloys, and the present ML models (GPR2, GPR4, OptGPR, and WM\_NNet) are suggested to study these alloys. However, it is interesting to examine the present ML models to predict stacking fault energy in other fcc-based alloys beyond Al23X, Ni23X, and Pt23X. As a test, Figure 8 illustrates the capability of the present ML models (demonstrated using OptGPR) to predict the values of new alloys by means of a method similarly to the 3-fold cross validation. That is, the predicted values of Al23X are based on the OptGPR model trained using the data in Ni23X and Pt23X; and so on. Figure 8 shows that the predicted values in Al23X have the largest MEA value (34.8 mJ/m2; see also Table 6). The largest MEA value in Al23X is due mainly to the less accurate ML predictions of in, for example, (i) Al23La, Al23Sr, and Al23K (see Figure 8), where the alloying elements La, Sr, and K are excluded in Ni23X and Pt23X for ML training; and (ii) Al23Co, Al23Fe, Al23Mn, and Al23Ni with magnetic alloying elements (Co, Fe, Mn, and Ni) and their DFT predictions show outliers (see Figure 7). In addition, the host element Al is a simple metal while the host elements (Ni and Pt) used in ML training are transition metals. For the ML predicted values in Ni23X and Pt23X, the MEA values are relatively small (16.7 and 17.8 mJ/m2, respectively; see Table 6) due partially to the facts that (i) the alloying elements are included in both the training and the predicting alloys and (ii) both the host elements Ni and Pt are transition metals. It is interesting to find that the largest outliers are for Ni23Y and Pt23Y, where the element Y is usually excluded in both the Ni- and Pt-based alloys.

Figure 9a illustrates the DFT-calculated values of Pt23X in comparison with those of Ni23X from both DFT calculations and ML predictions (demonstrated using the OptGPR model, i.e., the same values as shown in Figure 8). It can be seen that (i) the values of Ni23X predicted by both DFT and OptGPR are in the similar region and (ii) there exist good linear relationships between the values of Pt23X and Ni23X with R2 > 0.85 (see Figure 9a); indicating that the influence of alloying elements on stacking fault energy has similar trends in Ni and Pt, which are both transition metals.

Figure 9b illustrates the DFT-calculated values of Pt23X compared with those from ML predictions (demonstrated using the OptGPR model, more results given in the Supplemental csv file) for fcc-based dilute alloys of Al23X, Ni23X, Pt23X, Cu23X, Ag23X, Au23X, Rh23X, Pd23X and Ir23X; where the alloying elements X for the new alloys are the same as those in the Pt23X (see Figure 2). The values of stacking fault energy for the host elements, , are shown in Table 1. The ML models are trained using the values of Al23X, Ni23X, and Pt23X for the new dilute alloys (i.e., except for Al23X, Ni23X, and Pt23X with their results in Figure 8). The DFT-predicted values of Pt23X show a good linear relationship with R2 = 0.81 with respect to the predicted values from the OptGPR model; see Figure 9b. It is expected that the MAE values of the ML predicted in Cu23X, Ag23X, Au23X, Rh23X, Pd23X and Ir23X are less than 17 mJ/m2 according to the MAE values in Table 6, since (i) two transition metal alloys (Ni23X and Pt23X) are employed in the training datasets, and (ii) all new host elements (Cu, Ag, Au, Rh, Pd, and Ir) are transition metals.

# 4. Summary

The present work yields valuable insights into fundamental factors underlying stacking fault energy (γSFE) in terms of supervised machine learning (ML). We perform a comprehensive study of γSFE using the ML approach, including, (a) 85 data points in Al23X, Ni23X, and Pt23X with 43 solute elements (X’s) from first-principles calculations using the alias shear approach; (b) 49 elemental descriptors, including the quantities related to the periodic table, atomic sizes, physical properties, thermochemical properties, elastic properties, electronic structure, and host elements; and (c) 21 regression algorithms in the two mathematical software programs of MATLAB (version R2019b) and Wolfram Mathematica (version 12.0).

Illustrated with stacking fault energy, this present work provides a pathway to analyze statistical significance of descriptors, statistically best ML algorithms, and statistical outlier of DFT calculations. It indicates that (i) the optimal subset of descriptors to regulate γSFE includes 14 elemental attributes, i.e., the ReDes14, according to statistical significance of descriptors to decrease the MAE value in ML predictions; (ii) the top descriptor to regulate γSFE is covalent radius, followed by the unexpected influence related to electronic structure such as the numbers of p and d valence electrons; (iii) the alloys with alloying elements close to Ni and Co in the periodic table possess higher γSFE values; (iv) the top four worst outliers of DFT predictions of γSFE are for Al23La, Pt23Au, Ni23Co, and Al23Be based on the analyses between DFT calculations and ML predictions; and (v) in the case of the present γSFE, the best ML models are the Gaussian process regressions (GPR2, GPR4, and OptGPR) with an average MAE < 8 mJ/m2 (the goodness-of-fit R2 > 0.95) when k-fold ≥ 10 for cross validation.

Beyond detailed investigation in Al23X, Ni23X, and Pt23X, the present ML models are also employed to predict stacking fault energies in fcc-based dilute alloys of Cu23X, Ag23X, Au23X, Rh23X, Pd23X, and Ir23X. It is expected that the MAE values to predict stacking fault energies in new transition metal alloys are less than 17 mJ/m2; and the influence of alloying elements on stacking fault energy possesses similar trends for each host element (at least for transition metals) in comparison with those on, for example, Pt23X (or Ni23X).

**Acknowledgements**

This work was financially supported by the Rare and Precious Metals Material Genetic Engineering Project of Yunnan Province (2020). The authors at the Pennsylvania State University (Penn State) acknowledge partial support from the U. S. Department of Energy (DOE) via Award Nos. DE-FE0031553 and DE-EE0008456, and the Seed Grant from Institute for CyberScience at Penn State. We also would like to thank Dr. Sangkeun Lee at the Oak Ridge National Laboratory for his help on the ASCENDS code.

**References**

[1] Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A and Kim C 2017 Machine learning in materials informatics: recent applications and prospects *npj Comput. Mater.* **3** 54

[2] Mueller T, Kusne A G and Ramprasad R 2016 Machine Learning in Materials Science *Reviews in Computational Chemistry* vol 29 (Wiley) pp 186–273

[3] Butler K T, Davies D W, Cartwright H, Isayev O and Walsh A 2018 Machine learning for molecular and materials science *Nature* **559** 547–55

[4] Schmidt J, Marques M R G, Botti S and Marques M A L 2019 Recent advances and applications of machine learning in solid-state materials science *npj Comput. Mater.* **5** 83

[5] Ward L, Agrawal A, Choudhary A and Wolverton C 2016 A general-purpose machine learning framework for predicting properties of inorganic materials *npj Comput. Mater.* **2** 16028

[6] Jose R and Ramakrishna S 2018 Materials 4.0: Materials big data enabled materials discovery *Appl. Mater. Today* **10** 127–32

[7] Goldsmith B R, Esterhuizen J, Liu J-X, Bartel C J and Sutton C 2018 Machine learning for heterogeneous catalyst design and discovery *AIChE J.* **64** 2311–23

[8] Ulissi Z W, Tang M T, Xiao J, Liu X, Torelli D A, Karamad M, Cummins K, Hahn C, Lewis N S, Jaramillo T F, Chan K and Nørskov J K 2017 Machine-Learning Methods Enable Exhaustive Searches for Active Bimetallic Facets and Reveal Active Site Motifs for CO 2 Reduction *ACS Catal.* **7** 6600–8

[9] Ren F, Ward L, Williams T, Laws K J, Wolverton C, Hattrick-Simpers J and Mehta A 2018 Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments *Sci. Adv.* **4** eaaq1566

[10] Fujimura K, Seko A, Koyama Y, Kuwabara A, Kishida I, Shitara K, Fisher C A J, Moriwake H and Tanaka I 2013 Accelerated Materials Design of Lithium Superionic Conductors Based on First-Principles Calculations and Machine Learning Algorithms *Adv. Energy Mater.* **3** 980–5

[11] Pei Z, Yin J, Hawk J A, Alman D E and Gao M C 2020 Machine-learning informed prediction of high-entropy solid solution formation: Beyond the Hume-Rothery rules *npj Comput. Mater.* **6** 50

[12] Kaufmann K, Maryanovsky D, Mellor W M, Zhu C, Rosengarten A S, Harrington T J, Oses C, Toher C, Curtarolo S and Vecchio K S 2020 Discovery of high-entropy ceramics via machine learning *npj Comput. Mater.* **6** 42

[13] Hautier G, Fischer C C, Jain A, Mueller T and Ceder G 2010 Finding Nature’s Missing Ternary Oxide Compounds Using Machine Learning and Density Functional Theory *Chem. Mater.* **22** 3762–7

[14] Lee J, Seko A, Shitara K, Nakayama K and Tanaka I 2016 Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques *Phys. Rev. B* **93** 115104

[15] Pilania G, Gubernatis J E and Lookman T 2017 Multi-fidelity machine learning models for accurate bandgap predictions of solids *Comput. Mater. Sci.* **129** 156–63

[16] Faber F, Lindmaa A, von Lilienfeld O A and Armiento R 2015 Crystal structure representations for machine learning models of formation energies *Int. J. Quantum Chem.* **115** 1094–101

[17] Ward L, Liu R, Krishna A, Hegde V I, Agrawal A, Choudhary A and Wolverton C 2017 Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations *Phys. Rev. B* **96** 024104

[18] Long C J, Hattrick-Simpers J, Murakami M, Srivastava R C, Takeuchi I, Karen V L and Li X 2007 Rapid structural mapping of ternary metallic alloy systems using the combinatorial approach and cluster analysis *Rev. Sci. Instrum.* **78** 072217

[19] Jinnouchi R, Karsai F and Kresse G 2019 On-the-fly machine learning force field generation: Application to melting points *Phys. Rev. B* **100** 014105

[20] Kusne A G, Gao T, Mehta A, Ke L, Nguyen M C, Ho K M, Antropov V, Wang C Z, Kramer M J, Long C and Takeuchi I 2014 On-the-fly machine-learning for high-throughput experiments: Search for rare-earth-free permanent magnets *Sci. Rep.* **4** 1–7

[21] Shin D, Yamamoto Y, Brady M P, Lee S and Haynes J A 2019 Modern data analytics approach to predict creep of high-temperature alloys *Acta Mater.* **168** 321–30

[22] Zhang Y and Ling C 2018 A strategy to apply machine learning to small datasets in materials science *npj Comput. Mater.* **4** 25

[23] Pilania G, Wang C, Jiang X, Rajasekaran S and Ramprasad R 2013 Accelerating materials property predictions using machine learning *Sci. Rep.* **3** 1–6

[24] Shang S L, Wang W Y, Wang Y, Du Y, Zhang J X, Patel A D and Liu Z K 2012 Temperature-dependent ideal strength and stacking fault energy of fcc Ni: A first-principles study of shear deformation *J. Phys. Condens. Matter* **24** 155402

[25] Shang S, Wang Y, Du Y, Tschopp M A and Liu Z-K 2014 Integrating computational modeling and first-principles calculations to predict stacking fault energy of dilute multicomponent Ni-base alloys *Comput. Mater. Sci.* **91** 50–5

[26] Shang S L, Zacherl C L, Fang H Z, Wang Y, Du Y and Liu Z K 2012 Effects of alloying element and temperature on the stacking fault energies of dilute Ni-base superalloys *J. Phys. Condens. Matter* **24** 505403

[27] Shang S L, Shimanek J, Qin S, Wang Y, Beese A M and Liu Z K 2020 Unveiling dislocation characteristics in Ni3Al from stacking fault energy and ideal strength: A first-principles study via pure alias shear deformation *Phys. Rev. B* **101** 024102

[28] An X H, Wu S D, Wang Z G and Zhang Z F 2019 Significance of stacking fault energy in bulk nanostructured materials: Insights from Cu and its binary alloys as model systems *Prog. Mater. Sci.* **101** 1–45

[29] de Campos M F 2008 Selected values for the stacking fault energy of face centered cubic metals *Mater. Sci. Forum* **591**–**593** 708–11

[30] Qi Y and Mishra R K 2007 Ab initio study of the effect of solute atoms on the stacking fault energy in aluminum *Phys. Rev. B* **75** 224105

[31] Zhang X, Grabowski B, Körmann F, Ruban A V., Gong Y, Reed R C, Hickel T and Neugebauer J 2018 Temperature dependence of the stacking-fault Gibbs energy for Al, Cu, and Ni *Phys. Rev. B* **98** 224106

[32] Hirth J P and Lothe J 1992 *Theory of dislocations* (Krieger Pub. Co)

[33] Smallman R E, Dillamore I L and Dobson P S 1966 The measurement of stacking fault energy *Le J. Phys. Colloq.* **27** C3-86-C3-93

[34] Gao Q, Zhang H, Yang R, Fan Z, Liu Y, Wang J, Geng X, Gao Y, Shang S, Du Y and Liu Z 2018 Effect of alloying elements on the stacking fault energies of dilute Al-based alloys *J. Min. Metall. Sect. B Metall.* **54** 185–96

[35] Shang S L, Wang W Y, Zhou B C, Wang Y, Darling K A, Kecskes L J, Mathaudhu S N and Liu Z K 2014 Generalized stacking fault energy, ideal strength and twinnability of dilute Mg-based alloys: A first-principles study of shear deformation *Acta Mater.* **67** 168–80

[36] Cheng Y X, Zhu L, Wang G, Zhou J, Elliott S R and Sun Z 2020 Vacancy formation energy and its connection with bonding environment in solid: A high-throughput calculation and machine learning study *Comput. Mater. Sci.* **183** 109803

[37] Wang X and Xiong W 2020 Stacking fault energy prediction for austenitic steels: thermodynamic modeling vs. machine learning *Sci. Technol. Adv. Mater.* **21** 626–34

[38] Wang X, Xu S, Jian W R, Li X G, Su Y and Beyerlein I J 2021 Generalized stacking fault energies and Peierls stresses in refractory body-centered cubic metals from machine learning-based interatomic potentials *Comput. Mater. Sci.* **192** 110364

[39] Vilalta P C, Sheikholeslami S, Saleme Ruiz K, Yee X C and Koslowski M 2021 Machine Learning for Predicting the Critical Yield Stress of High Entropy Alloys *J. Eng. Mater. Technol. Trans. ASME* **143**

[40] Arora G and Aidhy D S 2020 Machine Learning Enabled Prediction of Stacking Fault Energies in Concentrated Alloys *Metals (Basel).* **10** 1072

[41] Yu H, Cao S, Youssef S S, Ma Y-J, Lei J-F, Qi Y, Hu Q-M and Yang R 2021 Generalized stacking fault energies and critical resolved shear stresses of random α-Ti-Al alloys from first-principles calculations *J. Alloys Compd.* **850** 156314

[42] Dou Y, Luo H, Zhang J and Tang X 2019 Generalized Stacking Fault Energy of {10-11}<11-23> Slip System in Mg-Based Binary Alloys: A First Principles Study *Materials (Basel).* **12** 1548

[43] Qiu S, Zhang X-C, Zhou J, Cao S, Yu H, Hu Q-M and Sun Z 2020 Influence of lattice distortion on stacking fault energies of CoCrFeNi and Al-CoCrFeNi high entropy alloys *J. Alloys Compd.* **846** 156321

[44] Zhang Y, Guo J, Chen J, Wu C, Kormout K S, Ghosh P and Zhang Z 2019 On the stacking fault energy related deformation mechanism of nanocrystalline Cu and Cu alloys: A first-principles and TEM study *J. Alloys Compd.* **776** 807–18

[45] Xu G, Chong X, Zhou Y, Wei Y, Hu C, Zhang A, Zhou R and Feng J 2020 Effects of the alloying element on the stacking fault energies of dilute Ir-based superalloys: A comprehensive first-principles study *J. Mater. Res.* **35** 2718–25

[46] Cai T, Li K Q, Zhang Z J, Zhang P, Liu R, Yang J B and Zhang Z F 2020 Predicting the variation of stacking fault energy for binary Cu alloys by first-principles calculations *J. Mater. Sci. Technol.* **53** 61–5

[47] Chen S, Wang Q, Liu X, Tao J, Wang J, Wang M and Wang H 2020 First-principles studies of intrinsic stacking fault energies and elastic properties of Al-based alloys *Mater. Today Commun.* **24** 101085

[48] Dong Z, Li W, Chai G and Vitos L 2020 Strong temperature – Dependence of Ni -alloying influence on the stacking fault energy in austenitic stainless steel *Scr. Mater.* **178** 438–41

[49] Zhao B, Huang P, Zhang L, Li S, Zhang Z and Yu Q 2020 Temperature Effect on Stacking Fault Energy and Deformation Mechanisms in Titanium and Titanium-aluminium Alloy *Sci. Rep.* **10** 3086

[50] Hu P, Zhao W and Sun Z 2019 Vacancy effect on the generalized stacking fault energy of alloyed γ-Ni system: A first-principles study *Comput. Mater. Sci.* **166** 187–92

[51] Wang W Y, Shang S L, Wang Y, Mei Z-G, Darling K A, Kecskes L J, Mathaudhu S N, Hui X D and Liu Z-K 2014 Effects of alloying elements on stacking fault energies and electronic structures of binary mg alloys: A first-principles study *Mater. Res. Lett.* **2** 29–36

[52] Han J, Su X M, Jin Z H and Zhu Y T 2011 Basal-plane stacking-fault energies of Mg: A first-principles study of Li- and Al-alloying effects *Scr. Mater.* **64** 693–6

[53] Jahnatek M, Hafner J, Krajci M, Jahnátek M, Hafner J and Krajčí M 2009 Shear deformation, ideal strength, and stacking fault formation of fcc metals: A density-functional study of Al and Cu *Phys. Rev. B* **79** 224103

[54] Kresse G and Furthmüller J 1996 Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set *Phys. Rev. B* **54** 11169–86

[55] Kresse G and Joubert D 1999 From ultrasoft pseudopotentials to the projector augmented-wave method *Phys. Rev. B* **59** 1758–75

[56] Perdew J P, Burke K and Ernzerhof M 1996 Generalized Gradient Approximation Made Simple *Phys. Rev. Lett.* **77** 3865–8

[57] Blöchl P E, Jepsen O and Andersen O K 1994 Improved tetrahedron method for Brillouin-zone integrations *Phys. Rev. B* **49** 16223–33

[58] Bučko T, Hafner J and Ángyán J G 2005 Geometry optimization of periodic systems using internal coordinates *J. Chem. Phys.* **122** 124508

[59] Shang S-L, Wang Y, Kim D and Liu Z-K 2010 First-principles thermodynamics from phonon and Debye model: Application to Ni and Ni3Al *Comput. Mater. Sci.* **47** 1040–8

[60] Seko A, Togo A and Tanaka I 2018 Descriptors for Machine Learning of Materials Data *Nanoinformatics* (Singapore: Springer Singapore) pp 3–23

[61] Anon Wolfram Mathematica: Modern Technical Computing

[62] Anon The Periodic Table

[63] Zhang R F, Zhang S H, He Z J, Jing J and Sheng S H 2016 Miedema Calculator: A thermodynamic platform for predicting formation enthalpies of alloys within framework of Miedema’s Theory *Comput. Phys. Commun.* **209** 58–69

[64] Miedema A R and Niessen A K 1983 The enthalpy of solution for solid binary alloys of two 4d-transition metals *Calphad* **7** 27–36

[65] Anon Periodic Table of Elements - PubChem

[66] Anon Elastic properties of the elements (data page) - Wikipedia

[67] Anon AZOM Materials

[68] Zunger A 1980 Systematization of the stable crystal structure of all AB-type binary compounds: A pseudopotential orbital-radii approach *Phys. Rev. B* **22** 5839–72

[69] Samsonov G V. 1968 *Handbook of the Physicochemical Properties of the Elements* (New York: Springer)

[70] Dinsdale A T 1991 SGTE data for pure elements *Calphad* **15** 317–425

[71] Kittel C 2005 *Introduction to Solid State Physics* (Hoboken, NJ: John Wiley & Sons, Inc.)

[72] Shang S-L, Zhou B-C, Wang W Y, Ross A J, Liu X L, Hu Y-J, Fang H-Z, Wang Y and Liu Z-K 2016 A comprehensive first-principles study of pure elements: Vacancy formation and migration energies and self-diffusion coefficients *Acta Mater.* **109** 128–41

[73] Anderson M S, Swenson C A and Peterson D T 1990 Experimental equations of state for calcium, strontium, and barium metals to 20 kbar from 4 to 295 K *Phys. Rev. B* **41** 3329–38

[74] Rasmussen C E and Williams C K I 2006 *Gaussian processes for machine learning* (MIT Press)

[75] Liu H, Ong Y-S, Shen X and Cai J 2020 When Gaussian Process Meets Big Data: A Review of Scalable GPs *IEEE Trans. Neural Networks Learn. Syst.* 1–19

[76] Knudde N, Raes W, De Bruycker J, Dhaene T and Stevens N 2020 Data-Efficient Gaussian Process Regression for Accurate Visible Light Positioning *IEEE Commun. Lett.* **24** 1705–9

[77] Peng J, Lee S, Williams A, Haynes J A and Shin D 2020 Advanced data science toolkit for non-data scientists – A user guide *Calphad* **68** 101733

[78] Sedgwick P 2012 Pearson’s correlation coefficient *BMJ* **345** e4483

[79] Reshef D N, Reshef Y A, Finucane H K, Grossman S R, McVean G, Turnbaugh P J, Lander E S, Mitzenmacher M and Sabeti P C 2011 Detecting Novel Associations in Large Data Sets *Science (80-. ).* **334** 1518–24

[80] Lee B-J, Shim J-H and Baskes M I 2003 Semiempirical atomic potentials for the fcc metals Cu, Ag, Au, Ni, Pd, Pt, Al, and Pb based on first and second nearest-neighbor modified embedded atom method *Phys. Rev. B* **68** 144112

[81] Jin Z H, Dunham S T, Gleiter H, Hahn H and Gumbsch P 2011 A universal scaling of planar fault energy barriers in face-centered cubic metals *Scr. Mater.* **64** 605–8

[82] Ogata S 2002 Ideal Pure Shear Strength of Aluminum and Copper *Science (80-. ).* **298** 807–11

[83] Nakashima P N H, Smith A E, Etheridge J and Muddle B C 2011 The Bonding Electron Density in Aluminum *Science (80-. ).* **331** 1583–6

**TABLES AND TABLE CAPTIONS**

Table 1. Stacking fault energy, γSFE, values (in mJ/m2) of fcc elements by DFT-based first-principles calculations and experimental measurements.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Elem. | The present work (Calc.) | | | | Calc. | Experimental results in the literature | | | | |
|  |  |  |  | [81] | [29] | [32] | [33] | [80] | [31] |
| Ag | 16 a | 15 b | 17 c | 15 d | 16 e | 16 f,g | 16 i | 21 j,k | < 22 m |  |
| Al | 128 a | 132 b | 123 c | 129 d | 112 e | 150 f,h | 166 i | 130 j,k | < 166 m | 110 ~ 280 n |
| Au | 26 a | 25 b | 26 c | 28 d | 25 e | 32 f,g | 32 i | 55 j,k | < 55 m |  |
| Cu | 40 a | 42 b | 37 c | 38 d | 36 e | 41 f,g | 45 i | 85 j,k | < 78 m | 24 ~ 169 n |
| Ir | 348 a | 348 b | 348 c | 354 d | 334 e |  | 300 i | 300 j,l |  |  |
| Ni | 131 a | 132 b | 131 c | 137 d | 133 e | 125 f,g | 125 i | 450 j,k | 125 m | 90 ~ 450 n |
| Pd | 139 a | 141 b | 136 c | 140 d | 134 e | 80 f,g | 180 i | 170 j,k | 100 m |  |
| Pt | 272 a | 272 b | 271 c | 283 d | 286 e | 45 f,g | 322 i | 95 j,k | 110 m |  |
| Rh | 196 a | 194 b | 198 c | 202 d |  | 200 f,g | 750 i | 330 j,k |  |  |

a Recommended values in the present work based on the average of and results.

b : The present DFT results calculated by pure alias shear deformation using a 6-atom orthorhombic cell.

c : The present DFT results calculated by pure alias shear deformation using a 24-atom orthorhombic cell.

d : The present DFT results calculated by simple alias shear with relaxations using a 24-atom orthorhombic cell.

e Calculated results based on the DFT-based climbing-image nudged elastic band method [81].

f Selected experimental data by de Campos [29].

g Measured data based on the weak beam method.

h Measured data using high-resolution transmission electron microscopy.

i Selected experimental data by Hirth and Lothe [32].

j Selected experimental data by Smallman et al. [33].

k Measured data based on the indirect texture method and these results may be not reliable [29].

l Measured data based on the stage 3 of the stress-stress curve and these data have low reliability [29].

m Selected experimental data by Lee et al. [80].

n Cited experimental data by Zhang et al. [31], which are scattered and in a wide range of values.

Table 2. Predicted values of ΔγSFE (mJ/m2) in dilute Al23X, Ni23X, and Pt23X with respect to the γSFE values of pure Al, Ni and Pt, respectively (see Table 1), by DFT-based calculations or machine learning (ML). The ML values are labelled by superscript \*, which are the average of the GPR2, GPR4, OptGPR, and WM\_NNet (see their definitions in Table 4) predictions with their standard deviations in parentheses. The unperformed predictions are marked by “Null” or empty; and the green (low) — yellow (middle) — red (high) color scale has been used to mark the ΔγSFE data. For the present DFT calculations of Pt23X, the ‘pv’ or ‘sv’ after an atomic symbol X indicate that the p or s states are treated as valence states, respectively.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Li** | **Be** |  |  |  |  |  |  |  |  |  |  | **B** | **C** |
| -2 | 33 |  |  |  |  |  |  |  |  |  |  |  |  |
| Null | Null |  |  |  |  |  |  |  |  |  |  |  |  |
| Null | Null |  |  |  |  |  |  |  |  |  |  |  |  |
| **Na** | **Mg** |  |  |  |  |  |  |  |  |  |  | **Al** | **Si** |
| -52 | -8 | ΔγSFE of Al23X from DFT or ML  ΔγSFE of Ni23X from DFT or ML | | | | | | | | | | 0 | -22 |
| Null | -34\* (3) | -28 | -27 |
| Null | -14\* (1) | ΔγSFE of Pt23X from DFT or ML | | | | | | | | | | -15 | -36 |
| **K** | **Ca** | **Sc\_sv** | **Ti\_pv** | **V\_pv** | **Cr\_pv** | **Mn\_pv** | **Fe\_pv** | **Co** | **Ni\_pv** | **Cu\_pv** | **Zn** | **Ga** | **Ge** |
| -134 | -66 | -75 | -73 | -60 | 7 | 44 | 44 | 83 | 55 | 27 | 2 | -19 | -37 |
| Null | -89\* (9) | -74 | -65 | -70 | -42 | -24 | -27 | -22 | 0 | -16 | -23 | Null | Null |
| Null | -75\* (3) | -63 | -60 | -66 | -19 | -5 | 5 | 2 | 22 | 18 | -5 | Null | Null |
| **Rb** | **Sr** | **Y\_sv** | **Zr\_sv** | **Nb\_pv** | **Mo\_pv** | **Tc\_pv** | **Ru\_pv** | **Rh\_pv** | **Pd** | **Ag** | **Cd** | **In** | **Sn** |
|  | -146 | Null | -114 | -95\* (4) | -56\* (1) | -16\* (1) | -2\* (4) | 21\* (3) | 44\* (1) | 15 | -14 | -36 | -58 |
|  | Null | -106 | -95 | -92 | -94 | -80 | -54 | -29 | -13 | -35\* (2) | Null | Null | Null |
|  | Null | -122 | -102 | -105 | -92 | -43 | -9 | 6 | -7 | -27 | Null | Null | Null |
| **Cs** | **Ba** | **La** | **Hf\_pv** | **Ta\_pv** | **W\_pv** | **Re\_pv** | **Os\_pv** | **Ir** | **Pt** | **Au** | **Hg** | **Tl** | **Pb** |
|  |  | -191 | -98 | -92\* (4) | -73\* (2) | -46\* (9) | -25\* (15) | -6\* (14) | 13\* (2) | -16\* (2) |  |  | -76 |
|  |  | Null | -84 | -82 | -88 | -85 | -60 | -34 | -9 | -36\* (6) |  |  | Null |
|  |  | Null | -87 | -98 | -104 | -61 | -17 | 11 | 0 | -26 |  |  | Null |

Table 3. Iterations between sequential descriptor selection (Step 1) and ML verification (Step 2) of the ranked descriptors from Step 1, aiming to quantify the statistical significance of the descriptors. Each iteration will result in one generation of the ranked descriptors.

|  |  |
| --- | --- |
| Step 1 | Sequential descriptor selection (SDS) using the function of “sequentialfs” in MATLAB to select a subset of descriptors in ReDes49 (see the supplemental csv file) that best predict the ΔγSFE values by sequentially selecting descriptors until there is no improvement in prediction. Here, (i) the GPR4 algorithm was used to predict ΔγSFE in terms of a 5-fold cross validation; (ii) the SDS processes were performed 1000 times in order to rank the descriptors in ReDes49 by statistical significance; (iii) the forward SDS process by adding descriptors to a null list was adopted, and (iv) the fully verified descriptors in Step 2, which decrease the MAE values, were kept in the SDS process when applicable. |
| Step 2 | ML verification of the ranked descriptors in Step 1 using the GPR4 algorithm with a 5-fold cross validation. Here, the descriptors were sequentially added in the verification process according to their statistical significance from high to low, and the verifications were performed 1000 times in order to get an average MAE and its uncertainty for each subset of descriptors. For example, some results are shown in Figure 4 and the supplemental Figure S2. |

Table 4. Twenty-one regression algorithms for supervised machine learning in the software packages of MATLAB (version R2019b) and Wolfram Mathematica (WM, version 12.0).

|  |  |
| --- | --- |
| MATLAB software | WM software |
| * **LR1** to **LR2** 🡪 Linear regressions without (the LR1) and with (the LR2) the robust option * **Tree1** to **Tree3** 🡪 Tree algorithms with the minimum leaf sizes of 4 (the Tree1), 12 (the Tree2), and 36 (the Tree3) * **SVM1 to SVM3** 🡪 Support vector machines with the kernel functions being the linear (the SVM1), the 2-order polynomial (the SVM2), and the 3-order polynomial (the SVM3) functions * **GPR1 to GPR4** 🡪Gaussian process regressions with the kernel functions being the squared exponential (GPR1), the matern52 (GPR2), the exponential (GPR3), and the rational quadratic (GPR4) functions * **OptGPR**a **🡪** Optimized GPR with its hyperparameters (e.g., the basis function and kernel function) determined by Bayesian optimization. Rational quadratic is the optimized kernel function in the present work but with different hyperparameters such as the sigma value to estimated noise standard deviation and the basis function in comparison with those in GPR4 | * **WM\_DT** 🡪 Decision tree * **WM\_GBT** 🡪 Gradient boosted trees * **WM\_LR** 🡪 Linear regression * **WM\_NN** 🡪 Nearest neighbors * **WM\_NNet 🡪** Neural network * **WM\_RF** 🡪 Random forest * **WM\_GP 🡪** Gaussian process |

aThe optimizable algorithms for Tree and SVM cannot yield lower MAE values for the present case and hence ignored.

Table 5. Average MAE values (in mJ/m2) and their standard deviations for predicting ΔγSFE using the algorithms in both MATLAB (version R2019b) and Wolfram Mathematica (WM, version 12.0). As an example, Figure 6 shows a complete examination of the average MAE and MaxAE values together with their uncertainties for the GPR2 algorithm to predict new ΔγSFE values.

|  |  |  |
| --- | --- | --- |
| Software | Algorithm | MAE for prediction with standard deviation in parentheses |
| Matlab | GPR2 | 8.4 (0.7)a 🡪 7.7 (0.5)b 🡪 7.3 (0.2)c |
| Matlab | GPR4 | 8.6 (0.7)a 🡪 7.9 (0.5)b 🡪 7.4 (0.2)c |
| Matlab | OptGPR | 8.5 (0.7)a 🡪 7.8 (0.5)b 🡪 7.3 (0.2)c |
| WM | WM\_NNet | 8.9 (2.9)b |

a Average value for k-fold of 5 with training of 1000 times for the GPR algorithms.

b Average value for k-fold of 10 and training of 1000 times for the GPR algorithms (or 50 times for the WM\_NNet).

c Average value for k-fold of 30 with training of 1000 times for the GPR algorithms.

Table 6. MAE values (in mJ/m2) to predict ΔγSFE of new alloys by using different ML algorithms.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| New alloys | GPR2 | GPR4 | OptGPR | WM\_NNet |
| Al23X a | 36.1 | 36.7 | 34.8 | 32.9 |
| Ni23X b | 17.1 | 16.0 | 16.7 | 20.1 |
| Pt23X c | 20.4 | 21.5 | 17.8 | 24.2 |

a Employed ML models trained by the ΔγSFE values of Ni23X and Pt23X.

b Employed ML models trained by the ΔγSFE values of Al23X and Pt23X.

c Employed ML models trained by the ΔγSFE values of Al23X and Ni23X.

**FIGURES AND FIGURE CAPTIONS**

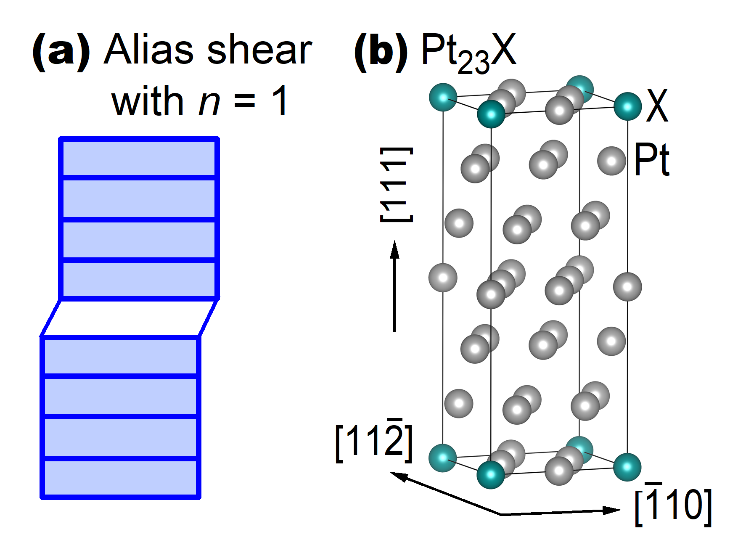
****

Figure 1. (a) Schematic diagram of alias shear with atoms slid in one plane (*n* = 1, the unshaded area) during the initial stage of shear, and (b) A 24-atom orthorhombic supercell of Pt23X with its lattice vectors aorth, borth, and corth parallel to the , , and directions of the conventional fcc lattice, respectively.

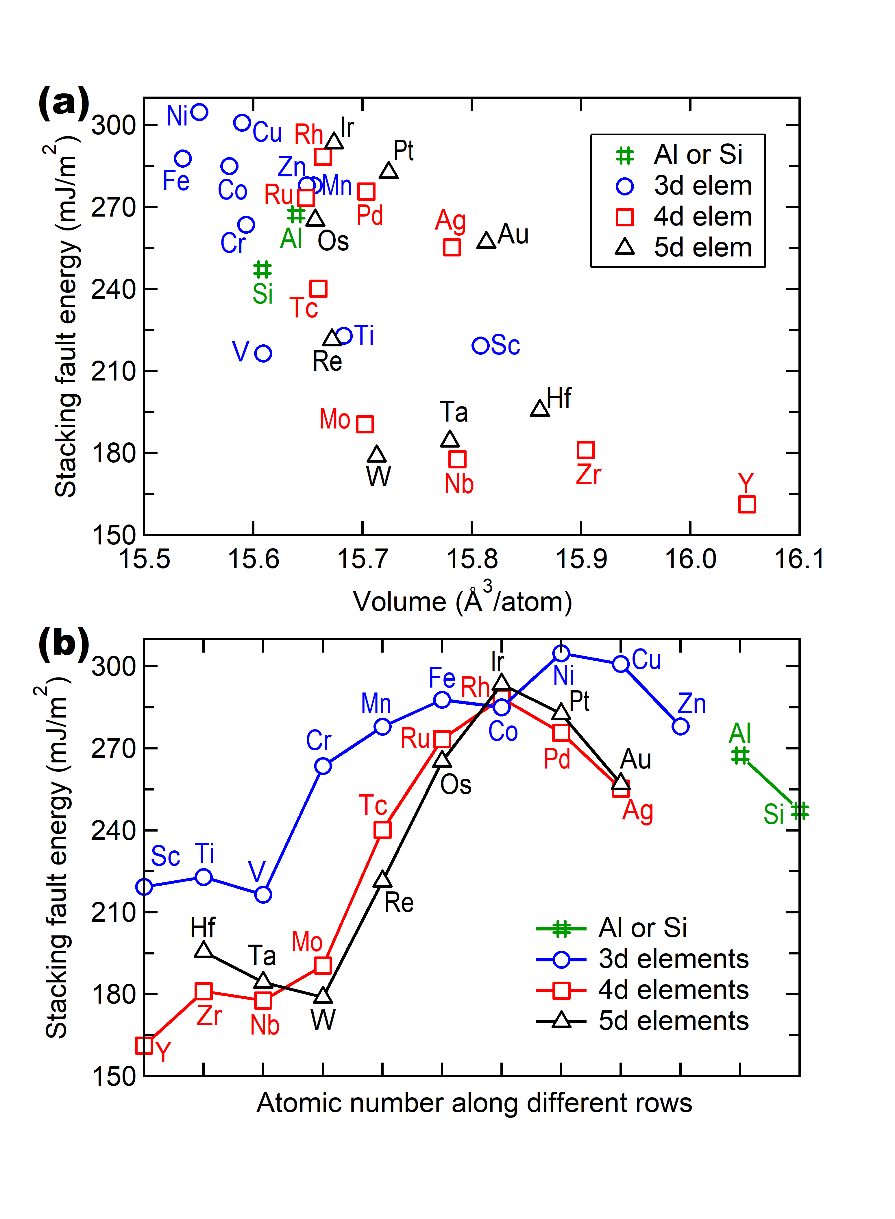


Figure 2. Stacking fault energies of Pt23X based on the present first-principles calculations plotted as a function of (a) equilibrium volume of Pt23X and (b) atomic number along different rows; see the γSFE values in Table 2 and the supplemental csv file.

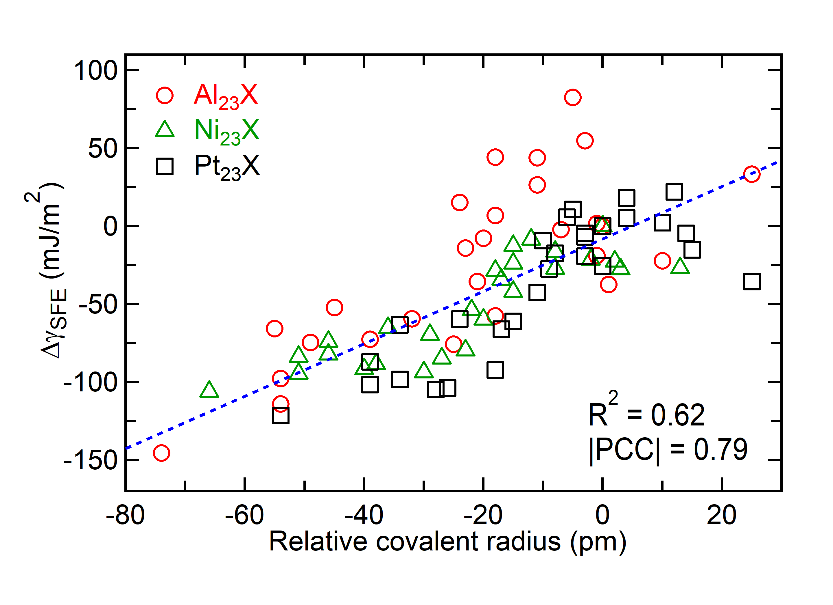


Figure 3. Relationship of the top one descriptor of covalent radius with respect to the ΔγSFE values. The dashed lines show the linear fitting with its R2 as well as its |PCC| value included.

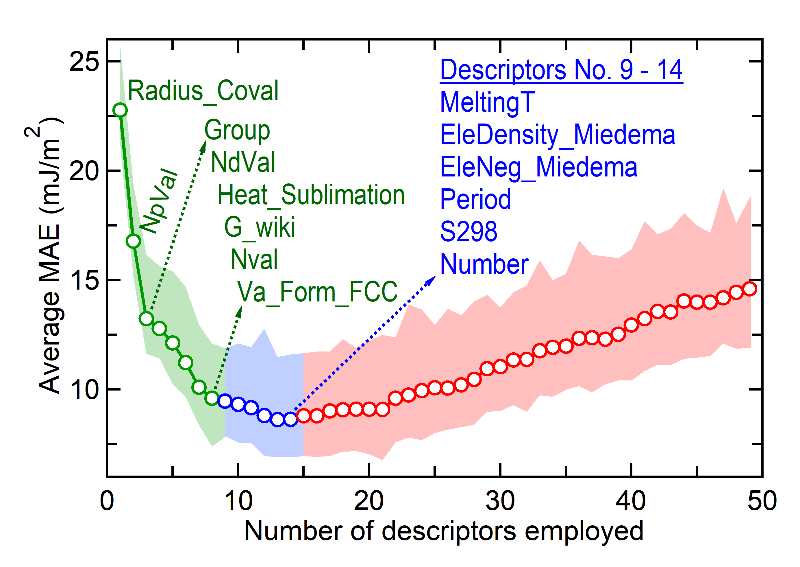


Figure 4. Average MAE values and their uncertainties (the color regions) from sequentially adding the final generation of the ranked descriptors based on their statistical significance; see the methodology in Table 3. The combination of the top 14 descriptors (shown in the green and blue colors, i.e., the ReDes14 in the supplemental csv file) results in the lowest MAE value. A complete list of the ranked descriptors is shown in the supplemental Table S1.

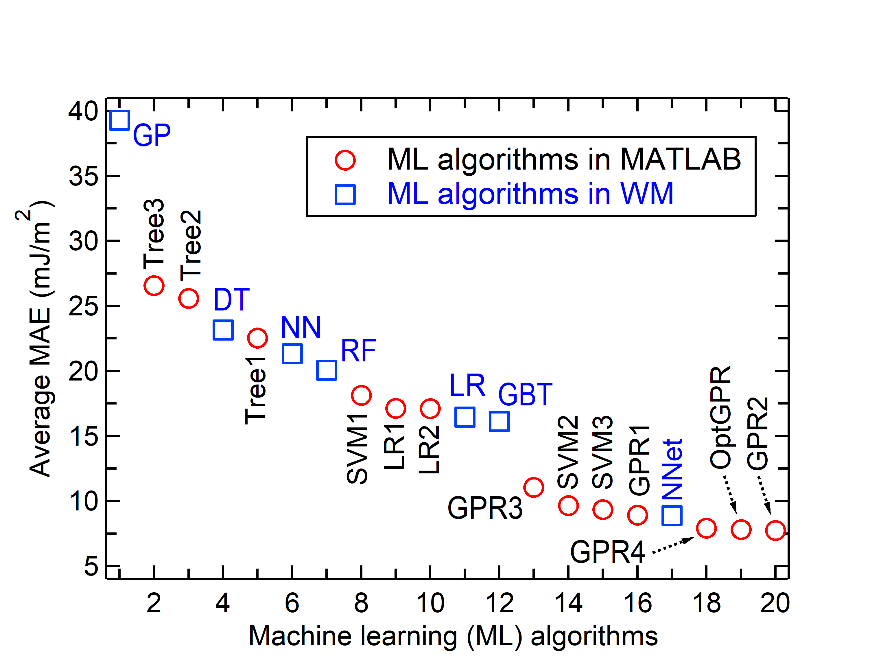


Figure 5. Examination of the ML algorithms (listed as *x*-axis) to predict new ΔγSFE values using the MATLAB software with 5-fold cross validation and the WM software with 10-fold cross validation. Here the optimal subset of descriptors in ReDes14 are used, and the details of these ML algorithms are shown in Table 4.

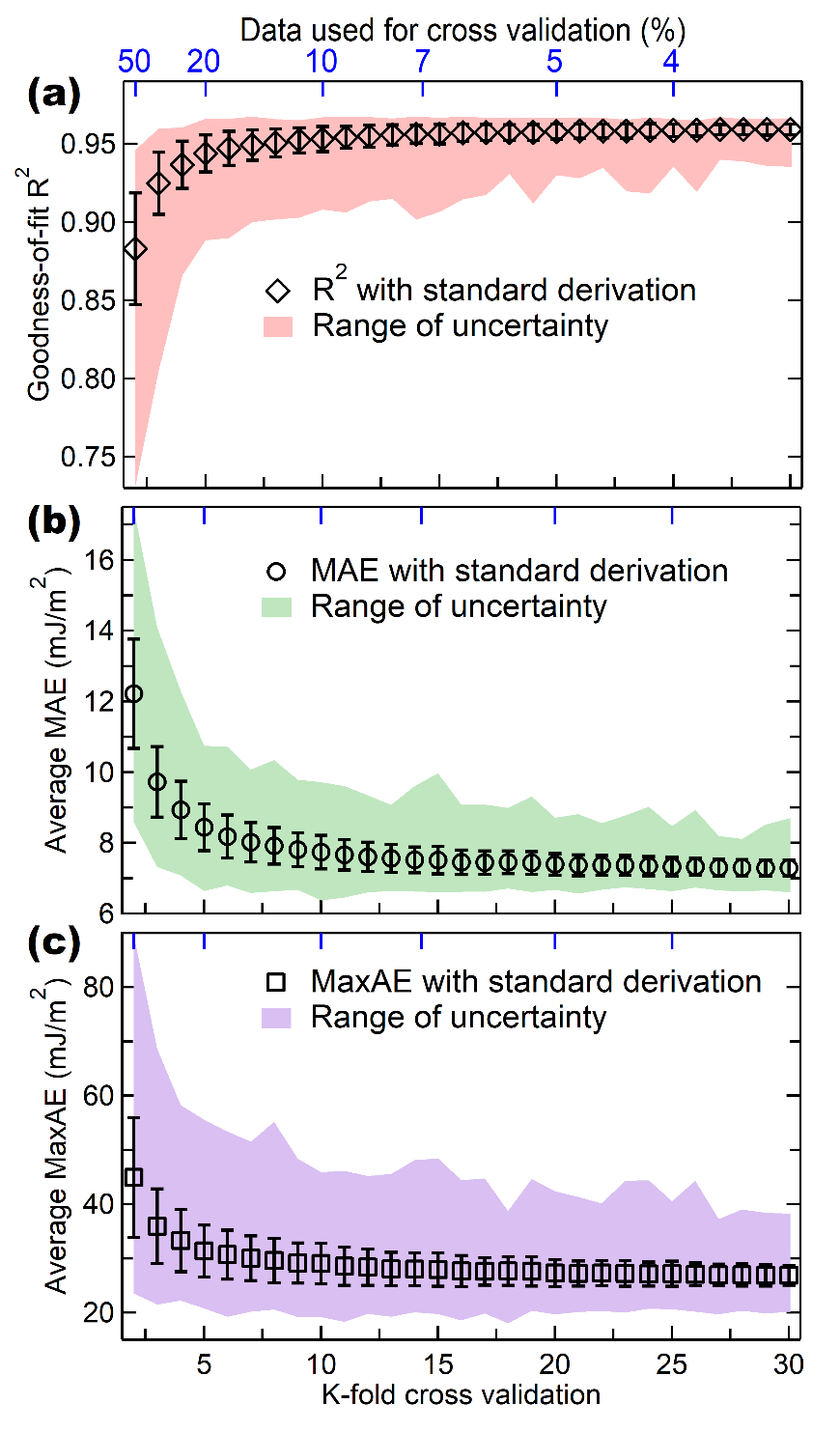


Figure 6. Average MAE and MaxAE values together with the goodness-of-fit for linear regression (i.e., the R2 values) to predict new ΔγSFE using the GPR2 algorithm as a function of k-fold cross validation from 2 to 30. Here, (i) the ReDes14 descriptors are used and (ii) the ML predictions are performed for 1000 times in order to gain the average values of R2 (a), MAE (b), and MaxAE (c), as well as their standard derivations and the range of uncertainties.

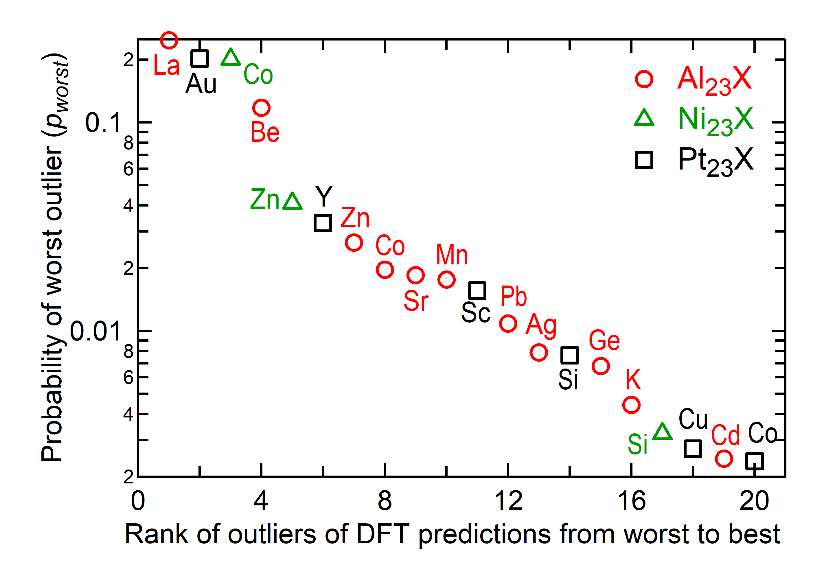


Figure 7. The top 20 worst outliers of DFT predictions of γSFE ranked by occurrence probability ofthe outliers using the MaxAE values between DFT calculations and ML predictions. The total worst probability *p*worst = 1.

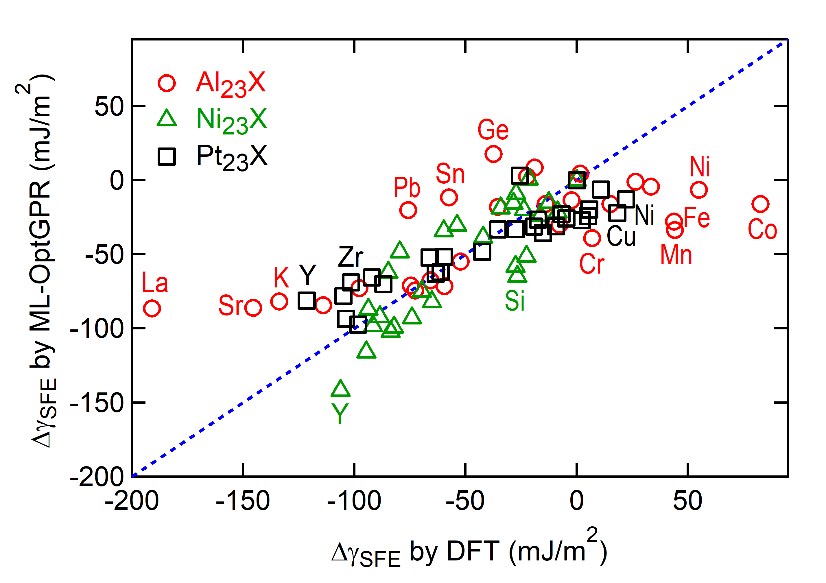


Figure 8. Comparison of DFT calculated ΔγSFE and ML predicted ΔγSFE values for new alloys of Al23X, Ni23X and Pt23X using the OptGPR model; see the methods and the MAE values in Table 6.

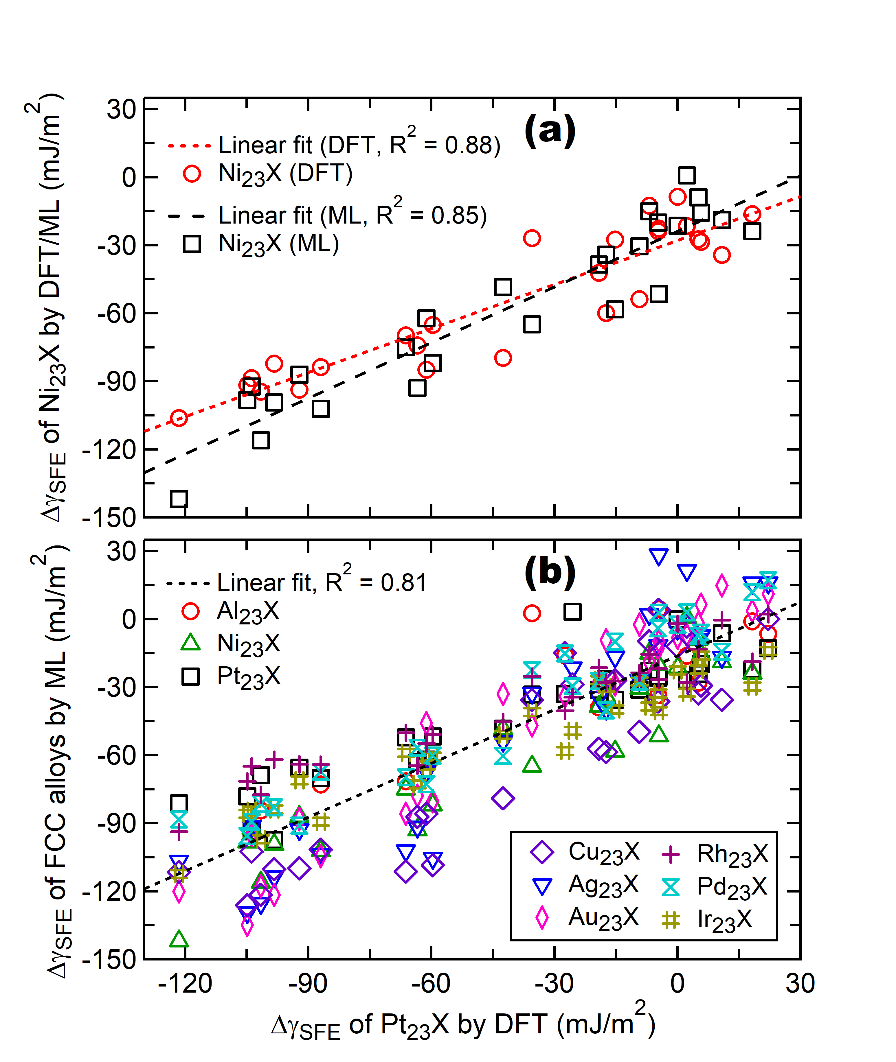


Figure 9. Relationship between the ΔγSFE values of Pt23X from DFT calculations and the ΔγSFE values of (a) Ni23X from DFT and ML predictions and (b) fcc-based dilute alloys predicted by the OptGPR model. Note that the ML models were trained by the ΔγSFE values of Al23X, Ni23X, and Pt23X; while the ΔγSFE values of Al23X, Ni23X, and Pt23X from ML are the same as those in Figure 8 (see methods given in Table 6).