Petascale supercomputing to accelerate the design of high-temperature alloys

Dongwon Shin\textsuperscript{a}, Sangkeun Lee\textsuperscript{b}, Amit Shyam\textsuperscript{a} and J. Allen Haynes\textsuperscript{a}

\textsuperscript{a}Materials Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, USA; \textsuperscript{b}Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, USA

**ABSTRACT**

Recent progress in high-performance computing and data informatics has opened up numerous opportunities to aid the design of advanced materials. Herein, we demonstrate a computational workflow that includes rapid population of high-fidelity materials datasets via petascale computing and subsequent analyses with modern data science techniques. We use a first-principles approach based on density functional theory to derive the segregation energies of 34 microalloying elements at the coherent and semi-coherent interfaces between the aluminium matrix and the $\theta'$-Al$_2$Cu precipitate, which requires several hundred supercell calculations. We also perform extensive correlation analyses to identify materials descriptors that affect the segregation behaviour of solutes at the interfaces. Finally, we show an example of leveraging machine learning techniques to predict segregation energies without performing computationally expensive physics-based simulations. The approach demonstrated in the present work can be applied to any high-temperature alloy system for which key materials data can be obtained using high-performance computing.

**1. Introduction**

A recent Oak Ridge National Laboratory (ORNL) study by Shyam et al. [1] demonstrated that it is possible to exert a remarkable stabilizing influence on thermodynamically metastable strengthening precipitates ($\theta'$-Al$_2$Cu) within aluminium-copper (Al-Cu) alloys to temperatures of at least 300 °C via microalloying with manganese (Mn) and zirconium (Zr). An extensive experimental investigation using atom probe tomography and scanning transmission electron microscopy revealed that trace amounts of microalloyed solute atoms segregated at the interfaces and inhibited the coarsening process that leads to detrimental phase transformation ($\theta'$-$\theta$) at elevated temperatures. This exciting experimental observation of an extended temperature regime of precipitate stability offers numerous opportunities to design new classes of alloys with significantly improved mechanical properties at high temperatures by understanding and then manipulating the interfacial stability of key precipitates.

Shyam et al. [1] also showed that a hierarchy and synergy exist in terms of the ability of individual or combined elements to stabilize the key interfaces in $\theta'$ precipitates at high temperatures. They demonstrated that either Mn or Zr additions alone provide stabilization of $\theta'$ only to a certain extent, but the critical temperature – below which stability and strength can be
preserved – can be further elevated by adding Mn and Zr together. The underlying origin of the observed elemental hierarchy, which would be extremely useful for high-temperature alloy design, is currently unknown. Identifying such favourable combinations of microalloying elements solely via experimental studies would be prohibitively time consuming and costly because of the complex multi-component nature of most high-temperature alloys. Accurate theoretical guidance and prediction of the effects of microalloying additions will have an across-the-board impact on the understanding of the mechanical behaviour of the next generation of high-temperature alloys.

It has been steadily demonstrated during the past couple of decades that first-principles density functional theory (DFT) calculations can provide accurate predictions of energetics, diffusion kinetics, and lattice mismatch/strain in alloys [2]. However, most of these type DFT calculations have been focused on elucidating the underlying physics and chemistry in single-phase materials. To directly guide the design of multi-phase high-temperature alloys, more complex models that can better represent realistic materials systems and key interfaces are needed. Usually, such models are supercells with a large number of atoms – of the order of a hundred atoms – and the size of supercells has often been limited by the available computing power. In addition, constructing a large DFT database is desirable to identify element(s) that can promote improved physical/chemical properties of alloys. However, performing a series of DFT calculations of large supercells in a high-throughput manner within a reasonable timeframe requires an extensive amount of computing resources in a short period. Moreover, analysing large datasets populated from massive physics-based calculations to generate materials hypotheses is also a daunting task.

In the current work, we used a defect supercell approach in a high-throughput manner to construct a large first-principles database of the segregation energies of 34 microalloying solutes at both coherent and semi-coherent interfaces between an Al matrix and $\theta'$-Al$_2$Cu precipitate after Biswas et al. [3]. The pure FCC Al and $\theta'$-Al$_2$Cu phases were fully relaxed, respectively, before being combined to create supercell models for coherent and semi-coherent interfaces. Next, the two generated supercells were fully relaxed with a periodic boundary condition invoked in all directions to induce a strain effect between the two phases. Finally, when solute atoms were introduced within two supercells, the supercell was forced to retain its volume and shape, while the internal positions of individual atoms were allowed to be relaxed. In addition, the periodic boundary condition was only invoked in the directions normal to the long axis of the supercell.

We considered the segregation of 34 elements; the supercell models are illustrated in Figure 1. Only one solute was introduced in each supercell, and a solute was progressively moved from the bulk toward the interface to derive the solute segregation energy as a function of the location from the interface. We considered all the crystallographically distinctive lattice sites at each platelet. As shown in Figure 1, four platelets were used to represent the Al matrix at both interfaces. They are denoted

2. Computational approaches

2.1. First-principles calculations

2.1.1. Solute segregation energy

We used a supercell approach to simulate the solute segregated coherent and semi-coherent interfaces between the Al matrix and $\theta'$-Al$_2$Cu precipitate after Biswas et al. [3]. The pure FCC Al and $\theta'$-Al$_2$Cu phases were fully relaxed, respectively, before being combined to create supercell models for coherent and semi-coherent interfaces. Next, the two generated supercells were fully relaxed with a periodic boundary condition invoked in all directions to induce a strain effect between the two phases. Finally, when solute atoms were introduced within two supercells, the supercell was forced to retain its volume and shape, while the internal positions of individual atoms were allowed to be relaxed. In addition, the periodic boundary condition was only invoked in the directions normal to the long axis of the supercell.

We considered the segregation of 34 elements; the supercell models are illustrated in Figure 1. Only one solute was introduced in each supercell, and a solute was progressively moved from the bulk toward the interface to derive the solute segregation energy as a function of the location from the interface. We considered all the crystallographically distinctive lattice sites at each platelet. As shown in Figure 1, four platelets were used to represent the Al matrix at both interfaces. They are denoted

![Figure 1. Relaxed supercells used for the Al/$\theta'$ interfacial segregation calculations. Vertical lines in each figure represent interfaces. Top: coherent interface (108 atoms); bottom: semi-coherent interface (168 atoms). The red atom (indicated with an arrow on the fourth row) in the semi-coherent supercell shows the position of a solute at a ‘bulk’ site in the Al matrix, which is sufficiently far from the interface. Platelets are labelled according to their positions relative to the interface planes. See the text for details.](image-url)
as Al\textsubscript{x}, where \( x \) represents the locations of solutes, \( i \) the interface, and \( i−1 \) and \( i−2 \) the first and second platelets away from the interface on the matrix side. The Al\textsubscript{i+3} platelet was assumed to be sufficiently far away from the interface to be considered ‘bulk’ and was denoted Al\textsubscript{b}.

The solute segregation energy can be described by the difference between the two total energies as shown in equation (1):

\[
\Delta E_{\text{seg}} = E(\text{Al}_i) - E(\text{Al}_b)
\]

where \( E \) represents the total energy from first-principles DFT electronic structure calculations. The details of the DFT approach to deriving the segregation energies of solutes at the interfaces between Al and \( \theta'-\text{Al}_2\text{Cu} \) are presented in Ref. [4].

2.1.2. Petascale supercomputing

Eighteen supercell calculations (5 and 13 supercells for coherent and semi-coherent interfaces, respectively, to consider crystallographically distinctive sites in each platelet) per element were required to derive solute segregation with the DFT approach described above. Thus, several hundred DFT supercell calculations in total were required to construct a comprehensive solute segregation energy database to consider 34 elements. If regular high-performance computing (HPC) resources had been used to run the several hundred supercell DFT calculations, the same number of jobs would have been submitted in serial as illustrated in Figure 2.

To accommodate many users on limited resources and to maximize efficiency, most computing centres have adopted job queueing systems based on a priority policy. In general, the more resources users request and spend, the sooner they lose their priority. Hence, a user may run several hundred calculations fast in the early stages by running multiple jobs simultaneously while the priority is high. However, the progress will slow down as the accumulated usage becomes larger, until it reaches a periodic priority reset process. Although the computing power and capacity of currently available HPC resources have increased steadily over the past decade, the demand for access to supercomputing resources also has increased exponentially. Thus, implementing a large number of expensive calculations on regular HPC resources has become even more challenging in recent years.

In this regard, the leadership computing facilities of the US Department of Energy (DOE) offer unique opportunities by supporting computationally intensive large-scale projects with large amounts of dedicated time on supercomputers. The Cray XK7 supercomputer Titan, hosted at the Oak Ridge Leadership Computing Facility (OLCF), has 18,688 compute nodes that offer a peak performance of around 20 petaflops. As a DOE Leadership Computing Facility, the OLCF has a mandate that a large portion of Titan’s computing time be allocated to exceptionally large jobs. To ensure the OLCF complies with DOE directives, users are encouraged to run jobs on Titan that are as large as their code will warrant, as shown in Table 1. As illustrated in Figure 2, several hundred supercell calculations can be grouped together as a number of ensemble jobs. This unique allocation policy at OLCF enables researchers to tackle problems in which extensive theoretical calculations can provide insight into the design and development of advanced materials.

In addition, we took advantage of the graphic processing unit (GPU) hybrid architecture of Titan and exploited the GPU version of the widely used DFT code Vienna Ab initio Simulation Package (VASP) [5,6], which offers \( \sim 2–3 \times \) faster performance than its counterpart CPU version. Currently, GPU-VASP supports only real-space projection, but extensive benchmark calculations against the CPU-VASP code supporting reciprocal-space projection have shown that the difference in the solute segregation energy is less than 0.1
meV for the current supercell calculations. All the DFT supercell calculations were performed using projector augmented wave \cite{7} potentials and the generalized gradient approximation \cite{8}. We used Perdew-Burke-Ernzerhof for the exchange–correlation functional \cite{9}.

2.1.3. Materials descriptors from DFT calculations
We included three DFT-derived materials descriptors in the correlation analysis: DFT solute volume and the mixing energies in the Al and Cu sublattices within $\theta'$-Al$_2$Cu via special quasi-random structures (SQSs) \cite{10}. We calculated the solute volume within Al by comparing the DFT volume difference with and without a solute in the Al supercell, as described in previous studies \cite{11,12}.

![Figure 3](image-url)  
**Figure 3.** DFT segregation energies of 34 solutes at each platelet (Al$_{i-2}$, Al$_{i-1}$ and Al$_i$) at the coherent and semi-coherent interfaces between the Al matrix and $\theta'$-Al$_2$Cu. Only the lowest segregation energies at given platelets are shown.

![Figure 4](image-url)  
**Figure 4.** DFT solute segregation energies plotted with selected descriptors: (a) molar volume at the coherent interface (Al$_{i-2}$); (b) DFT solute volume in Al at the coherent interface (Al$_{i-2}$); (c) atomic radius at the semi-coherent interface (Al$_i$); and (d) mixing energy of solute at the Al sublattice in $\theta'$-Al$_2$Cu from DFT SQSs calculation at the semi-coherent interface (Al). Vertical lines in each figure represent the respective values for Al, while the DFT solute volume of Al itself and the mixing within the Al sublattice in $\theta'$-Al$_2$Cu are treated as zero.
Sci. Technol. Adv. Mater. 18 (2017) 832  D. SHIN et al.

Al\textsubscript{108} (3 × 3×3) and Al\textsubscript{107}X supercells were used to obtain the solute volumes corresponding to 34 elements. SQSs can mimic random mixing between solute atoms and the respective Al/Cu sublattices of θ'-Al\textsubscript{2}Cu by satisfying the atomic arrangement quantified as correlation functions. Therefore, SQSs can serve as structural templates to derive the mixing energies by switching atomic identities within the context of DFT electronic structure calculations. SQS mixing energies in the Al and Cu sublattices are denoted as Al\textsubscript{mix}(SQS) and Cu\textsubscript{mix}(SQS), respectively, hereafter.

2.2. Materials informatics

2.2.1. Correlation Analysis

Using the generated DFT database of solute segregation energies in Al-Cu alloys, we investigated the results to identify key descriptors by performing correlation analysis. We analysed the segregation energies of 34 elements in both coherent and semi-coherent interfaces with 17 atomistic descriptors in five different groups:

- atomic size/volume: atomic number, weight, radius, DFT solute volume, molar volume and covalent radius (single and double bond);
- atomic structure: electron affinity;
- physical property: density at 25 °C;
- atomic interaction: Pauling electronegativity and chemical hardness;
- thermodynamic properties: melting point, boiling point, enthalpy of fusion (Δ\textsubscript{H}\textsubscript{fusion}), enthalpy of vaporization (Δ\textsubscript{H}\textsubscript{vaporization}), and SQS mixing energies in Al and Cu sublattices denoted as Al\textsubscript{mix}(SQS) and Cu\textsubscript{mix}(SQS), respectively.

In analysing the dataset for the DFT solute segregation energy combined with the 17 atomistic descriptors, we applied an advanced technique called maximal information coefficient (MIC) analysis [13], as well as

![Figure 5](image_url). Correlation coefficients between the solute segregation energies derived from DFT calculations at the coherent interface (Al\textsubscript{i}, Al\textsubscript{i−1}, and Al\textsubscript{i−2} platelets) and materials descriptors with MIC (top) and Pearson (bottom) approaches.
generalized linear models, Kernel ridge regression, nearest neighbors, ensemble methods, etc. To cover the variety of different machine learning models, we picked two from generalized linear models (Linear regression [15], Bayesian ridge regression [16]), Kernel ridge regression [17] (itself a separate category), one from nearest neighbors (Nearest neighbour regression [18]), and one from ensemble methods (Random forest regression [19]). A brief introduction to the individual machine learning models used in the current study is presented below.

2.2.2. Machine learning
We explored the feasibility of developing data-driven approaches, such as data mining and machine learning, to predict segregation energies without performing computing-intensive tasks (e.g. DFT and molecular dynamic simulations). The numerical approach to predicting output variable value for given input variables by estimating the relationship between input variables and output variables is referred to as regression in the machine learning technique. We used a Python-based open source data analytics toolkit, scikit-learn [14], which contains a wide variety of machine learning algorithms. The library classifies regression models in several categories such as conventional Pearson's correlation coefficient (PCC) analysis. We used both the MIC and PCC approaches to crosscheck whether the identified descriptors are consistent and to select features to be used to train machine learning models.

Figure 6. Correlation coefficients between the solute segregation energies derived from DFT calculations at the semi-coherent interface (Al, Al, and Al platelets) and materials descriptors with MIC (top) and Pearson (bottom) approaches.
We arbitrarily selected the top ten ranking features to be used within the model training based on our correlation analyses. The DFT solute segregation database generated in the current work was used as an input dataset for training these machine learning models.

3. Results and discussion

3.1. DFT solute segregation energies

Solute segregation energies at both the coherent and semi-coherent interfaces at each platelet are plotted in Figure 3. This plot can be categorized into four quadrants, which combine favourable and unfavourable solute segregation as negative and positive segregation energies, respectively, at the two different interfaces. Notably, both Mn and Zr belong in the upper left quadrant in which segregation is favourable at the semi-coherent but not at the coherent interface at Al. Scandium (Sc), which has been experimentally reported to stabilize θ by solute segregation and thus provide high coarsening resistance at high temperatures [20,21], also belongs to the same quadrant. Therefore, we can correlate the improved high-temperature stability of Al-Cu alloys to microalloying elements (Mn, Zr, and Sc) and their anisotropic solute segregation energies at the two different interfaces.

distribution. To overcome the overfitting, based on Bayesian inference, it incorporates prior distributions for the estimated parameters for the model.

**Kernel Ridge regression** is a model to estimate the conditional expectation of a random variable to find a non-linear relation between a pair of random variables. It uses the kernel method for simplifying the computation of inner products in a high-dimensional space and learns a linear model in the implicit feature space induced by a kernel and the dataset.

**((k-))Nearest neighbours regression** is a non-parametric method that can be used when the input variables are continuous variables. It is one of the simplest machine learning algorithms, as it simply outputs the average of the values of given data points, k nearest neighbours. It is considered to be a lazy learning method, as it defers the computation and uses only a portion of relevant datasets.

**Random forest regression** is an ensemble learning method that constructs multiple decision trees; it learns recursive decision rules inferred from the data features represented as a tree structure at training time and outputting mean prediction of the individual trees. Using multiple decision trees is one technique to overcome the decision tree’s overfitting to the training dataset.

We arbitrarily selected the top ten ranking features to be used within the model training based on our correlation analyses. The DFT solute segregation database generated in the current work was used as an input dataset for training these machine learning models.

3. Results and discussion

3.1. DFT solute segregation energies

Solute segregation energies at both the coherent and semi-coherent interfaces at each platelet are plotted in Figure 3. This plot can be categorized into four quadrants, which combine favourable and unfavourable solute segregation as negative and positive segregation energies, respectively, at the two different interfaces. Notably, both Mn and Zr belong in the upper left quadrant in which segregation is favourable at the semi-coherent but not at the coherent interface at Al. Scandium (Sc), which has been experimentally reported to stabilize θ by solute segregation and thus provide high coarsening resistance at high temperatures [20,21], also belongs to the same quadrant. Therefore, we can correlate the improved high-temperature stability of Al-Cu alloys to microalloying elements (Mn, Zr, and Sc) and their anisotropic solute segregation energies at the two different interfaces.

| Platelets | Values | Correlation analysis | Machine learning models |
|-----------|--------|----------------------|------------------------|
| Coherent(AI) |        |                      | RF         | LR | NN | KR | BR |
|           |        |                      | R  | MIC | 0.799 | 0.856 | 0.820 | 0.842 | 0.920 |
|           |        |                      | Pearson  | 0.785 | 0.909 | 0.804 | 0.786 | 0.923 |
|           |        |                      | R² | MIC | 0.639 | 0.733 | 0.672 | 0.709 | 0.847 |
|           |        |                      | Pearson  | 0.616 | 0.826 | 0.646 | 0.618 | 0.851 |
|           |        |                      | RMSE | MIC | 0.239 | 0.227 | 0.224 | 0.238 | 0.176 |
|           |        |                      | Pearson  | 0.248 | 0.173 | 0.249 | 0.269 | 0.155 |
| Coherent(AI–1) |        |                      | R  | MIC | 0.574 | 0.564 | 0.710 | 0.429 | 0.395 |
|           |        |                      | Pearson  | 0.572 | 0.505 | 0.672 | 0.594 | 0.398 |
|           |        |                      | R² | MIC | 0.329 | 0.318 | 0.505 | 0.184 | 0.129 |
|           |        |                      | Pearson  | 0.328 | 0.255 | 0.452 | 0.353 | 0.158 |
|           |        |                      | RMSE | MIC | 0.213 | 0.233 | 0.192 | 0.264 | 0.256 |
|           |        |                      | Pearson  | 0.233 | 0.259 | 0.192 | 0.227 | 0.253 |
| Coherent(AI–2) |        |                      | R  | MIC | 0.757 | 0.408 | 0.800 | 0.535 | 0.550 |
|           |        |                      | Pearson  | 0.631 | 0.509 | 0.666 | 0.526 | 0.629 |
|           |        |                      | R² | MIC | 0.573 | 0.167 | 0.640 | 0.286 | 0.302 |
|           |        |                      | Pearson  | 0.399 | 0.259 | 0.443 | 0.277 | 0.396 |
|           |        |                      | RMSE | MIC | 0.127 | 0.228 | 0.108 | 0.193 | 0.146 |
|           |        |                      | Pearson  | 0.154 | 0.187 | 0.144 | 0.193 | 0.141 |
| Semi-coherent(AI) |        |                      | R  | MIC | 0.886 | 0.909 | 0.788 | 0.683 | 0.910 |
|           |        |                      | Pearson  | 0.846 | 0.908 | 0.918 | 0.834 | 0.886 |
|           |        |                      | R² | MIC | 0.786 | 0.825 | 0.621 | 0.467 | 0.829 |
|           |        |                      | Pearson  | 0.715 | 0.824 | 0.843 | 0.695 | 0.785 |
|           |        |                      | RMSE | MIC | 0.334 | 0.309 | 0.410 | 0.484 | 0.327 |
|           |        |                      | Pearson  | 0.354 | 0.314 | 0.374 | 0.456 | 0.295 |
| Semi-coherent(AI–1) |        |                      | R  | MIC | 0.786 | 0.575 | 0.331 | 0.694 | 0.406 |
|           |        |                      | Pearson  | 0.793 | 0.609 | 0.475 | 0.674 | 0.669 |
|           |        |                      | R² | MIC | 0.618 | 0.331 | 0.282 | 0.481 | 0.165 |
|           |        |                      | Pearson  | 0.629 | 0.371 | 0.226 | 0.455 | 0.448 |
|           |        |                      | RMSE | MIC | 0.107 | 0.198 | 0.153 | 0.192 | 0.209 |
|           |        |                      | Pearson  | 0.151 | 0.169 | 0.221 | 0.194 | 0.186 |
| Semi-coherent(AI–2) |        |                      | R  | MIC | 0.843 | 0.784 | 0.585 | 0.681 | 0.608 |
|           |        |                      | Pearson  | 0.531 | 0.858 | 0.379 | 0.793 | 0.783 |
|           |        |                      | R² | MIC | 0.710 | 0.614 | 0.342 | 0.463 | 0.370 |
|           |        |                      | Pearson  | 0.282 | 0.736 | 0.335 | 0.629 | 0.614 |
|           |        |                      | RMSE | MIC | 0.067 | 0.093 | 0.117 | 0.095 | 0.118 |
|           |        |                      | Pearson  | 0.123 | 0.075 | 0.105 | 0.085 | 0.080 |
to stabilize θ′ at elevated temperatures, and a hierarchy exists in such combinations. It was demonstrated that either Mn or Zr additions alone could provide stabilization of θ′ only to a certain extent; however, the critical temperature to which θ′ can be stabilized can significantly be elevated by adding both Mn and Zr, while also limiting or eliminating other elements. Hence, it would be worthwhile to experimentally test combinations of elements in the first quadrant in which Mn and Zr are located. A detailed discussion on identifying microalloying elements to θ′-Al2Cu at high temperatures in types of interfaces (and along various platelet positions) of θ′ precipitates. Based on the observed relationship here, we may propose the use of microalloying elements other than the ones that already have been experimentally verified to improve the high-temperature strength and stability of cast Al-Cu alloys.

In selecting microalloying elements, one can consider a strategy of adding multiple elements to create a synergistic effect. Shyam et al. [1] showed experimentally that there is evident positive (as well as other negative) synergies in microalloying with more than one element to stabilize θ′ at elevated temperatures, and a hierarchy exists in such combinations. It was demonstrated that either Mn or Zr additions alone could provide stabilization of θ′ only to a certain extent; however, the critical temperature to which θ′ can be stabilized can significantly be elevated by adding both Mn and Zr, while also limiting or eliminating other elements. Hence, it would be worthwhile to experimentally test combinations of elements in the first quadrant in which Mn and Zr are located. A detailed discussion on identifying microalloying elements to θ′-Al2Cu at high temperatures in

Figure 7. Scatter plots for the five machine learning models (RF: random forest, LR: linear regression, NN: nearest neighbour, KR: kernel ridge, BR: Bayesian ridge). The x- and y-axes represent the actual and predicted solute segregation energies of 34 elements at the coherent interfaces (a) Al_i, (b) Al_{i-1}, and (c) Al_{i-2}.
between the DFT solute segregation energies and some select descriptors. Notably, these linear relationships were dominant at Al$_1$–2 at the coherent interface and Al$_1$ at the semi-coherent interface.

Afterward, we performed more systematic correlation and regression analyses to quantify the relationship between the solute segregation energies and descriptors at both coherent and semi-coherent interfaces using the MIC and PCC approaches. Overall, the two different correlation analysis methods were in good agreement with each other (see Figures 5 and 6). Both the MIC and conjunction with the physical mechanisms associated with the DFT segregation energy database is provided in Ref. [4].

### 3.2. Correlation analysis

We started by plotting all 17 descriptors with respect to the DFT segregation energies at three different platelets, Al$_1$, Al$_{1-1}$, and Al$_{1-2}$, at both the coherent and semi-coherent interfaces for visual analysis. Selected plots are shown in Figure 4. We identified a linear relationship between the DFT solute segregation energies and some select descriptors. Notably, these linear relationships were dominant at Al$_{1-2}$ at the coherent interface and Al$_1$ at the semi-coherent interface.

Afterward, we performed more systematic correlation and regression analyses to quantify the relationship between the solute segregation energies and descriptors at both coherent and semi-coherent interfaces using the MIC and PCC approaches. Overall, the two different correlation analysis methods were in good agreement with each other (see Figures 5 and 6). Both the MIC and
PCC approaches found that size- and volume-related descriptors (e.g., molar volume, atomic radius and DFT volume) are strongly correlated with solute segregation at both interfaces. A similar size effect – i.e. that large solutes tend to bind more strongly with vacancies in Al [11] and Mg [12], respectively – has previously been reported in DFT studies. This observation shows that a size argument can also be applied in solute segregation.

On the other hand, some elements exhibited a tendency to be segregated from the interface, even though they were smaller in size/volume than Al at the semi-coherent interface. This observation suggests that the size/volume effect is important, but it alone cannot fully explain the segregation behaviour of solute atoms at the semi-coherent interface. For example, solubilities of microalloying elements within θ', represented as mixing energies from SQS calculations in the Al and Cu sublattices (Almix(SQS) and Cu mix(SQS)), are strongly correlated with solute segregation energies. As shown in Figures 5 and 6, both MIC and PCC methods identified that size, and solubility-related descriptors (e.g. atomic radius, volume, DFT solute volume and SQSs) are strongly correlated with the segregation at Almix in the coherent and Al in the semi-coherent interface, respectively.

3.3. Machine learning

Choosing appropriate features to be included in the training of machine learning models can be of its own topic (feature selection). Our intention in the current work is to introduce an emerging computational approach to the structural materials community, rather than how to exercise it in detail. Hence, we arbitrarily selected ten top ranking features out of 17 from both the MIC and PCC analyses and used them to train the machine learning models. The performance of the correlation analyses and the accuracy of the machine learning models are summarized in Table 2. Scattered plots comparing the actual and predicted segregation energies of the coherent and semi-coherent interfaces are presented in Figures 7 and 8.

Overall, the quality of prediction was satisfactory only in select cases. This is mainly attributable to the lacking instances of 34 elements in each platelet (Al, Al1−1/2, and Al2−2 at both the coherent and semi-coherent interfaces). However, the agreement between the actual and predicted segregation energies was quite good even with the limited datasets. For example, most of the machine learning models predicted the segregation energies quite reliably regardless of the correlation analyses at the Al platelet of the semi-coherent interface, as shown in Figure 8(a). The machine learning predictions at the Al platelet of the coherent interface, as shown in Figure 7(a), were also quite good. Remarkably, Al platelets at both coherent and semi-coherent interfaces were analysed as having strong correlations with the respective descriptors, as shown in Figures 5 and 6 (see the third bars of each descriptor in both figures). It is encouraging that reliable data-driven predictive models can be developed in an automated fashion even with a relatively small number of pure theoretical calculations, if strong correlations exist between specific descriptors and materials properties. With recent progress in the data science and open source communities, advanced data analytics tools are now readily available to researchers in other domains. Although successful prediction of segregation energies is limited to a small number of cases, the authors contend that the presented workflow has tremendous potential for application in design and development of improved high-temperature alloys.

4. Conclusions

We considered a total of 34 elements as potential microsegregants to improve the stability of θ'-Al2Cu at elevated temperatures within the framework of high-throughput DFT supercell calculations. Several hundred defect supercells were computed in an ensemble to take advantage of the petascale computing resources at the OLCF. The populated large DFT dataset in the current work was fused with various atomistic descriptors and interrogated to reveal correlations. The correlation analysis results were used to select features to be used in the training of machine learning models. Even with a small training dataset, the select machine learning models developed could reliably predict segregation energies. The accuracy of machine learning models may be further improved by incorporating an expanded DFT dataset of other alloying elements and considering more atomistic features, which may have higher correlations with segregation energies. The computational workflow demonstrated in the current work can be applied to accelerate the design of other high-temperature alloy systems in which high-fidelity datasets can be rapidly populated via supercomputer calculations followed by modern data analytics.

Acknowledgements

This research was sponsored by the Laboratory Directed Research and Development Program of ORNL, managed by UT-Battelle, LLC, for the US DOE. This research used resources of the OLCF at ORNL, which is supported by the Office of Science of the US DOE under contract DE-AC05-00OR22725. Early calculations were supported by the US Department of Energy, Office of Energy Efficiency and Renewable Energy, Vehicle Technologies Office, Propulsion Materials Program.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Dongwon Shin http://orcid.org/0000-0002-5797-3423
References

[1] Shyam A, Roy S, Shin D, et al. Cast aluminum alloys with extreme microstructural stability at elevated temperature. Under Review.

[2] Kohn W, Sham LJ. Self-consistent equations including exchange and correlation effects. Phys Rev. 1965;140:A1133–A1138.

[3] Biswas A, Siegel DJ, Wolverton C, et al. Precipitates in Al–Cu alloys revisited: atom-probe tomographic experiments and first-principles calculations of compositional evolution and interfacial segregation. Acta Mater. 2011;59:6187–6204.

[4] Shin D, Shyam A, Lee S, et al. 2017. Solute segregation at the Al/θ′-Al2Cu interface in Al-Cu alloys. Acta Mater. 2017; 141: 327–340

[5] Kresse G, Furthmüller J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. Comput Mater Sci. 1996;6:15–50.

[6] Kresse G, Furthmüller J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. Phys Rev B. 1996;54:11169–11186.

[7] Kresse G, Joubert D. From ultrasoft pseudopotentials to the projector augmented-wave method. Phys Rev B. 1999;59:1758–1775.

[8] Perdew JP, Wang Y. Accurate and simple analytic representation of the electron-gas correlation energy. Phys Rev B. 1992;45:13244–13249.

[9] Perdew JP, Burke K, Ernzerhof M. Generalized gradient approximation made simple. Phys Rev Lett. 1996;77:3865–3868.

[10] Zunger A, Wei S-H, Ferreira LG, et al. Special quasirandom structures. Phys Rev Lett. 1990;65:353–356.

[11] Wolverton C. Solute-vacancy binding in aluminum. Acta Mater. 2007;55:5867–5872.

[12] Shin D, Wolverton C. First-principles study of solute–vacancy binding in magnesium. Acta Mater. 2010;58:531–540.

[13] Reshef DN, Reshef YA, Finucane HK, et al. Detecting novel associations in large data sets. Science. 334:1518–1524.

[14] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine Learning in Python. J Mach Learn Res. 2011;12:2825–2830.

[15] Neter J, Kutner MH, Nachtsheim CJ, et al. 1996. Applied linear statistical models vol 4. Chicago: Irwin.

[16] Yu K, Tresp V, Schwaighofer A. Learning Gaussian processes from multiple tasks. Proc Int Conf Gaussian processes. 2005;22:1012–1019.

[17] Zhang Y, Duchi J, Wainwright M. Divide and conquer kernel ridge regression. Conf Learn Theory. 2013;30:1–26.

[18] Cherkassky V, Ma Y. Comparison of model selection for regression. Neural Comput. 2003;15:1691–1714.

[19] Segal MR. Machine learning benchmarks and random forest regression. Biostatistics. 2004: 1–14.

[20] Yang C, Zhang P, Shao D, et al. The influence of Sc solute partitioning on the microalloying effect and mechanical properties of Al-Cu alloys with minor Sc addition. Acta Mater. 2016;119:68–79.

[21] Chen BA, Liu G, Wang RH, et al. Effect of interfacial solute segregation on ductile fracture of Al-Cu-Sc alloys. Acta Mater. 2013;61:1676–1690.