CALPHAD: Computer Coupling of Phase Diagrams and Thermochemistry 72 (2021) 102231

Contents lists available at ScienceDirect

Calphad

journal homepage: http://www.elsevier.com/locate/calphad

Towards high-throughput microstructure simulation in compositionally complex alloys via machine learning

Yue Li [a,b,c,1], Bjørn Holmedal [c], Boyu Liu [a], Hongxiang Li [a,***], Linzhong Zhuang [a], Jishan Zhang [a], Qiang Du [b,*], Jianxin Xie [a,**]

[a] Beijing Advanced Innovation Center for Material Genome Engineering, State Key Laboratory for Advanced Metals and Materials, University of Science and Technology Beijing, 100083, Beijing, China
[b] SINTEF Industry, 0314, Oslo, Norway
[c] Norwegian University of Science and Technology, 7491, Trondheim, Norway

ARTICLE INFO

Keywords:
Materials informatics
Machine learning
High-throughput computing
Microstructure simulation
Tabulation

ABSTRACT

The coupling of computational thermodynamics and kinetics has been the central research theme in Integrated Computational Material Engineering (ICME). Two major bottlenecks in implementing this coupling and performing efficient ICME-guided high-throughput multi-component industrial alloys discovery or process parameters optimization, are slow responses in kinetic calculations to a given set of compositions and processing conditions and the quality of a large amount of calculated thermodynamic data. Here, we employ machine learning techniques to eliminate them, including (1) intelligent corrupt data detection and re-interpolation (i.e. data purge/cleaning) to a big tabulated thermodynamic dataset based on an unsupervised learning algorithm and (2) parameterization via artificial neural networks of the purged big thermodynamic dataset into a non-linear equation consisting of base functions and parameterization coefficients. The two techniques enable the efficient linkage of high-quality data with a previously developed microstructure model. This proposed approach not only improves the model performance by eliminating the interference of the corrupt data and stability due to the boundedness and continuity of the obtained non-linear equation but also dramatically reduces the running time and demand for computer physical memory simultaneously. The high computational robustness, efficiency, and accuracy, which are prerequisites for high-throughput computing, are verified by a series of case studies on multi-component aluminum, steel, and high-entropy alloys. The proposed data purge and parameterization methods are expected to apply to various microstructure simulation approaches or to bridging the multi-scale simulation where handling a large amount of input data is required. It is concluded that machine learning is a valuable tool in fueling the development of ICME and high throughput materials simulations.

1. Introduction

The demand for advanced materials has been ever-increasing, but the time to develop, produce, and deploy a new material is often 20 years or more [1,2]. The rapidly evolving high-throughput computing (HTC) technique and Integrated Computational Material Engineering (ICME) are widely adopted in the materials field and are regarded as promising solutions for accelerating this timeframe [2–5]. One of the prerequisites for HTC and ICME to succeed in materials discovery and process optimization lies in the robustness and computational efficiency in the individual task so that a large number of tasks can be easily integrated and executed within a feasible time scale.

Microstructure simulation, built on the coupling of computational thermodynamics and kinetics, is a crucial part of any ICME framework. Up to now, various microstructure simulation approaches have been put forward, including the phase field approach which can provide a detailed description of the spatial distributions of microstructure features [6–8], and the Kampmann–Wagner numerical (KWN) model which

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail addresses: hxli@skl.ustb.edu.cn (H. Li), qiang.du@sintef.no (Q. Du), jxxie@ustb.edu.cn (J. Xie).
[1] Yue Li is currently working at Max-Planck-Institut für Eisenforschung GmbH, Düsseldorf, Germany.

https://doi.org/10.1016/j.calphad.2020.102231
Received 27 August 2020; Received in revised form 3 November 2020; Accepted 22 November 2020
Available online 3 December 2020
0364-5916/© 2020 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

can address multi-scale multi-component industrial problems due to its computation efficiency [9–11]. Nevertheless, these microstructure modeling approaches require access to a large amount of thermodynamic data, usually in the format of phase diagram data. These data are usually generated by CALPHAD software such as Thermo-Calc or Pandat software [12–14]. Often one can call the CALPHAD software and calculate thermodynamic data in every grid point and at every time step as the strategy adopted by DICTRA (Thermo-calc's diffusion phase transformation module) and MICRESS (an implementation of phase field method) software [15,16]. However, with this so-called in-situ coupling

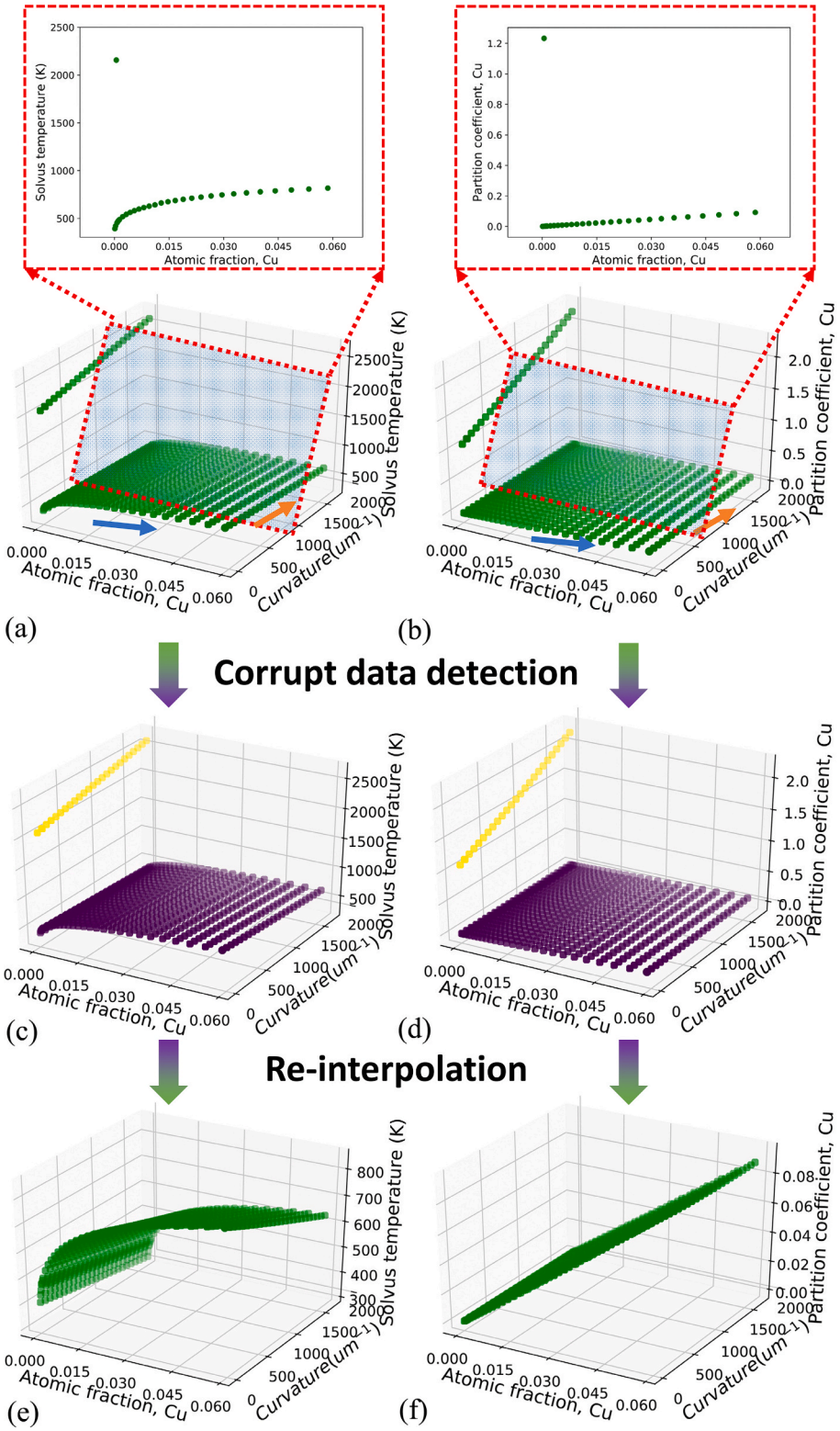


Fig. 1. Tabulated thermodynamic dataset of $\theta(Al_2Cu)$ phase in Al-Cu alloys for aging simulations: (a) and (b) are the raw data corresponding to solvus temperature and Cu partition coefficient, respectively; (c) and (d) are the corresponding datasets after intelligent detection (yellow points represent corrupt data); (e) and (f) are the corresponding re-interpolation data. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

it is time-consuming in handling multi-component industrial alloy. A popular approach for efficient data access is the tabulation or mapping file technique [17–20] which is to generate these thermodynamic data in pre-defined grid points and then to apply a multilinear interpolation procedure to calculate the values among these grid points. Due to its efficiency, this technique has been applied not only to the phase field models [21,22] and KWN models [10,18,23] but also in many other disciplines, e.g., combustion modeling [19,24], indoor airflow simulation in building application [25], and organ simulation in biology [26].

However, when incorporating these tabulated thermodynamic data into a microstructure model, several challenges are encountered. Firstly, the numerical quality of these tabulated data such as smoothness must be guaranteed, which is essential for the convergence of the iterative algorithms used in the numerical solution of microstructure models. Fig. 1 (a) and (b) show the tabulated Gibbs-Thomson phase diagram data of the $\theta(Al_2Cu)$ phase in the Al-Cu system for an aging treatment simulation [10,27], which is generated by the TQ module of Thermo-Calc software (TCAl 4.0 database). It contains four axes, i.e., curvature, Cu concentration, solvus temperature, and Cu partition coefficient. Note that the curvature is the reciprocal of the radius of a precipitate and will affect its growth kinetic in microstructure simulation via the Gibbs-Thomson effect [10]. Clearly, some incorrect or corrupt data exists at very low Cu concentrations shown in Fig. 1 (a) and (b), which would hinder the subsequent iterative calculations if the calculations need access to these data points. Similar corrupt data is also found in a complex multi-component Fe-based system as mentioned later. The occurrence of these corrupt data is due to that some of the boundary was not the focus when building the corresponding thermodynamic database. Note that the occurrence of these corrupt data is rare thanking to the great efforts from the CALPHAD community but still not avoidable owing to the extrapolation algorithms from simple to complex systems used in the CALPHAD software to perform thermodynamic computations. The corrupt data would be more likely to occur in a compositionally complex alloy system whose tabulation file has higher dimensions and a larger amount of data. Due to the sheer amount of these data (usually up to hundreds of MB in binary format file), it is impossible to manually identify and remove these corrupt data. Up to now, an intelligent automated thermodynamic data purge/cleaning procedure is still lacking, which will be treated in this work. It should be noted that the previous treatment is either to try different initial parameters in CALPHAD software or to optimize the thermodynamic database.

Moreover, although the tabulation technique greatly shortens the simulation time as compared to the direct coupling method, the demand for computer physical memory is high because a large amount of data must be loaded into the computer memory. For example, as mentioned later in the case studies, the required computer memory is increased by 11700% (the size of the called binary file is 1587 MB) when simulating the solidification process of an Al-Co-Cr-Fe-Ni-Ti high entropy alloy using the KWN model as compared to the direct coupling method. This limits the application of the tabulation technique towards HTC in which shared memory mode parallel computing is often adopted. Thus, it is valuable if an efficient and low-memory-usage coupling technique is developed to bridge this identified gap. A straightforward method is to parameterize the tabulation data into an equation. Artificial neural networks (ANN), as a powerful regression algorithm, can handle this task, as demonstrated by Strandlund more than 10 years ago [12]. However, given the rapid recent progress in ANN, it is justifiable to re-visit this topic and explore what the latest ANN technique can offer on data purge and regression analysis within the context of microstructure modeling.

Here we report our exploration with materials informatics [1,28] to solve the above-mentioned two bottlenecks. A flowchart of the proposed machine learning framework is shown in Fig. 2. Firstly, a machine learning algorithm is adopted to make intelligent data purge, including corrupt data detection and recalibration/re-interpolation. Then, these recalibrated data are trained by ANN, and a non-linear equation with a small parameterization dataset (only about 4 KB) is obtained to represent the whole tabulated data. Note that the calibration is conducted outside of the CALPHAD software. Finally, the equation with the parameterization dataset is coupled with the microstructure model using a straightforward method to enable high-throughput microstructure simulations. Successful case studies are presented in three compositionally complex alloy systems, including aluminum, steel, and high-entropy alloys.

2. Methods

An unsupervised algorithm, which is named “Density-based spatial clustering of applications with noise” (DBSCAN) [29], is applied to purge data. The ANN and DBSCAN algorithms, as implemented in Tensorflow software [30], is adopted. The obtained non-linear equation with the parameterization dataset is coupled with the previously developed solidification and precipitation KWN models [9,23] using a FORTRAN-based Dynamic-Link Library with the “Plug and Play” feature. Almost all figures in the paper are made using the matplotlib [31].

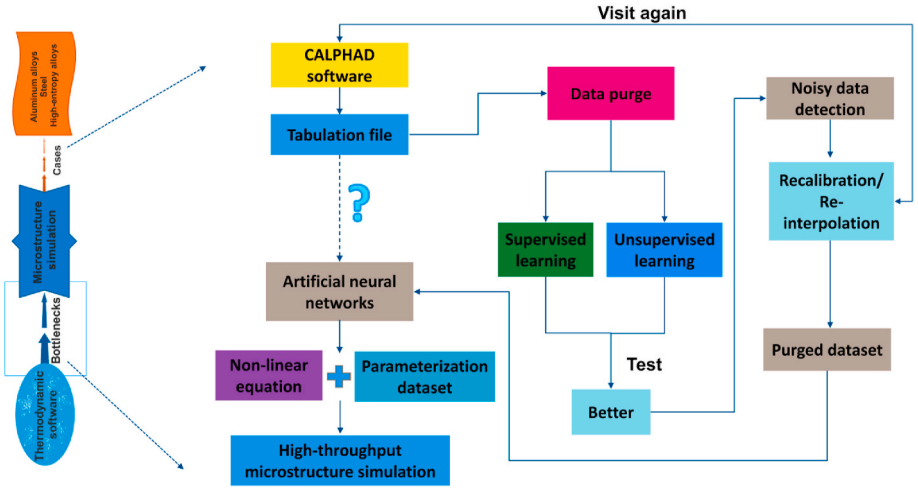


Fig. 2. Flowchart of the proposed machine learning framework to enable high-throughput microstructure simulations in compositionally complex alloy systems.

3. Results

3.1. Data purge

We firstly take binary Al-Cu alloy as an example to illustrate the methodology and then extend it to multi-component alloy systems. As shown in Fig. 1 (a) and (b), the tabulated phase diagram data are divided into two clusters: one with few data points, called corrupt data, and the other with the majority data points, called normal data. The corrupt data, if not filtered, have a great influence on regression analysis using ANN (the detail of the ANN algorithm is introduced later) as shown in Fig. S1 in the Supplementary Information. Hence, the data purge is necessary before making the regression analysis using ANN.

There are two mainstream methodologies to perform clustering analysis, i.e. supervised and unsupervised learning algorithms. For the former, represented by the "Support Vector Machine" (SVM) algorithm [32], some pre-labeled training data are required. However, the good choice of the labeled training data, which still requires heavy human intervention (see Fig. S2 in the Supplementary Information), is critical for the good performance of this algorithm. In contrast, unsupervised learning algorithms do not need to label data in advance. Here, the unsupervised algorithm DBSCAN is applied to the Al-Cu system. It is a density-based classification algorithm and only two hyperparameters need to be determined: the ε (Eps) neighborhood of every point and the minimum number of points required to form a dense region (minPts). This algorithm consists of the following steps to deal with the tabulated data of $\theta(Al_2Cu)$ phase in Al-Cu system (Fig. 1):

1) Input the entire tabulation dataset, and choose the data with the curvature equal to zero ($K_1 = 0$);
2) Choose two columns, i.e., Cu content and solvus temperature, as shown in Fig. 1 (a);
3) Make standardization using StandardScaler [29,33] on the data in the two columns, and use the DBSCAN algorithm to detect all corrupt data by tuning the values of ε (Eps) and minPts, as shown in Fig. 1 (c); It should be noted that StandardScalar is to standardize features by removing the mean and scaling to unit variance [33].
4) Make inverse standardization for the two columns and repair the corrupt data (note that the solvus temperature and Cu partition coefficient are repaired simultaneously as explained later), as shown in Fig. 1 (e) and (f);
5) Choose the next group data with the curvature of K_{i+1} ($K_{i+1} = K_i + \Delta K$), and go to steps 2–4;
6) Until all curvatures are chosen, input the repaired dataset, and choose the data with the minimum Cu content (C_1), as shown in Fig. 1 (a);
7) Choose two columns, i.e., curvature and solvus temperature, and perform the same program of steps 3–4;
8) Choose the next group data with the Cu content of C_{i+1} ($C_{i+1} = C_i + \Delta C$), and go to step 7;
9) Until all Cu contents are chosen, a complete repaired dataset is obtained.

Note that the corrupt data simultaneously occur at the axes corresponding to the solvus temperature and Cu partition coefficient in this example (Fig. 1 (c) and (d)). Thus, the two axes are scanned simultaneously, and only two cycles are performed to complete the data purge of this binary system. To ensure that all corrupt data can be detected, the two hyperparameters are adjusted each cycle. A two-step algorithm is proposed to make data repair. The first step is to manually calculate the data corresponding to corrupt points by the console mode of Thermo-Calc software instead of its TQ module. The first step also provides an interface for inputting data from other sources. If the first step does not work, a linear interpolation/extrapolation will be performed to overwrite the corrupt data. The purged tabulated data are shown in Fig. 1 (e) and (f). Note that the change of steps 1–5 and 6–9 will not affect the result.

3.2. Parameterization of tabulated data based on ANN

We previously had employed the table look-up and interpolation procedure to achieve efficient access to these thermodynamic quantities in microstructure modeling [10,34,35]. However, as mentioned above, this efficient access is at the expense of large computer physical memory usage, which would put restrictions when the microstructure simulation was embedded in a parallel-running-mode high-throughput ICME simulation. If the tabulated thermodynamic dataset could be parameterized, e.g., the solvus temperature (y_1) and Cu partition coefficient (y_2) can be expressed by the Cu content (x_1) and curvature (x_2), less memory would be demanded in a HTC case study. The ANN algorithm can perform this parameterization as demonstrated below.

As shown in Fig. 3, a fully connected three-layer ANN is shown, including one input layer, one hidden layer, and one output layer. The number of input and output nodes are defined by the problem to be solved while the number of hidden nodes is arbitrary. Assuming there are N input nodes, J hidden nodes, and K output nodes, the non-linear regression function is expressed as

$$y_k = \theta_k + \sum_{j=1}^{J} W_{jk}\, sigmoid\left(\theta_j + \sum_{n=1}^{N} \omega_{nj} x_n\right) \tag{1.1}$$

$$sigmoid(x) = 1/(1+\exp(-x)) \tag{1.2}$$

The *sigmoid* function is the used activation function. The sets of weights W_{jk} and ω_{nj} and the sets of biases θ_k and θ_j can be determined by training the ANN. When one hidden layer is chosen, the total number of weights and biases is $(N+1)J + (J+1)K$. The neural network can be trained by using $x_1, x_2, \ldots, x_N$ to reproduce the target values, $t_1, t_2, \ldots, t_K$. The loss function is given by

$$E = \frac{1}{2S} \sum_{s=1}^{S} \sum_{k=1}^{K} (y_{ks} - t_{ks})^2 \tag{2}$$

where S is the total training number of each target value, t_k. We choose the conventional backpropagation algorithm [36] combined with the gradient descent optimizer to update the weights and biases (the learning rate is set as 0.1). 70% of the whole data is used as the training dataset, and the remaining as the test dataset. It is worth noting that the training and test datasets have been completely disordered, and standardized using StandardScaler before ANN training.

The ANN algorithm is applied to train the tabulated dataset of the θ

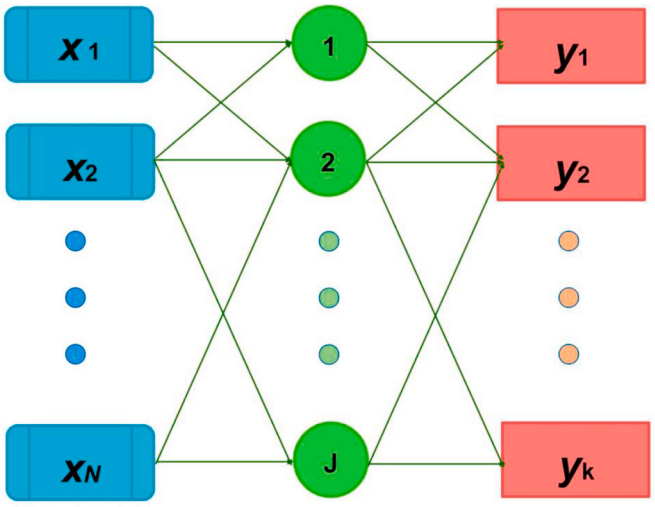


Fig. 3. Schematic diagram of the used three-layer ANN.

phase in the Al-Cu system, and the comparisons of the predictions and target values are shown in Fig. 4. Note that the predictions contain both the output results using the training and test datasets. Good consistency can be found between the targets and regression results with both the training loss and test loss (Equation (2)) up to approximately 0.007. Moreover, the test loss is always slightly higher than the training loss in all iterations (Fig. 4 (e)), suggesting that there is no overfitting.

After making the ANN regression, we can create a parameterization dataset combined with **Equation** 1.1 to express the discrete tabulated phase diagram data points with good accuracy. The parameterization dataset consists of four parts: weights (W) and biases (θ), structural parameters (N, J, and K), activation function, and standardization parameters. The size of the parameterization dataset is only about 4 KB, which is far smaller than that of the tabulated dataset (391 KB in binary form). The data compression will be more significant in a multi-component alloy system.

3.3. The generalization to multi-component alloy systems

The current proposed thermodynamic data purge and parameterization techniques can be generalized to multi-component alloy systems. For the system with N components, N cycles are performed to check its data quality if corrupt data occur at y_1, y_2, …, y_K simultaneously like Fig. 1. This data purge method is further applied to some multi-component alloy systems, including non-stoichiometric FCC phase precipitating from the liquid in the Al-Zn-Mg-Cu system (from Thermo-Calc, TCAl 4.0 database), non-stoichiometric FCC phase precipitating from BCC matrix in the Fe-Cu-Al-Mn-Ni system (from Thermo-Calc, TCFe 8.0 database), and non-stoichiometric FCC phase precipitating from the liquid in the Al-Co-Cr-Fe-Ni-Ti high entropy system (from Thermo-Calc, TCFe 8.0 database). The results are shown in Fig. 5. For the Al-Zn-Mg-Cu system, no corrupt data are found. Thus, the ANN algorithm ($N = 4$, $J = 10$, and $K = 4$) is directly applied to the raw thermodynamic dataset (an example of this dataset is shown in Fig. S3), and the comparisons of the predictions and target values are shown in Fig. 5 (a). The regression results match well with the targets with both

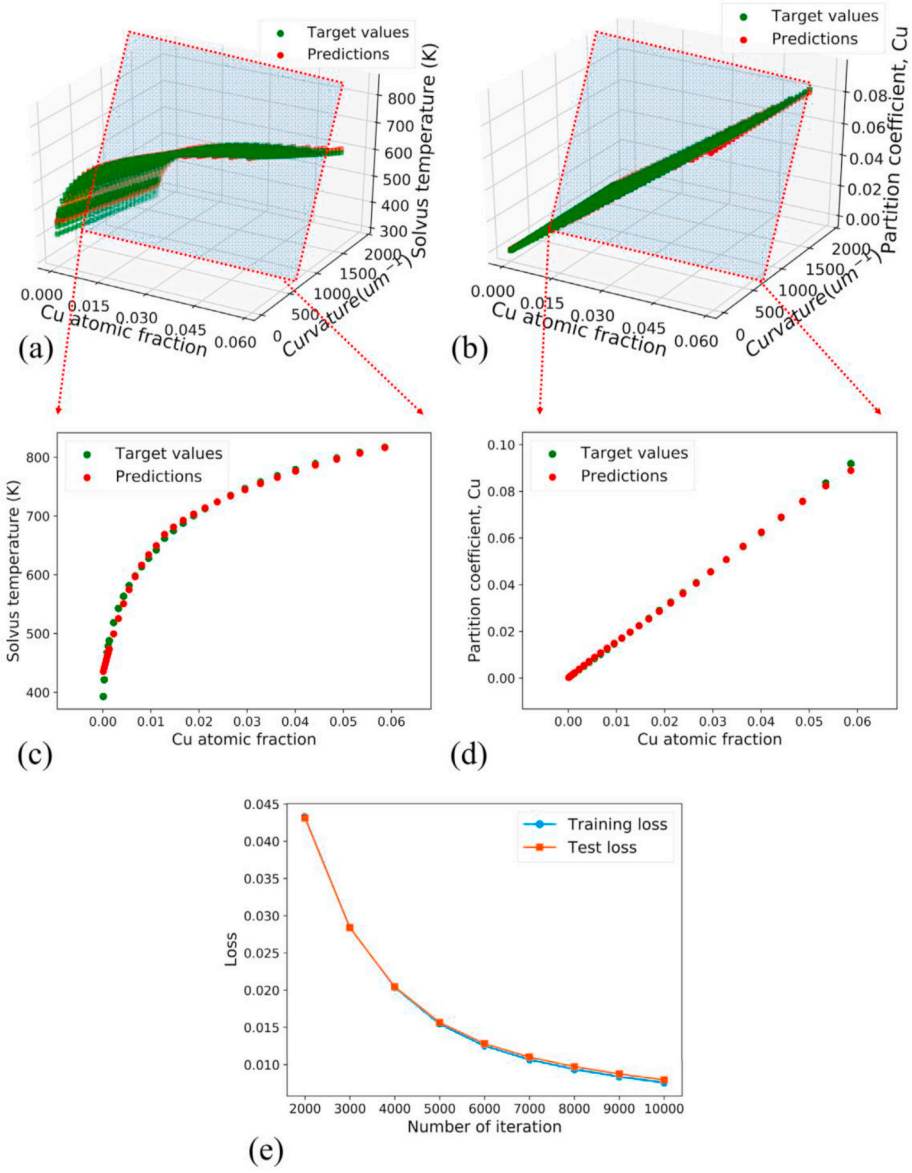


Fig. 4. Comparison of predictions and target values based on the ANN algorithm for θ phase in Al-Cu alloys: data corresponding to (a, c) solvus temperature and (b, d) Cu partition coefficient; and (e) the variances of the training and test loss values (after standardization) with the iteration number. Note that the used ANN has N = 2 input nodes, J = 10 hidden nodes, and K = 2 output nodes.

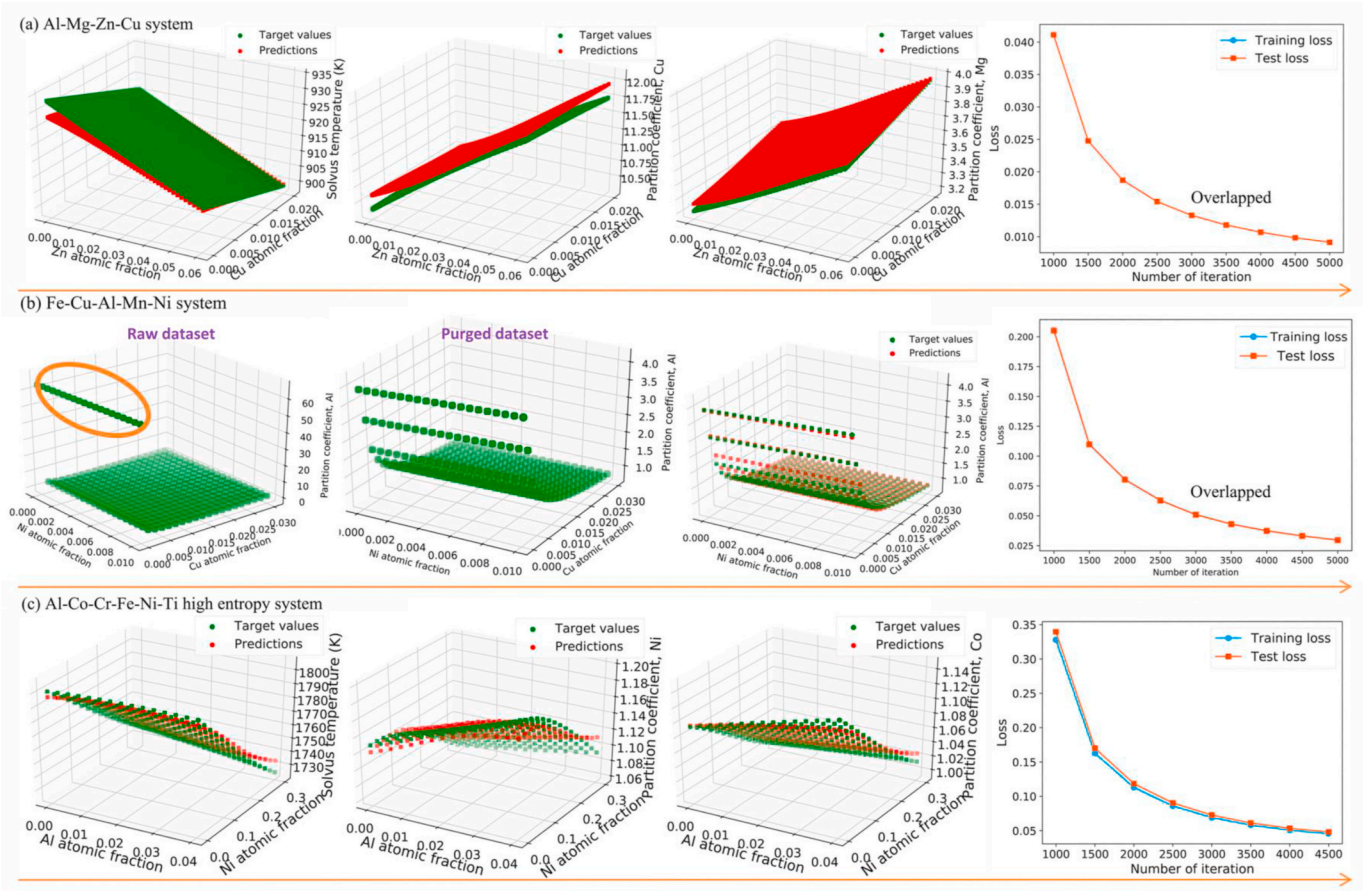


Fig. 5. Part of the tabulated thermodynamic datasets and the corresponding data purge and parameterization results of the non-stoichiometric (a) FCC phase precipitating from liquid matrix in Al-Zn-Mg-Cu, (b) FCC phase precipitating from BCC matrix in Fe-Cu-Al-Mn-Ni, and (c) FCC phase precipitating from liquid matrix in Al-Co-Cr-Fe-Ni-Ti systems. Note that the solvus temperature in (a) and (c) actually corresponds to liquidus.

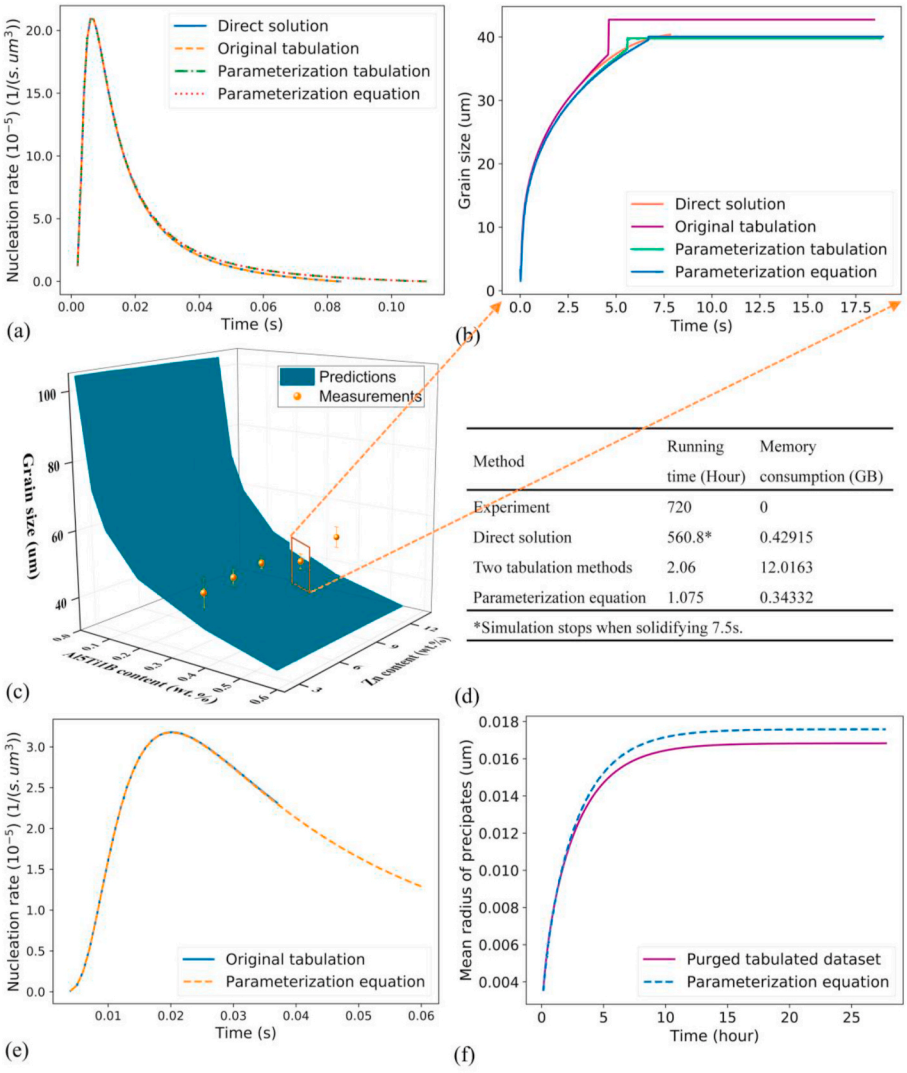
the training loss and test loss below 0.01. For the Fe-Cu-Al-Mn-Ni system (Fig. 5 (b)), some corrupt data exist in its tabulated dataset. The proposed data purge and parameterization techniques have been applied, and the corresponding purged and ANN regression ($N = 5, J = 10$, and $K = 5$) results with the training loss and test loss being about 0.025 are shown in Fig. 5 (b). For the Al-Co-Cr-Fe-Ni-Ti high entropy system (Fig. 5 (c)), no corrupt data exist, which could be because the phase diagram above solidus is simpler to be calculated than that below solidus for an alloy. So the ANN algorithm ($N = 6, J = 25$, and $K = 6$) has been applied to deal with the raw dataset, the training and test loss being about 0.05. Note that the tabulated datasets of these non-stoichiometric phases are more complex than those stoichiometric phases (like the θ phase in the Al-Cu system), which further suggests the power of the adopted ANN regression. After making the ANN regression, the corresponding parameterization datasets are generated to represent these tabulation data. The sizes of the tabulation files (binary form) of Al-Zn-Mg-Cu, Fe-Cu-Al-Mn-Ni, and Al-Co-Cr-Fe-Ni-Ti systems are significantly reduced from 791 MB, 819 MB, and 1587 MB to about 4 KB, respectively.

3.4. *The application of the data purge and parameterization method in microstructure modeling*

After obtaining the base functions and parameterization datasets, one can couple them with different microstructure models via a Dynamic-Link Library file in the so-called "Plug and Play" manner. Here, as an example, we apply the proposed method to the KWN model to evaluate the feasibility and demonstrate the advantages. Note that the KWN model was previously coupled with the CALPHAD thermodynamic data via the tabulation technique [37]. Three cases are performed, including as-cast grain size predictions in the inoculated Al-*x*Zn-2Mg-2Cu alloys (wt.%) and $Al_{3.32}Co_{27.27}Cr_{18.18}Fe_{18.18}Ni_{27.27}Ti_{5.78}$ high entropy alloy (at.%) [38], and aging precipitation of spherical FCC phases from BCC matrix at 900 K in the Fe-1Cu-1Al-0.5Mn-0.5Ni alloy (at.%).

The previously-developed KWN model for as-cast grain size prediction [23] is applied to the Al-*x*Zn-2Mg-2Cu alloys inoculated by different amounts of Al-5Ti-1B grain refiners. The measured effective cooling rate is about 67 K/s [39]. The input parameters for this simulation are listed in Table S1 in the Supplementary Information. Four data access techniques are performed, including the direct coupling method, using the original tabulation file, using the new tabulation file generated from the parameterization equation, and using the parameterization equation. The comparisons are shown in Fig. 6 (a) and (b). For the nucleation stage (Fig. 6 (a)), the four methods predict almost the same results, suggesting a good accuracy of the proposed parameterization technique. This is attributed to the quite low regression loss values shown in Fig. 5 (a). For the growth stage (Fig. 6 (b)), the two tabulation methods cannot run smoothly as compared to the direct coupling and parameterization methods. This is because the iterations have accessed the data which are beyond the range of the tabulation file. Note that the simulation cannot continue to run after about 7.5s using the direct coupling, as shown in Fig. 6 (b). This may be because the iterations have called the corrupt data points which are not met in the generated tabulation file. The grain sizes predicted by the four methods are almost the same (about 40 μm) and match well with the measured result (55μm) [39].

Fig. 6 (c) shows the results of as-cast grain size predictions using the proposed parameterization technique (including 30 cases). It is found



Method	Running time (Hour)	Memory consumption (GB)
Experiment	720	0
Direct solution	560.8*	0.42915
Two tabulation methods	2.06	12.0163
Parameterization equation	1.075	0.34332

*Simulation stops when solidifying 7.5s.

Fig. 6. Evolutions of (a) nucleation rate and (b) grain size with time during solidification in Al-9Zn-2Mg-2Cu alloy inoculated by 0.4 wt% Al5Ti1B master alloys using four different thermodynamic datasets visiting methods; (c) as-cast grain size predictions of Al-*x*Zn-2Mg-2Cu alloys (x=2–12) inoculated by different amounts of Al-5Ti-1B master alloys using the proposed parameterization technique (30 cases are performed and some experimental results [39] are also given); (d) comparisons of computation performance in predicting the results of (c) using four different thermodynamic data visiting methods (the period to make so experiments is estimated as 30 days); (e) evolutions of nucleation rate with time during solidification in $Al_{3.32}Co_{27.27}Cr_{18.18}Fe_{18.18}Ni_{27.27}Ti_{5.78}$ high entropy alloy using different thermodynamic data visiting methods; (f) mean radius of spherical FCC phases precipitating from BCC matrix aging at 900K in the Fe-1Cu-1Al-0.5Mn-0.5Ni alloy (at.%).

that the predictions match well with the experimental results [39]. Fig. 6 (d) shows the corresponding computation performance using various techniques. Firstly, compared with the direct method, the running time is decreased by 99% using the two tabulation methods, while the computer memory usage is increased by 2600%, which is attributed to loading the large tabulation file into the memory. Secondly, the proposed parameterization technique not only makes the running time decreased by 99.8% as compared to the direct method but also keeps a very low usage of computer memory which is decreased by 97% as compared to the two tabulation methods. Note that the binary tabulation file is significantly compressed down to only 0.004‰ of the original size using the parameterization technique.

The KWN model is further applied to the $Al_{3.32}Co_{27.27}Cr_{18.18}Fe_{18.18}Ni_{27.27}Ti_{5.78}$ high entropy alloy (at.%) [38] to predict the grain nucleation during solidification. The input parameters for this simulation are listed in Table S2 in the Supplementary Information. A comparison of the predicted nucleation rates using the tabulation and parameterization techniques is shown in Fig. 6 (e). Obviously, their predictions match well with each other, and the consumed computer memory is significantly reduced by 99% using the parameterization equation as compared to that using the tabulation technique.

The KWN model [9] is applied to simulate the precipitation of spherical FCC phases from BCC matrix aging at 900 K in the Fe-1Cu-1Al-0.5Mn-0.5Ni alloy (at.%). The input parameters for this simulation are listed in Table S3 in the Supplementary Information. When the direct coupling method or original tabulation file was applied, the simulation broke down, which is attributed to the existence of corrupt data as shown in Fig. 5 (b). However, the simulation ran well using the purged tabulation file or the parameterization equation. The comparisons of their simulation results are shown in Fig. 6 (f). Clearly, they match well with each other, which is due to the good regression accuracy (Fig. 5 (b)). Compared with using the purged tabulation file, the consumed computer memory is significantly reduced by 98% using the parameterization equation. It is believed that the computation speed using the parameterization equation would be significantly increased as compared to using the direct coupling method assuming the simulation using the direct coupling method could run well.

4. Discussion

In this study, data purge and parameterization techniques are proposed, which ensures the quality of thermodynamic data and enables the microstructure simulation of compositionally complex alloys efficiently and thus paves the way for the ICME-guided high-throughput materials discovery and process parameters optimization. The existence of corrupt thermodynamic data in compositionally complex alloy systems largely hinders microstructure simulation, such as in the above-mentioned Fe-1Cu-1Al-0.5Mn-0.5Ni case. We also intentionally select Al-Co-Cr-Fe-Ni-Ti high entropy system (Fig. 5 (c)), in which by chance no corrupt data exists after several trial and error phase diagram calculations. The

proposed data purge technique can intelligently and efficiently identify corrupt data points in multi-component alloy systems, and then in situ repair these data by interpolation/extrapolation with few human interventions. Usually what users need to do is to specify two suitable hyperparameters for the DBSCAN algorithm in every cycle, i.e., ε (Eps) and minPts. The purged tabulated dataset is further trained using a three-layer ANN, and a non-linear equation (Equation (1.1)) with a parameterization dataset can be obtained to represent a large amount of tabulated data with very high accuracy. The proposed parameterization method based on machine learning solves the seemingly irreconcilable dilemma, i.e., the efficiency issue associated with the direct coupling method and the memory issue associated with the tabulation method. Note that the ANN training process indeed requires substantial computational resources for a compositionally complex alloy system, but according to the authors' rich experience in the microstructure simulation field [40] it is worth doing for multi-component systems because this not only greatly saves much time and computational resources in adjusting model parameters and performing a high-throughput microstructure simulation but also effectively eliminates the interference originating from continuity and boundedness questions, as mentioned below. Note that for one alloying system, the obtained parameterization dataset can be applied to deal with different compositions and heat treatment histories.

Apart from the excellent computation performance in terms of high computation speed and low memory cost, the proposed parameterization method also poses other benefits, as shown in Fig. 6 (b). This is attributed to the continuity and boundedness of this obtained non-linear equation. The two features are also required by the microstructure simulation where the iterations are applied. The first is to verify its continuity. It has been proved in mathematics that any derivable function must be continuous at every point in its domain [41]. The derivable function of Equation (1.1) is expressed by:

$$y'(n,k) = \sum_{j=1}^{J} sigmoid\left(\theta_j + \sum_{n=1}^{N} \omega_{nj} x_n\right) \omega_{nj} W_{kj}^{T} \tag{3}$$

where $y'(n,k)$ is the derivative of y_k in Equation (1.1) with respect to the variable x_n. W^T is the transpose of the matrix W. Clearly, $y'(n,k)$ is the sum of several $sigmoid(x)$ functions with the number of J. It is easy to prove that these $sigmoid(x)$ functions in Equation (3) are derivable. Therefore, their sum is derivable [41] and finally, Equation (1.1) is a continuous function. The second is to verify the boundedness of Equation (1.1). These $sigmoid(x)$ functions are bounded and their linear sum is bounded [41]. Fig. S4 in the Supplementary information gives an intuitive description of the two features of the non-linear equation in the Al-Mg-Zn system. When the solute concentrations are far beyond the maximum solute concentrations in the original tabulation file, the predicted values are well controlled in certain zones. Moreover, continuity is also observed easily in this example.

The previous tabulation technique only generates the tabulated thermodynamic quantities data in certain ranges and an out-of-range extrapolation can return with negative values for the partition coefficient. Thus the simulation will become discontinuous and even break down (like Fig. 6 (b)) when the calculation occasionally attempts to access the data points beyond these ranges during an iteration. Owing to the proved continuity and boundedness features of this obtained non-linear equation (like Fig. S4), the proposed parameterization technique can generate reasonable values in numerical sense beyond the training concentration ranges, and thus makes an out-of-bound iteration smoother and run without breaking down like Fig. 6 (b) using the proposed parameterization technique. Besides the high computational speed, very low memory consumption, and high robustness, data compression with rather a high accuracy is another advantage of the proposed parameterization technique, which is beneficial for the transfer of thermodynamic data among different computers.

The Dynamic-Link Library file plays an important role in coupling the obtained non-linear equation and parameterization dataset with a microstructure model. Its "Plug and Play" feature makes researchers easily modify it to different alloy systems. This programming technique is applicable to other microstructure simulation methods [6–8].

As compare to Strandlund's work more than 10 years ago [12], the present work firstly verified that data purge is necessary before making regression analysis using ANN. Secondly, the continuity and boundedness features of the obtained non-linear equation by ANN are highlighted. By comparing with other methods, it is pointed out that the two features make iteration smoother and run without breaking down. Thirdly, the present work highlights these proposed methods by applying them to microstructure simulations of compositionally complex alloys. The high computational robustness, efficiency, and accuracy, which are prerequisites for high-throughput computing, are verified by a series of case studies.

Some improvements should be made in the future. Firstly, the DBSCAN algorithm is only applied to two-dimensional space and thus several cycles are performed to finish the data purge. Whether a multi-dimensional density-based clustering algorithm could be developed is worth exploring. Secondly, it is necessary to optimize the ANN algorithm to lower the computational cost and increase training speed in compositionally complex alloy systems, such as using stochastic gradient descent (SGD) [42]. Thirdly, although some case studies are presented using the KWN model in this study, the proposed data purge and parameterization techniques are also applicable to the phase field approach and other fields using the tabulation technique as mentioned in the introduction. Finally, the access to CALPHAD software is not open enough and the thermodynamic models used in CALPHAD software to perform thermodynamic computation need to be further improved to generate higher quality tabulated thermodynamic datasets. All of these cannot be achieved without the co-efforts from the scientists in CALPHAD, microstructure modeling, and machine learning research communities.

5. Conclusions

In this paper, a machine learning framework is proposed which not only ensures the quality of the used thermodynamic datasets but also accelerates the response in a microstructure model to a given set of compositions and processing conditions with high robustness and computational accuracy. Moreover, the straightforward coupling method between the microstructure module and parameterization dataset eliminates the two bottlenecks restricting the application of ICME to multi-component alloy systems and bridges the gap between the ICME framework and HTC in compositional complex alloy systems. The proposed data purge and parameterization methods are expected to apply to other microstructure simulation approaches like the phase field method or to bridging the multi-scale simulation where handling a large amount of input data is a prerequisite.

Declaration of competing interest

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

Acknowledgments

This work was supported by the Major State Research and Development Program of China (No. 2016YFB0300801) and the National Natural Science Foundation of China (No. 51671017 and No. 51971020). Y. Li is supported by the China Scholarship Council and Research Council of Norway as a joint training Ph. D student in SINTEF/NTNU, Norway. The authors would like to thank Dr. Cong Zhang from University of Science and Technology Beijing for providing TCFe 8.0 database.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.calphad.2020.102231.

Author contributions

Q. Du, J.X. Xie and H.X. Li planned the study. Y. Li and Q. Du implemented the data purge and parameterization algorithm and its dynamic link with the microstructure model. Y. Li and B.Y Liu performed the computational design calculations. All authors discussed the results and were involved in the writing of the manuscript, as well as taking accountability for all aspects of the work in this manuscript.

Data availability

The data required to reproduce these findings are available from the authors upon a reasonable request.

References

[1] G.J. Mulholland, S.P. Paradiso, Perspective: materials informatics across the product lifecycle: selection, manufacturing, and certification, Apl. Mater. 4 (5) (2016), 053207.

[2] S. Curtarolo, G.L.W. Hart, M.B. Nardelli, N. Mingo, S. Sanvito, O. Levy, The high-throughput highway to computational materials design, Nat. Mater. 12 (2013) 191.

[3] M.K. Horton, J.H. Montoya, M. Liu, K.A. Persson, High-throughput prediction of the ground-state collinear magnetic order of inorganic materials using Density Functional Theory, Npj Comput. Mater. 5 (1) (2019) 64.

[4] M. Aykol, S. Kim, V.I. Hegde, D. Snydacker, Z. Lu, S. Hao, S. Kirklin, D. Morgan, C. Wolverton, High-throughput computational design of cathode coatings for Li-ion batteries, Nat. Commun. 7 (2016) 13779.

[5] H. Hafiz, A.I. Khair, H. Choi, A. Mueen, A. Bansil, S. Eidenbenz, J. Wills, J.-X. Zhu, A.V. Balatsky, T. Ahmed, A high-throughput data analysis and materials discovery tool for strongly correlated materials, Npj Comput. Mater. 4 (1) (2018) 63.

[6] Y. Yang, O. Ragnvaldsen, Y. Bai, M. Yi, B.-X. Xu, 3D non-isothermal phase-field simulation of microstructure evolution during selective laser sintering, Npj Comput. Mater. 5 (1) (2019) 81.

[7] K. Kim, A. Roy, M.P. Gururajan, C. Wolverton, P.W. Voorhees, First-principles/Phase-field modeling of θ′ precipitation in Al-Cu alloys, Acta Mater. 140 (2017) 344–354.

[8] Z.-H. Shen, J.-J. Wang, Y. Lin, C.-W. Nan, L.-Q. Chen, Y. Shen, High-throughput phase-field design of high-energy-density polymer nanocomposites, Adv. Mater. 30 (2) (2018), 1704380.

[9] Q. Du, K. Tang, C.D. Marioara, S.J. Andersen, B. Holmedal, R. Holmestad, Modeling over-ageing in Al-Mg-Si alloys by a multi-phase CALPHAD-coupled Kampmann-Wagner Numerical model, Acta Mater. 122 (2017) 178–186.

[10] Y. Li, B. Holmedal, H. Li, L. Zhuang, J. Zhang, Q. Du, Precipitation and strengthening modeling for disk-shaped particles in aluminum alloys: size distribution considered, Materialia 4 (2018) 431–443.

[11] C.L. Liu, Q. Du, N.C. Parson, W.J. Poole, The interaction between Mn and Fe on the precipitation of Mn/Fe dispersoids in Al-Mg-Si-Mn-Fe alloys, Scripta Mater. 152 (2018) 59–63.

[12] H. Strandlund, High-speed thermodynamic calculations for kinetic simulations, Comput. Mater. Sci. 29 (2) (2004) 187–194.

[13] W. Cao, S.-L. Chen, F. Zhang, K. Wu, Y. Yang, Y. Chang, R. Schmid-Fetzer, W. Oates, PANDAT software with PanEngine, PanOptimizer and PanPrecipitation for multi-component phase diagram calculation and materials property simulation, Calphad 33 (2) (2009) 328–342.

[14] B. Sundman, U.R. Kattner, M. Palumbo, S.G. Fries, OpenCalphad - a free thermodynamic software, Integr. Mater. Manuf. Innov. 4 (1) (2015) 1–15.

[15] H. Chen, T. Barman, Thermo-Calc and DICTRA modelling of the β-phase depletion behaviour in CoNiCrAlY coating alloys at different Al contents, Comput, Mater. Sci. 147 (2018) 103–114.

[16] P. Das, P. Dutta, Phase field modelling of microstructure evolution and ripening driven grain growth during cooling slope processing of A356 Al alloy, Comput. Mater. Sci. 125 (2016) 8–19.

[17] X. Doré, H. Combeau, M. Rappaz, Modelling of microsegregation in ternary alloys: application to the solidification of Al–Mg–Si, Acta Mater. 48 (15) (2000) 3951–3962.

[18] Q. Du, D. Eskin, L. Katgerman, An efficient technique for describing a multi-component open system solidification path, Calphad 32 (3) (2008) 478–484.

[19] S.B. Pope, Computationally efficient implementation of combustion chemistry using in situ adaptive tabulation, Combust. Theor. Model. 1 (1) (1997) 41–63.

[20] M.-F. Zhu, W. Cao, S.-L. Chen, C.-P. Hong, Y.A. Chang, Modeling of microstructure and microsegregation in solidification of multi-component alloys, J. Phase Equilibria Diffus. 28 (1) (2007) 130–138.

[21] B. Böttger, R. Altenfeld, G. Laschet, G.J. Schmitz, B. Stöhr, B. Burbaum, An ICME process chain for diffusion brazing of alloy 247, Integr. Mater. Manuf. Innov. 7 (2) (2018) 70–85.

[22] B. Böttger, J. Eiken, M. Apel, Multi-ternary extrapolation scheme for efficient coupling of thermodynamic data to a multi-phase-field model, Comput. Mater. Sci. 108 (2015) 283–292.

[23] Q. Du, Y. Li, An extension of the Kampmann–Wagner numerical model towards as-cast grain size prediction of multicomponent aluminum alloys, Acta Mater. 71 (2014) 380–389, 0.

[24] T. Lu, C.K. Law, Toward accommodating realistic fuel chemistry in large-scale computations, Prog. Energy Combust. Sci. 35 (2) (2009) 192–215.

[25] W. Tian, T.A. Sevilla, D. Li, W. Zuo, M. Wetter, Fast and self-learning indoor airflow simulation based on in situ adaptive tabulation, J. Build. Perform. Simu. 11 (1) (2018) 99–112.

[26] J.O. Dada, P. Mendes, ManyCell: A Multiscale Simulator for Cellular Systems, International Conference on Computational Methods in Systems Biology, Springer, Berlin, Heidelberg, 2012, pp. 366–369.

[27] Q. Du, M. Perez, W.J. Poole, M. Wells, Numerical integration of the Gibbs–Thomson equation for multicomponent systems, Scripta Mater. 66 (7) (2012) 419–422.

[28] J. Hill, G. Mulholland, K. Persson, R. Seshadri, C. Wolverton, B. Meredig, Materials science with large-scale data and informatics: unlocking new opportunities, MRS Bull. 41 (5) (2016) 399–409.

[29] G. Bonaccorso, Machine Learning Algorithms, Packt Publishing Ltd, Birmingham, 2017.

[30] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, Tensorflow: a System for Large-Scale Machine Learning, 12th Symposium on Operating Systems Design and Implementation, 2016, pp. 265–283.

[31] J.D. Hunter, Matplotlib: a 2D graphics environment, Comput. Sci. Eng. 9 (3) (2007) 90–95.

[32] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

[33] StandardScalar, in: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[34] Q. Du, L. Jia, K. Tang, B. Holmedal, Modelling and Experimental Validation of Microstructure Evolution during the Cooling Stage of Homogenization Heat Treatment of Al-Mg-Si Alloys, Materialia, 2018.

[35] Q. Du, W. Poole, M. Wells, N. Parson, Microstructure evolution during homogenization of Al–Mn–Fe–Si alloys: modeling and experimental results, Acta Mater. 61 (13) (2013) 4961–4973.

[36] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep Learning, MIT press, Cambridge2016.

[37] Q. Du, W. Poole, M. Wells, A mathematical model coupled to CALPHAD to predict precipitation kinetics for multicomponent aluminum alloys, Acta Mater. 60 (9) (2012) 3830–3839.

[38] Y.-J. Chang, A.-C. Yeh, The formation of cellular precipitate and its effect on the tensile properties of a precipitation strengthened high entropy alloy, Mater. Chem. Phys. 210 (2018) 111–119.

[39] Y. Li, Z.R. Zhang, Z.Y. Zhao, H.X. Li, L. Katgerman, J.S. Zhang, L.Z. Zhuang, Effect of main elements (Zn, Mg, and Cu) on hot tearing susceptibility during direct-chill casting of 7xxx aluminum alloys, Metall. Mater. Trans. 50 (8) (2019) 3603–3616.

[40] K. Tang, Q. Du, Y. Li, Modelling microstructure evolution during casting, homogenization and ageing heat treatment of Al-Mg-Si-Cu-Fe-Mn alloys, Calphad 63 (2018) 164–184.

[41] D. Zill, W.S. Wright, M.R. Cullen, Advanced Engineering Mathematics, Jones & Bartlett Learning2011.

[42] S. Mei, A. Montanari, P.-M. Nguyen, A mean field view of the landscape of two-layer neural networks, Proc. Natl. Acad. Sci. U.S.A. 115 (33) (2018) E7665–E7671.