Computational Materials Science 188 (2021) 110133

Contents lists available at ScienceDirect

Computational Materials Science

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

Utilizing Gaussian processes to fit high dimension thermodynamic data that includes estimated variability

Richard Couperthwaite [a,*], Douglas Allaire [b], Raymundo Arróyave [a]

[a] Materials Science and Engineering Department, Texas A&M University, College Station, Texas, United States
[b] Department of Mechanical Engineering, Texas A&M University, College Station, Texas, United States

ARTICLE INFO

Keywords:
Thermodynamics
Thermo-Calc
Dual phase steel
Gaussian process

ABSTRACT

CALPHAD-based thermodynamic modeling is an integral component of any ICME framework applied to the accelerated development of alloys. The utility of this type of analysis is that it provides knowledge about the impact of chemistry and (to some degree) processing on the phase stability of alloys. This information can later be passed on to other computational tools which can be used to narrow the experimental space that needs to be explored to arrive at optimal alloy designs. Two major challenges arise with these techniques: (1) it is difficult to interface the outputs of such models with other computational tools without significant overhead; (2) CALPHAD-based predictions tend to be agnostic with regards to uncertainty. The latter challenge is because in commercial thermodynamic packages, it is often not possible to access the model parameters as they tend to be encrypted, making the associated thermodynamic databases essentially 'black boxes' and so methods that consider only the inputs to the models must be considered. In the current work, we develop surrogate models of CALPHAD-based phase stability predictions that fulfill two objectives: (1) they enable the offline evaluation of a component of the ICME model chain that can then be incorporated into a more complete alloy design scheme without the need to directly interface with a thermodynamic engine; (2) they allow for the consideration of uncertainty. We apply the framework to the investigation of the impact of chemistry and heat treatment on the phase constitution of commercial steel grades and evaluate the performance of this framework relative to direct thermodynamic calculations.

1. Introduction

There are many ways to approach materials design within the Integrated Computational Materials Engineering (ICME) framework. In these ICME frameworks, one of the approaches is to formulate process-structure–property (PSP) relationships that can then be inverted to discover regions in the alloy-processing space with optimal performance. One of the key components of these PSP relationships is the process-to-structure relationships. Formulation of the process-structure relations in any alloy system is usually achieved in one of two ways.

Experimental methods can determine the phase fraction and composition of an alloy. Various metallography techniques can determine the phase fractions [1–4]. And by using spectrometry methods it is also possible to measure phase compositions. This approach provides the most accurate measure of a material's microstructure. But, it is costly, in both material and time costs. Thermodynamic models can predict the equilibrium state of the material [5]. However, thermodynamic models are significantly less accurate unless they have been properly assessed and validated against experiments. The advantage of using thermodynamic models is that they come with a significant cost benefit compared to experimental methods.

Due to the inherent heterogeneity of materials, it is helpful to be able to account for uncertainty in the thermodynamic models. There is a fairly large body of work in the literature that considers parametric uncertainty in Thermodynamic modeling with work by Olbricht [6], Otis [7], Honarmandi [8] and Duong [9] being good examples of a fully Bayesian approach to quantifying parametric uncertainty in CALPHAD models.

All these approaches utilize Markov-chain Monte-Carlo (MCMC) sampling of the model parameters of the CALPHAD models used in thermodynamic simulations [8,6,7,10,11,9]. However, these methods rely on access to the thermodynamic models and good quality thermodynamic databases. In contrast, it is far more common to have access to commercially available thermodynamic software such as Thermo-

* Corresponding author.
E-mail address: richardcouperthwaite@tamu.edu (R. Couperthwaite).

https://doi.org/10.1016/j.commatsci.2020.110133
Received 24 July 2020; Received in revised form 16 October 2020; Accepted 17 October 2020
Available online 20 November 2020

Calc™ [12]. Utilizing such software tools comes with the drawback that access to the models and parameters is restricted, and so it is not possible to use the MCMC methods mentioned above to calculate the uncertainty in the model outputs. Therefore the current work proposes a more conventional surrogate modeling approach that is still capable of accounting for some sources of uncertainty.

The current work aims to develop a framework to transform a thermodynamic-based simulation framework connecting chemistry and phase constitution into a surrogate modeling scheme that can predict the volume fraction and phase compositions of a two-phase material. The inputs to this model are the composition of the material as well as the temperature of a single-stage intercritical annealing treatment.

A key aspect of the proposed work is the development of a process-structure model capable of offline evaluation (i.e. without the need to explicitly call a thermodynamic engine) that predicts the mean response accurately and also provides a measure of quantified uncertainty. This kind of model and the approach used in the current work will be generally applicable to any ICME approach requiring thermodynamic models. A further aspect of the current work is to insure that the models are computationally cheap to evaluate. The motivation for this is that design optimization frameworks typically require many function evaluations, and a computationally cheap model will add less overhead to the optimization framework.

Steel alloys are the focus of the current work since these materials are still of significant interest in many industries [3,13]. Steel production entails significant variability, weighing differences between batches, spillage, evaporation, and inaccurate temperature measurements are a selection of the many process parameters that can introduce variability. In most experimental or production processes it is very difficult if not impossible to account for these parameters fully and so we would classify this uncertainty as residual variability. While the authors acknowledge that this variability does involve some uncertainty that could be reduced by better processing methods, we assumed that significant process optimization has already been accomplished and any further reduction in uncertainty would be minimal.

To reiterate, the current work aims to achieve two goals. The first goal is to generate a surrogate for the thermodynamic model that ensures the models are cheap to query and accurately reproduce the thermodynamic model. In this way, the surrogate model becomes an offline model that allows the thermodynamic response to be used in an ICME approach without explicit calls to the thermodynamic engine. The second goal is to propose a method of propagating uncertainty through the surrogate using parametric variability of the input parameters. The distributions from which to sample the input parameters are defined by the controllable composition of elements in production-grade steel.

2. Methods

In the current work, we utilize the Thermo-Calc™ [12] model as a simulator model of a real heat-treatment process. This simulator model provides information on the volume fraction and phase compositions of steel materials using a CALPHAD based approach. The current work aims to build a statistically based emulator, or surrogate, model using Gaussian Processes (GPs) that can accurately replicate the Thermo-Calc™ and be used to probe the parametric variability of the model.

As a result, the current work is divided into two stages. The first stage is the generation of a surrogate model based on data obtained from Thermo-Calc™ [12]. As already indicated, the current work uses GP models for the surrogate models. The second stage is the propagation of parametric variability through the surrogate model. In this second part, the composition of production-grade steel alloys informs the distribution shape for the parametric variability.

2.1. Thermodynamic assessment with Thermo-Calc™

Thermo-Calc™ [12] utilizes the CALPHAD method to determine equilibrium phase fractions and compositions in multi-component systems. The current work uses the TCFE7 iron database for the thermodynamic data, and the Matlab interface for Thermo-Calc™ was used to complete the computations.

Table 1
DP980 and DF140T Nominal Composition and the Composition of the design space in the current work.

	C (wt.%)	Mn (wt.%)	Si (wt.%)	Fe (wt.%)	Temperature (°C)
DP980	0.09	2.15	0.60	bal.	–
DF140T	0.15	1.45	0.30	bal.	–
Model Input Bounds	0.0–1.0	0.0–3.0	0.0–2.0	bal.	650–850

The focus of the current work is dual-phase (martensite-ferrite) steel alloys containing C, Si, and Mn (Fe in balance). We assumene that these alloys to have been subjected to a single-stage intercritical annealing heat treatment followed by quenching. As such, the input space for the models in the current work is the composition (wt%) of the C, Si, and Mn and the temperature for the intercritical annealing heat treatment (T_{IA}).

For the composition, we selected two common dual-phase steels to guide the limits of the region of interest. These alloys are DP-980 and DF-140T and Table 1 shows the composition of both alloys. It is necessary to define upper and lower bounds to the composition of the elements in the alloy to avoid the computational space from becoming too large. Therefore, we chose bounds that encompassed both alloys and ensured that the results would also apply to a larger range of alloys to allow for possible comparison with results in the literature. For the intercritical annealing temperature, we chose a range such that it was possible to produce material with 100% ferrite and 100% austenite within the input space. Table 1 shows the chosen bounds for the composition and intercritical annealing temperature.

Thermo-Calc™ calculates equilibrium phases, and so it is not possible to obtain the martensite fraction from Thermo-Calc™[1] Therefore, we used the Koistinen–Marburger relation shown in Eq. (1) [14] to determine the fraction of austenite converted to martensite under different quenching temperature conditions. The Koistinen–Marburger relation requires the martensite start temperature (T_{ms}) and the quench temperature (T_Q). The martensite start temperature was calculated using the formula presented by Andrews [15] shown in Eq. (2). The quench temperature was assumed to be 25 °C.

$$V_f^{mart} = 1 - e^{(-0.011(T_{ms}-T_Q))}. \tag{1}$$

$$T_{ms}(K) = 812 - 423X_C - 30.4X_{Mn} - 0.075X_{Si}. \tag{2}$$

One of the key assumptions at this point is that due to the fast cooling during quenching there is insufficient time for diffusion to occur and so the composition of the martensite phase is the same as the high-temperature austenite phase.

To get the data for the construction of the emulator model, several different samplings of the design space were used. Firstly, uniform sampling with either 6, 7, 8, or 9 samples per dimension was used. This produced four data-sets of 1296, 2401, 4096, and 6561 samples. Secondly, Latin hypercube sampling of the space was used to generate the sampling points. Latin hypercube sampling is typically used when the sampling region is very large or has many dimensions. The approach subdivides each input dimension into the number of points required and then randomly combines these input values ensuring that each value on each dimension is used exactly once [16]. In this case, the number of samples obtained matched the data-set sizes of the uniform sampling.

[1] Thermo-Calc™ versions from 2019a do include the possibility for calculating the martensite volume fractions, however, this method can only be used with the TCFE9, or later, database and also does not provide the phase composition and so was not utilized in the current work.

The motivation for using both approaches was to compare which approach is capable of producing the most accurate surrogate model with the smallest number of training points. Since reducing the number of training points will greatly reduce the amount of time that it takes to invert the GP kernel matrix a smaller training sample size will speed up queries to the surrogate model.

There are 7 outputs of the Thermo-Calc™ model. The first is the volume fraction of the martensite phase (calculated from the Koistinen–Marburger relation). The next six outputs are the elemental weight fractions of Si, Mn, and C in both the martensite and ferrite phase. Since there is an assumption of no losses during heat treatment, the elemental composition of the two phases must obey a mass balance. As such, it won't be necessary to model the full composition of both phases. However, the composition of both phases was considered to determine which would create a better surrogate model.

2.2. Source of uncertainty

Uncertainty in both modeling and experimental work has multiple sources. For the current work, the sources of uncertainty in computer models will be discussed in detail. Some of the sources of uncertainty in experimental work have already been mentioned. The following is a summary of the sources of computational uncertainty identified in the work by Kennedy and O'Hagan [17], the interested reader is referred to their work for a more complete discussion.

Parameter Uncertainty. This is the uncertainty associated with not knowing the true value of the parameters of the model. The assumption behind this uncertainty is that there is a single true value for the parameter.

Parametric Variability. Parametric variability, in contrast to parametric uncertainty, is when the parameter in question does not have a unique value, but rather has a distribution of possible values.

Model Inadequacy. Model inadequacy is the discrepancy between the output of the model and typically the mean of a real-world result. This assumes that the model is being utilized with the correct values for all parameters.

Residual Variability. Residual variability encompasses two sources of uncertainty that are difficult if not possible to differentiate between. The first is that there could be missing inputs that if fully specified would reduce this variability, and the second is that the real world process itself might be stochastic.

Observation Error. Observation error applies to a real-world process where there is often an error associated with the measurement of the output.

Code Uncertainty. Despite it being theoretically possible to predict the outcome of a mathematical model, the complexity of many models and the requirements of running the model for hours or even days means that it is not possible to know the exact outcome from the model. This is classified as code or interpolation uncertainty.

The approach described by Kennedy and O'Hagan [17] is the basis for the approach used in the current work, however, for the current work the model is not calibrated against experimental results. The approach can be expanded to include experimental results for future work. Since experimental results are not considered in the current work, the sources of uncertainty that will be considered are observation error and code uncertainty. The uncertainty in the output of the surrogate model will be introduced by considering parametric variability in the inputs to the surrogate model.

The authors acknowledge that residual variability in the production of steels is very distinct from the parametric uncertainty or variability in the emulator, or surrogate, model. However, the current work proposes that the residual variability can provide information to inform the parametric variability of the surrogate model. How this is to be achieved is to use the compositional variation of production-grade steels to define the distributions for the input parameters to the surrogate model.

2.3. Gaussian process fitting of thermodynamic results

The current work aims to define a surrogate model or emulator of the Thermo-Calc™ model. This can be defined as the determination of a function f such that $f : \chi \longrightarrow \mathscr{Y}$. In this case, $Y(\boldsymbol{x}) \in \mathscr{Y} \subseteq \mathbb{R}$ is the univariate output of the Thermo-Calc™ model at a given input $\boldsymbol{x} \in \chi \subseteq \mathbb{R}^q$, where χ is the q-dimensional domain of interest or design space. The measurement, or observation, error ($\epsilon_{obs}(\boldsymbol{x})$) is defined as the uncertainty in the measurement of $Y(\boldsymbol{x})$, however, since the Thermo-Calc™ model is a deterministic model there is no error associated with the result. As a result, this term will be replaced by a uniform noise variance in the current work to ensure computational stability. This is discussed in further detail later.

$$Y(\boldsymbol{x}) = f(\boldsymbol{x}) + \epsilon_{obs}(\boldsymbol{x}). \tag{3}$$

The current work uses a GP for the emulator model. GPs have become one of the most widely used statistical models [18] since they provide the ability to analyze and quantify uncertainty in functions, provide excellent flexibility through the different covariance functions that can be employed as well as having attractive statistical properties.

A GP is a non-parametric statistical model that defines a stochastic process $f(\boldsymbol{x})$ such that all the finite distributions of the model are assumed to be multi-variate normal. As a result of this, the joint probability distribution of the outputs from the stochastic process for any finite set of inputs $\mathbb{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ may be modeled as an n-dimensional multivariate normal distribution:

$$p(f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)) \sim \mathscr{N}_n(\boldsymbol{\mu}, \boldsymbol{C}). \tag{4}$$

where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{C}$ is the covariance function. These are defined by a mean function $\mu(\cdot)$ and a covariance function $C(\cdot,\cdot)$ with the following properties:

$$\mu(\boldsymbol{x}_i) = \boldsymbol{\mu}_i = \mathbb{E}[f(\boldsymbol{x}_i)]. \tag{5}$$

$$C(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{C}_{i,j} = cov\left[f(\boldsymbol{x}_i), f(\boldsymbol{x}_j)\right]. \tag{6}$$

Considering this context, we will define a GP as $f(\cdot) \sim \mathscr{GP}(\mu, C)$. A more detailed explanation of this kind of stochastic process is provided in the work by Rasmussen and Williams [19].

The covariance function captures the spatial dependence between two different input locations, and along with the mean function plays a role in the final probability distribution of the outputs of the stochastic process. The probability distribution of the surrogate model outputs defines the interpolation uncertainty, or to use the terminology defined by Kennedy and O'Hagan, code uncertainty [17].

This approach for defining the GP model allows for the development of the model definition in Eq. (3) to include two sources of uncertainty. This approach defines $f(\boldsymbol{x})$ as the mean response of the GP, $\epsilon_{obs}(\boldsymbol{x})$ as the observational error, and $\epsilon_{code}(\boldsymbol{x})$ as the interpolation error from the GP. Where the distribution of the code error is defined by $\epsilon_{code}(\boldsymbol{x}) \sim \mathscr{N}(0, C)$.

$$Y(\boldsymbol{x}) = f(\boldsymbol{x}) + \epsilon_{obs}(\boldsymbol{x}) + \epsilon_{code}(\boldsymbol{x}). \tag{7}$$

The observational error ($\epsilon_{obs}(\boldsymbol{x})$) can be handled in two ways. The first is to measure the error of each observation explicitly, while the second is to use the observational error as an additional parameter in the building of the GP. This second approach is referred to as using a nugget [20].

There are many different possible covariance functions available for use with GPs. Rasmussen and Williams [19] provide definitions of many of the more commonly used covariance functions. Two of these covariance functions are utilized in the current work. The squared exponential and Matérn($\nu = 5/2$) covariance functions.

The squared exponential function calculates the covariance of the input space as a weighted euclidean distance between the input variables and can be parameterized according to Eq. (8), where n is the number of dimensions, σ_f^2 is the signal variance, and l is the

characteristic length scale or smoothness parameter.

$$C(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 exp\left(-\frac{1}{2}\sum_{h=1}^{n}\left[\frac{(x_{i,h} - x_{j,h})}{l_h}\right]^2 \right). \quad (8)$$

The squared exponential function was implemented in MATLAB using code based on the approach developed by Ghoreishi et al. [21].

The Matérn class of covariance functions is defined in a *single dimension* by Eq. (9), where K_v is a modified Bessel function [19]. However, it is more common to define the function by specifying a specific value for v. One of the more commonly used values is $v = 5/2$ [19]. This choice reduces the covariance function to that shown in Eq. (10). This equation shows a generic multi-dimension representation of the covariance function with $v = 5/2$, where n is the number of dimensions, and l_h is the characteristic length scale of dimension h.

$$C(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \left[\frac{2^{1-v}}{\Gamma(v)} \left(\frac{\sqrt{2v}\,(x_i - x_j)}{l} \right)^v K_v \left(\frac{\sqrt{2v}\,(x_i - x_j)}{l} \right) \right]. \quad (9)$$

$$C(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \sum_{h=1}^{n} \left(1 + \frac{\sqrt{5}\,(x_i - x_j)}{l_h} + \frac{5(x_i - x_j)^2}{3l_h^2} \right) exp\left(-\frac{\sqrt{5}\,(x_i - x_j)}{l_h} \right). \quad (10)$$

The implementation of the Matérn covariance function was done within Python using the "George.py" module [22].

Since the input parameters have different units, all the inputs were scaled to the interval $[0, 1]$. For mathematical convenience, the outputs from Thermo-Calc™ were standardized to have a mean of zero and a variance of 1. This approach allows us to specify the mean function as $\mu(\boldsymbol{x}) = \boldsymbol{0}$. The observation error term defined in Eq. (3) is added to the covariance function when calculating the output of the GP model. However, if the results do not contain a measurement error $\epsilon_{obs}(\cdot)$, as in the case of the output from a deterministic model, a small value, often referred to as a nugget [20], can be added in place of observation error to provide numerical stability in the calculation of the matrices and their inverses. Doing this assumes that the errors are all identically and independently distributed with a normal distribution of zero mean and σ_n^2 variance, $\epsilon(\boldsymbol{x}) \sim \mathcal{N}(0, \sigma_n^2)$. The variance of the errors (σ_n^2) is also referred to as the noise variance [19].

There are two standard approaches to optimizing the values of the hyper-parameter for GPs. The first is to use gradient-based approaches [19]. The second is to use Bayesian approaches such as Markov-chain Monte-Carlo methods [17]. It was chosen to use the gradient-based approach in the current work since the gradient-based approach is typically less computationally expensive.

While the gradient-based approach is usually less computationally costly than the Bayesian approach it has been noted that there is a possibility for there to be multiple local optima in the parameter space. Rasmussen and Williams [19] discuss this briefly and indicate that this is more likely to occur when there is less data available since there will be more combinations of the parameters that can provide a sufficient fit to the data.

Due to the chance of local optima in the hyper-parameter space, the optimization used a multi-start approach in an attempt to avoid having the optimization process get stuck in local optima. Since all the input values were scaled to the interval $[0, 1]$ the initial guesses for the length scale hyperparameters were selected from that interval. As discussed, to aid the inversion of the matrices a nugget term with $\sigma_n^2 = 0.05$ was included.

For the current work, it is important that the surrogate model is an accurate representation of the Thermo-Calc™ data, therefore, the results from the GP were validated against 10,000 data points calculated by uniform sampling with 10 samples for each dimension. The coefficient of determination (Eq. 12) was used as the measure of fit.

In addition to measuring the coefficient of determination, the results were plotted against the Thermo-Calc™ results for the two alloys of interest in the current work. For this comparison, the composition of the alloys was fixed and the temperature varied over the design range. Since there is no training data that directly corresponds to the Thermo-Calc™ data for these two alloys, this method helps provide visual confirmation of how well the model is predicting general values of the design space.

2.4. Uncertainty propagation

As discussed earlier, the compositional variation of production-grade steels informs the parametric variability. Since the published steel grades show the possible variation in the elemental composition it is possible to define a maximum and minimum value for any given input.

Two approaches were used in the current work, however, it is noted that these were chosen for convenience rather than being definitive methods for approaching this kind of problem. The first method was to assume that the input distribution was a uniform distribution between the maximum and minimum values for the input. This is considered a non-informative approach. The second method was to assume that the input was normally distributed and that the maximum and minimum values define a distance of two standard deviations away from the mean. This approach does make a strong assumption about the input distribution being normal.

Using these two distributions the parametric variability of the model is assessed by sampling from the input distributions and then calculating the mean and variance of the mean output from the GP. This is one of the simpler methods for obtaining the parametric variability since it doesn't take into account the code uncertainty of the GP. This also simplifies the definition of the parametric variability to be a multivariate normal distribution with mean 0 and variance σ ($\epsilon_{par} \sim \mathcal{N}(0, \sigma)$) and the model definition given in Eq. (3) and developed in Eq. (7) can be further developed to include the uncertainty due to parametric variability.

$$Y(\boldsymbol{x}) = f(\boldsymbol{x}) + \epsilon_{obs}(\boldsymbol{x}) + \epsilon_{code}(\boldsymbol{x}) + \epsilon_{par}(\boldsymbol{x}). \quad (11)$$

3. Results

3.1. Building of the surrogate model

The first tests involved building the GP model with the squared exponential kernel. The optimization of the GP hyperparameters was conducted with training points from both the uniform and Latin hypercube sampling. For each sampling method, the best performing hyperparameters were recorded for the seven output values from Thermo-Calc. The hyperparameters of interest in the current work are the characteristic length scale (l) and the signal variance (σ_f), since the noise variance, or nugget (σ_n), was kept constant.

To measure the accuracy of the GP model, the coefficient of determination was used. The Coefficient of determination uses the ratio between the residual sum of squares (Eq. 13) and the total sum of squares (Eq. 14), where y_i is a Thermo-Calc™ output, $\bar{y}$ is the mean of all Thermo-Calc™ outputs and f_i is the surrogate model value corresponding to the inputs of Thermo-Calc™ output y_i. The residual sum of squares is the square of the distance between the predicted values and the true values, while the total sum of squares is the distance between the true values and the mean of the true values. The coefficient of determination is defined for the interval $[0, 1]$ with a value of 1 indicating a perfect fit, and a value of 0 indicating that the prediction is no better than the mean. Any values outside of this range indicate that the predicted values from the model are further from the true values than the mean.

$$CoD = 1 - \frac{SS_{res}}{SS_{tot}}. \quad (12)$$

$$SS_{res} = \sum_i (y_i - f_i)^2. \quad (13)$$

Table 2
Coefficient of determination for the 10,000 test data points for each of the sample sets used in both uniform and LHS sampling.

	Uniform Data				LHS Data			
	1296	2401	4096	6561	1296	2401	4096	6561
V_f^{mart}	0.99	0.99	1.00	1.00	0.48	0.48	0.95	1.00
X_C^{mart}	0.98	0.99	0.99	0.99	0.57	0.42	1.00	0.97
X_{Si}^{mart}	0.99	1.00	1.00	1.00	0.57	0.53	0.41	1.00
X_{Mn}^{mart}	0.99	1.00	1.00	1.00	0.72	0.54	1.00	0.99
X_C^{ferr}	0.97	0.98	0.99	0.99	0.66	0.54	0.27	1.00
X_{Si}^{ferr}	0.99	1.00	1.00	1.00	0.46	0.31	1.00	1.00
X_{Mn}^{ferr}	0.98	0.99	0.99	1.00	0.70	0.62	1.00	1.00

$$SS_{tot} = \sum_i (y_i - \bar{y})^2. \tag{14}$$

The accuracy of each of the GPs using the optimum length scale and noise variance hyperparameter results are shown in Table 2. As can be seen, the accuracy of the GP increases with increasing sample size. This effect is more noticeable with the Latin hypercube training data than the uniform training data. What is interesting to note is that the uniform training data has a much higher accuracy against the test data than the same number of training points generated from a Latin hypercube sampling. As already noted, Latin hypercube sampling is typically used when only a very sparse sampling of the input space is possible. Therefore, for a 4-dimensional problem such as is considered in the current work, the sample size of 1000 to 7000 is sufficiently large for the uniform sampling to perform better. Choosing smaller sample sizes or increasing the dimensionality of the problem will almost certainly result in the Latin hypercube sampling performing better than the uniform sampling.

The scatter plots in Fig. 1 show how as the sample size increases the fit improves. However, this representation also provides a further observation. The fit deviates most significantly for higher temperature samples. As such, if the input space was reduced to lower temperatures, it might be possible to achieve the same accuracy with fewer training samples.

Using the data for the two alloys DP980 and DF140T, the trained GPs were used to predict the outputs from the Thermo-Calc model to assist in providing a visual representation of how well the GPs are performing. When comparing the GP output with that of Thermo-Calc for the two alloys specifically, the performance is not as good as the performance on the test points, particularly when it concerns the composition of the

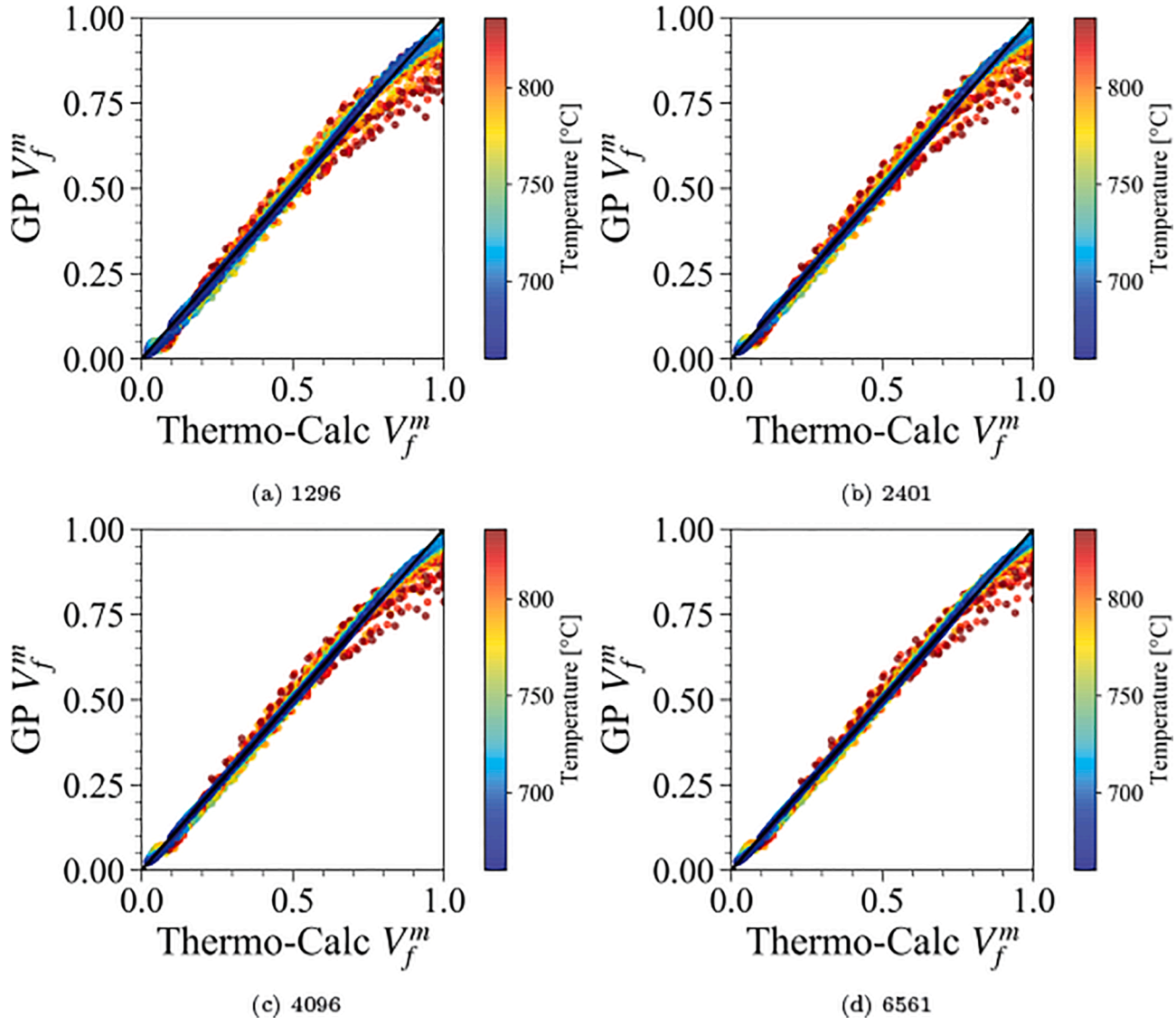

Fig. 1. Scatter plot of the 10,000 volume fraction results from Thermo-Calc™ with the results from the GP with a squared exponential covariance function.

Table 3
Coefficient of determination for the DP980 data set when using the samples as specified with the Squared Exponential covariance function.

	Uniform Data				LHS Data			
	1296	2401	4096	6561	1296	2401	4096	6561
V_f^{mart}	0.99	0.99	0.98	0.98	0.26	−0.22	0.92	0.97
X_C^{mart}	0.76	0.70	0.66	0.64	−1.25	−5.17	0.68	0.66
X_{Si}^{mart}	−159.84	−124.60	−100.29	−92.65	0.15	0.02	−2.21	−97.55
X_{Mn}^{mart}	−0.20	−0.47	−0.57	−0.64	−3.37	−7.40	−0.67	−0.47
X_C^{ferr}	0.07	−0.61	−1.30	−1.74	−11.61	−10.95	−4.18	−1.21
X_{Si}^{ferr}	−284.15	−195.05	−146.80	−130.84	0.00	−1.19	−112.38	−116.29
X_{Mn}^{ferr}	−51.08	−60.73	−61.58	−63.01	−31.91	−93.48	−53.73	−64.72

Table 4
Results from the Matérn Kernel fit using only the sample set with 6561 samples and uniform sampling.

	DF140T	DP980
V_f^{mart}	0.991	0.985
X_C^{mart}	0.999	0.998
X_{Si}^{mart}	0.952	0.899
X_{Mn}^{mart}	0.997	0.995
X_C^{ferr}	0.981	0.987
X_{Si}^{ferr}	0.79	0.719
X_{Mn}^{ferr}	0.91	0.869

phases in the alloy. Table 3 shows the results for the DP980 alloy. The DF140T alloy results showed a similar trend.

The results of the fit to the composition of the phases in the two test alloys is slightly concerning and also surprising. The fit to the test points for the composition results is almost perfect, while the fit to the composition values of the test alloys is very bad. As a test, the number of training points was increased to 10,000 and the results still showed the same problem.

The next step was to implement the GP model with a Matérn (5/2) covariance function. This approach resulted in a significantly better fit for the composition of the phases, Table 4. This indicates that the Matérn (5/2) is a better covariance function to use in the current work. Despite the better fit, Fig. 2 shows that the interpolation error of the composition values is still significant. The error has been truncated at zero since a negative composition has no physical meaning.

The results in Table 4 also show that the predictions for the martensite phase composition are better than for the ferrite phase. As such, it would be best to use the GPs to predict the martensite composition and then calculate the ferrite composition using the mass balance of the elements.

Since one of the aims of the current work is to have a fast surrogate model, the time taken to calculate the 10,000 samples was measured for each of the GPs constructed. This measurement was done by repeating the calculation of the test set 20 times and averaging the results. These results are shown in Table 5 in the appendix. These show that even for the largest training sample size the time taken to calculate 10,000 data points is reasonably small. Considering this, and the increased accuracy that using the larger training set provides, it was decided that the largest training set would be used in the subsequent analyses.

3.2. Parameter variability

Using the two methods described the parametric variability was added to the results from the surrogate model. As discussed in the methods, the approach in the current work was to predict the mean response at each sampling from the composition and temperature input space and then find the average and variance of this mean output. This is used to define the normally distributed parametric error. Following Eq. (11) this parametric error was added to the code or interpolation error and the 95% confidence interval was calculated. These results for the two sampling procedures are shown in Fig. 3.

As can be seen the parametric error approximately doubles the 95% confidence interval of the data. However, the uniform sampling results in a smaller confidence interval. The most likely reason for this is that using the interval between the bounds results in a smaller sampling region around the mean. The normal distribution allows values outside of the region in the uniform sampling approach.

Table 5
Time taken to build the GP and query 10,000 data points simultaneously.

Training Sample Size	Time to Test 10,000 points (s) MATLAB	Python
1296	5.23 ± 0.05	0.70 ± 0.02
2401	6.93 ± 0.11	1.53 ± 0.05
4096	10.70 ± 0.09	3.63 ± 0.05
6561	16.84 ± 0.24	7.83 ± 0.08

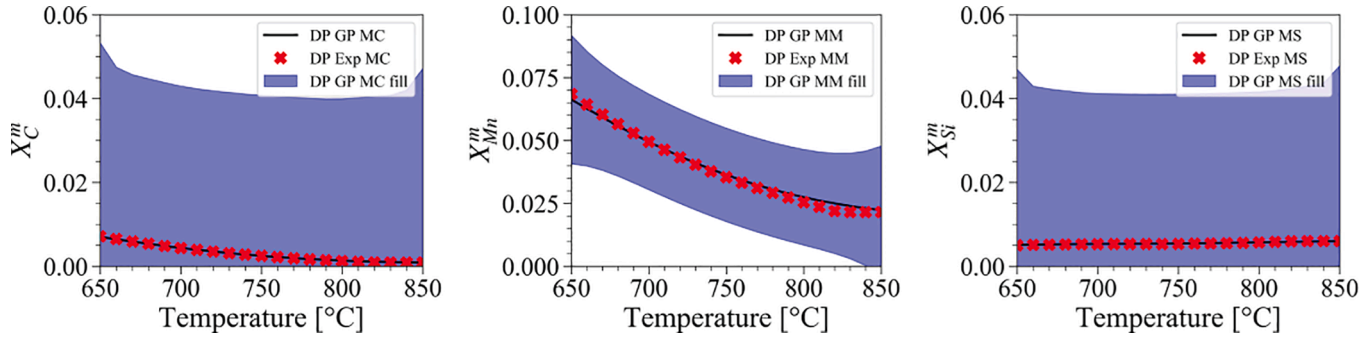

Fig. 2. Comparison of the GP outputs and the Thermo-Calc™ results for the elemental compositions of the martensite phase for the DF140T alloy. These show the very large confidence interval around the GP prediction for the phase composition.

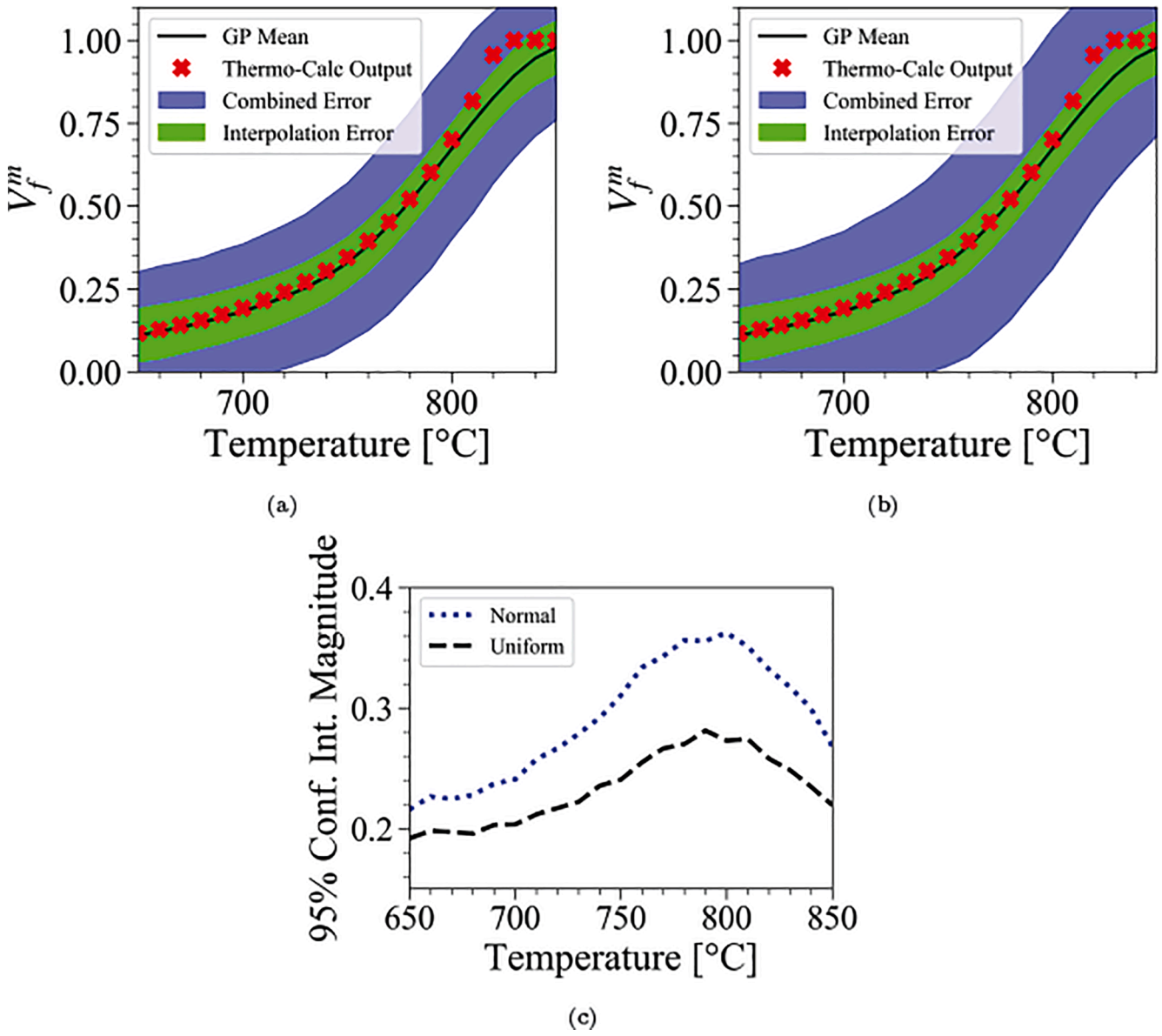

Fig. 3. Surrogate model outputs for the DP980 alloy showing 95% confidence intervals defined by interpolation error only (green), and the combination of interpolation error and parametric variability (blue) for samplings of the parametric error from (a) Uniform and (b) Normal Distributions with (c) showing a comparison of the magnitude of one side of the 95% confidence interval for the two sampling approaches.

4. Discussion

During the current work, namely fitting a GP surrogate model to Thermo-Calc™ results several noteworthy results were obtained. Firstly, it was found that using a squared exponential function for the GP provided a good fit to the volume fraction output of the code, but failed when fitting the composition of the phases. The exact reason for this is not known, however, this result does show that testing of multiple covariance functions is necessary to ensure the best fit for the results.

The second noteworthy finding was that Latin hypercube sampling produces GPs with a worse fit to the data when compared to the same number of samples from a uniform sampling procedure. This would possibly be a result of the Thermo-Calc™ output being relatively smooth over most of the domain, however, more testing would be required to confirm this. The second reason for the poor performance of the Latin hypercube approach is that the design space is still small enough to be effectively sampled by uniform sampling. Therefore, when expanding this work to larger design spaces with more dimensions, the Latin hypercube sampling will become a better approach since it won't be computationally possible to consider uniform sampling.

The procedure followed in the current work developed a set of GP surrogate models that were able to separately account for three sources of uncertainty in the modeling process. These sources were observation error, code uncertainty, and parametric variability. While in the current work, the observation error is neglected since the Thermo-Calc™ result is a deterministic result, this approach would be able to account for this error. This would be done by including the observation error as the noise variance (σ_n) for each of the observations.

The code uncertainty in the models of the current work is reasonable for the prediction of the volume fraction, however, it was observed that the code uncertainty for the elemental composition of the phases was significantly larger. This could potentially be decreased by using a larger sample, however, the time cost of the larger training set would need to be tested to determine the optimal size that can lower the interpolation uncertainty while not increasing the computational time to unreasonable levels.

Parametric variability was added to the surrogate model by using two distributions, uniform and normal, defined by the residual variability of production-grade steel. This approach was found to produce reasonable uncertainty to the results. Both distributions of the input parameters approximately doubled the size of the 95% confidence interval. However, the uniform distribution had a smaller 95% confidence interval. As a result, using the normal distribution will result in a more conservative estimate of the error.

5. Conclusions

The current method developed a set of GP surrogate models that can be easily integrated with ICME materials design approaches. These models provide basic information on the microstructure of a material following a simple heat-treatment process. Further, these models can be evaluated quickly, which means that they will not increase the computational time of an optimization approach significantly.

Using the residual uncertainty in the composition of production-grade steel materials, the distributions for the calculation of parametric uncertainty were defined. This provides the opportunity for propagating this uncertainty through structure-to-property models in the PSP chains used in ICME approaches.

While the surrogate model developed in this work can be easily integrated into any existing ICME approach, the intention is to integrate this model into the multi-information source fusion method presented by Ghoreishi et al. [21]. This is to expand the work on the multi-information source fusion approach to include materials composition and processing parameters since this was identified as an area for development in this approach.

As a final note, while these results are a useful addition to the modeling of material microstructures, the volume fraction of phases and phase composition are only some of the variables that are needed to fully determine the mechanical properties of a material. One of the other significant parameters that are necessary is the grain size of the phases. Therefore, while the current work has proven that it is possible to obtain good results from a GP fit to Thermo-Calc™ data, in addition to the expansion of the thermodynamic modeling explained earlier, it is necessary to expand the current work to include calculations of the grain size of the material.

6. Data Availability

On reasonable request, the data used for this work can be obtained from the corresponding author.

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.

Acknowledgements

The authors acknowledge Grant No. NSF-CMMI-1663130, *DEMS: Multi-Information Source Value of Information-Based Design of Multiphase Structural Materials*. RA also acknowledges Grant No. NSF-DGE-1545403.

Appendix A

A.1. Code timing comparison

The timing of the code was done by building and training the GP model and then querying 10,000 data points from the model. This measurement was repeated 20 times and the results were averaged. These results are shown in Table 5 in the appendix. These show that even for the largest training sample size the time taken to calculate 10,000 data points is reasonably small (17s for the MATLAB implementation and 8s for the Python Implementation. This shows that for this particular application, it is not necessary to sacrifice the accuracy of the largest dataset for a faster GP model since the times taken by the model built from the largest dataset are small enough to not significantly impact an optimization approach.

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.commatsci.2020.110133.

References

[1] S. Gündüz, Effect of chemical composition, martensite volume fraction and tempering on tensile behaviour of dual phase steels, Materials Letters 63 (27) (2009) 2381–2383, https://doi.org/10.1016/j.matlet.2009.08.015. URL:http://www.sciencedirect.com/science/article/pii/S0167577X09006132.

[2] V.L. de la Concepción, H.N. Lorusso, H.G. Svoboda, Effect of carbon content on microstructure and mechanical properties of dual phase steels, Procedia Materials Science 8 (2015) 1047–1056, https://doi.org/10.1016/j.mspro.2015.04.167.

[3] H. Ashrafi, M. Shamanian, R. Emadi, N. Saeidi, A novel and simple technique for development of dual phase steels with excellent ductility, Materials Science and Engineering: A 680 (2017) 197–202, https://doi.org/10.1016/j.msea.2016.10.098. URL:http://www.sciencedirect.com/science/article/pii/S0921509316313259.

[4] J. Sun, T. Jiang, Y. Sun, Y. Wang, Y. Liu, A lamellar structured ultrafine grain ferrite-martensite dual-phase steel and its resistance to hydrogen embrittlement, Journal of Alloys and Compounds 698 (2017) 390–399, https://doi.org/10.1016/j.jallcom.2016.12.224. URL:http://www.sciencedirect.com/science/article/pii/S0925838816341536.

[5] H. Bhadeshia, Computational design of advanced steels, Scripta Materialia 70 (2014) 12–17, https://doi.org/10.1016/j.scriptamat.2013.06.005. URL:http://www.sciencedirect.com/science/article/pii/S1359646213003072.

[6] W. Olbricht, N.D. Chatterjee, K. Miller, Bayes estimation: A novel approach to derivation of internally consistent thermodynamic data for minerals, their uncertainties, and correlations, Part I: Theory, Physics and Chemistry of Minerals 21 (1) (1994) 36–49, https://doi.org/10.1007/BF00205214. URL:https://doi.org/10.1007/BF00205214.

[7] R.A. Otis, Z.-K. Liu, High-throughput thermodynamic modeling and uncertainty quantification for ICME, JOM 69 (5) (2017) 886–892, https://doi.org/10.1007/s11837-017-2318-6. URL:https://doi.org/10.1007/s11837-017-2318-6.

[8] P. Honarmandi, R. Arroyave, Using Bayesian framework to calibrate a physically based model describing strain-stress behavior of TRIP steels, Computational Materials Science 129 (Supplement C) (2017) 66–81, https://doi.org/10.1016/j.commatsci.2016.12.015. URL:http://www.sciencedirect.com/science/article/pii/S0927025616306292.

[9] T.C. Duong, R.E. Hackenberg, A. Landa, P. Honarmandi, A. Talapatra, H.M. Volz, A. Llobet, A.I. Smith, G. King, S. Bajaj, A. Ruban, L. Vitos, P.E. Turchi, R. Arróyave, Revisiting thermodynamics and kinetic diffusivities of uranium-niobium with Bayesian uncertainty analysis, Calphad 55 (2016) 219–230, https://doi.org/10.1016/j.calphad.2016.09.006. URL:http://www.sciencedirect.com/science/article/pii/S036459161630061X.

[10] N.D. Chatterjee, R. Krüger, G. Haller, W. Olbricht, The Bayesian approach to an internally consistent thermodynamic database: theory, database, and generation of phase diagrams, Contributions to Mineralogy and Petrology 133 (1) (1998) 149–168, https://doi.org/10.1007/s004100050444. URL:https://doi.org/10.1007/s004100050444.

[11] M. Stan, B. Reardon, A Bayesian approach to evaluating the uncertainty of thermodynamic data and phase diagrams, Calphad 27 (3) (2003) 319–323, https://doi.org/10.1016/j.calphad.2003.11.002. URL:http://www.sciencedirect.com/science/article/pii/S0364591603000804.

[12] J.-O. Andersson, T. Helander, L. Höglund, P. Shi, B. Sundman, Thermo-Calc & DICTRA, computational tools for materials science, Calphad 26 (2) (2002) 273–312, https://doi.org/10.1016/S0364-5916(02)00037-8. URL:http://www.sciencedirect.com/science/article/pii/S0364591602000378.

[13] H. Bhadeshia, S.R. Honeycombe, Steels (third ed.), Butterworth-Heinemann, Oxford, 2006. doi:10.1016/B978-075068084-4/50003-0. URL:http://www.sciencedirect.com/science/article/pii/B9780750680844500030.

[14] D. Koistinen, R. Marburger, A general equation prescribing the extent of the austenite-martensite transformation in pure iron-carbon alloys and plain carbon steels, Acta Metallurgica 7 (1) (1959) 59–60, https://doi.org/10.1016/0001-6160(59)90170-1. URL:http://www.sciencedirect.com/science/article/pii/0001616059901701.

[15] K. Andrews, Empirical formulae for the calculation of some transformation temperatures, Journal of the Iron and Steel Institute 203 (1965) 721–727.

[16] M.D. McKay, R.J. Beckman, W.J. Conover, A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics 21 (2) (1979) 239–245, publisher: [Taylor & Francis Ltd, American

Statistical Association, American Society for Quality]. doi:10.2307/1268522. URL: http://www.jstor.org/stable/1268522.
[17] M.C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (3) (2001) 425–464, https://doi.org/10.1111/1467-9868.00294. URL:https://doi.org/10.1111/1467-9868.00294.
[18] A. O'Hagan, Polynomial chaos: A tutorial and critique from a statistician ' s perspective, 2013.
[19] C.E. Rasmussen, C.K. Williams, Gaussian Processes for Machine Learning, The MIT Press, 2006.
[20] I. Andrianakis, P.G. Challenor, The effect of the nugget on Gaussian process emulators of computer models, Computational Statistics & Data Analysis 56 (12) (2012) 4215–4228, https://doi.org/10.1016/j.csda.2012.04.020. URL:http://www.sciencedirect.com/science/article/pii/S0167947312001879.
[21] S.F. Ghoreishi, A. Molkeri, A. Srivastava, R. Arroyave, D. Allaire, Multi-Information Source Fusion and Optimization to Realize ICME: Application to Dual-Phase Materials, Journal of Mechanical Design 140 (11) (2018) 111409–111409–14. https://doi.org/10.1115/1.4041034.
[22] S. Ambikasaran, D. Foreman-Mackey, L. Greengard, D.W. Hogg, M. O'Neil, Fast Direct Methods for Gaussian Processes. URL:http://arxiv.org/abs/1403.6015.