Berg Huettenmaenn Monatsh (2022) Vol. 167 (1): 29–35
<https://doi.org/10.1007/s00501-021-01182-3>
 © The Author(s) 2021

BHM Berg- und
 Hüttenmännische
 Monatshefte

Applications of Data Driven Methods in Computational Materials Design

Christoph Dösinger¹, Tobias Spitaler¹, Alexander Reichmann¹, Daniel Scheiber², and Lorenz Romaner¹

¹Department Werkstoffwissenschaft, Lehrstuhl für Metallkunde und metallische Werkstoffe, Montanuniversität Leoben, Leoben, Austria

²Materials Center Leoben Forschung GmbH, Leoben, Austria

Received November 21, 2021; accepted November 23, 2021; published online December 21, 2021

Abstract: In today's digitized world, large amounts of data are becoming available at rates never seen before. This holds true also for materials science where high-throughput simulations and experiments continuously produce new data. Data driven methods are required which can make best use of the information stored in large data repositories. In the present article, two of such data driven methods are presented. First, we apply machine learning to generalize and extend the results obtained from computationally intense density functional theory (DFT) simulations. We show how grain boundary segregation energies can be trained with gradient boosting regression and extended to many more positions in the grain boundary for a complete description. The second method relies on Bayesian inference, which can be used to calibrate models to give data and quantification of the model uncertainty. The method is applied to calibrate parameters in thermodynamic models of the Gibbs energy of Ti-W alloys. The uncertainty of the model parameters is quantified and propagated to the phase boundaries of the Ti-W system.

Keywords: Data driven materials science, Machine learning, Bayesian inference, Uncertainty quantification and propagation, Grain boundary segregation, CALPHAD modelling

Anwendungen von datengetriebenen Methoden in computergestütztem Materialdesign

Zusammenfassung: Die zunehmenden Datenmengen, die in der modernen Materialwissenschaft mittels Simulationen und Experimenten generiert werden, erfordern die An-

wendung und Entwicklung von datengetriebenen Methoden, um die enkodierten Informationen optimal zu nutzen. In diesem Artikel werden zwei solcher Methoden vorgestellt. Zunächst wenden wir maschinelles Lernen an, um die Ergebnisse aus rechenintensiven Dichtefunktionaltheoriesimulationen zu verallgemeinern und zu erweitern. Wir zeigen, wie die Segregationsenergien von metallischen Korngrenzen mit der sogenannten Gradient-Boosting-Regression trainiert und auf viele weitere Positionen ausgedehnt werden können, um eine vollständige Beschreibung zu erhalten. Die zweite Methode beruht auf der Bayes'schen Inferenz, die zur Kalibrierung von Modellen verwendet werden kann, um Daten und eine Quantifizierung von Modellunsicherheiten zu erhalten. Die Methode wird hier zur Kalibrierung von Parametern in thermodynamischen Modellen der Gibbs-Energie von Ti-W-Legierungen angewendet. Es wird gezeigt, wie die Unsicherheit der Modellparameter quantifiziert werden kann und wie sich diese auf die Vorhersage der Phasengrenzen des Ti-W-Systems überträgt.

Schlüsselwörter: Datengetriebene Materialwissenschaft, Maschinelles Lernen, Bayes'sche Inferenz, Quantifizierung und Fortpflanzung von Unsicherheiten, Korngrenzensegregation, CALPHAD-Modellierung

1. Introduction

Data-driven science is becoming more and more an integral part of modern material development. Through improvements in experimental techniques and computational power, increasing amounts of data are being generated and more data is available in research facilities as well as in global data repositories [1]. To be able to process and analyse this data, new tools are required which allow detecting patterns and anomalies as well as extracting new relationships between materials quantities which could not be discovered otherwise. Computational materials science is

L. Romaner (✉)
 Department Werkstoffwissenschaft, Lehrstuhl für Metallkunde
 und metallische Werkstoffe,
 Montanuniversität Leoben,
 Roseggerstraße 12,
 8700 Leoben, Austria
lorenz.romaner@unileoben.ac.at

adopting novel data-driven approaches to speed up calculations and improve its predictive power.

In this article we will present two applications of data-driven methods on the basis of two specific examples. The first example describes how computationally expensive calculations based on density functional theory (DFT) can be replaced with regression methods to obtain a more complete description of segregation phenomena in metallic materials. While DFT allows for scanning the periodic table for favourable or unfavourable solute elements at grain boundaries, the description is still idealized and does not fully account for the high degree of complexity which can be encountered at grain boundaries. For this reason, it is required to train machine learning methods to a consistent set of data encoding the DFT results and extend from there to all atomic sites in the grain boundary.

The second example demonstrates how uncertainty quantification can be applied to calibrating physical models to complex materials phenomena and processes. So far, parameter calibration has been mostly uncertainty-agnostic and computational predictions have been rarely provided with an error bar. However, approaches based on Bayesian inference provide a convenient framework for uncertainty quantification, which is crucial in the verification of simulation results [2]. We will demonstrate how such a methodology can be applied to calibrate a thermodynamic model to experimental data of binary W-Ti alloys and to determine their degree of uncertainty. In this way, the regions in the phase diagram can be identified where further experimental or theoretical data are required for a consistent improvement of the thermodynamic model.

2. Methodology of Data Driven Methods

2.1 Machine Learning of Segregation Energies

In general, the term machine learning (ML) covers different methods including regression, classification or clustering. Common to all these methods is their application on an existing data set which should be large enough to encode the relevant information. To make these data usable, they are represented by a set of features or descriptors which are either taken directly from the data sets or obtained by pre-processing of the data. ML methods are generally divided into supervised and unsupervised learning algorithms. In supervised learning the ML algorithm is handed, in addition to the features, also the target quantity which is then used to train the model. In contrast, for unsupervised learning, no target values are provided. Supervised ML methods cover a variety of methods ranging from the simplest, i.e. linear regression, to more sophisticated methods such as decision tree-based methods or artificial (deep) neural networks. In this work, linear regression (LR) and a particular version of a tree-based ensemble method, gradient boosting regression (GBR), are adopted as implemented in the free Python library *scikit-learn* [3]. A comparison of the performance of the two regression techniques gives valuable insights into the quality of the employed features.

To carry out LR, a function

$$f(\vec{x}; \vec{w}) = \vec{x} \vec{w} + w_0 \quad (1)$$

is fitted to data points $(\vec{x}_i, y_i)$ by varying the coefficients $w_0, \vec{w}$ and minimizing the residual sum of squares (RSS)

$$\min_{\vec{w}} \sum_i (\vec{x}_i \vec{w} + w_0 - y_i)^2. \quad (2)$$

The $\vec{x}_i$ are the features of each data point and the y_i are the corresponding target quantities, in the present case the segregation energies. The expression $\vec{x}_i \vec{w}$ denotes the scalar product between two vectors.

The advantage of a LR is that it can be evaluated quickly, even for large datasets. As the name says, it can only describe linear relations. Therefore, if a feature performs well with LR, it is particularly suitable for ML of the investigated data.

GBR is one of the so-called ensemble methods where several models are trained on a dataset which then give reliable results. For the gradient boosting regression, several models $h(\vec{x}; a)$ are summed together for the model $F(\vec{x})$ [4]:

$$F_M(\vec{x}) = \sum_{m=0}^M \beta_m h(\vec{x}; a_m). \quad (3)$$

The expansion coefficients β_m and the parameters a_m are iteratively fitted to the residual $\tilde{y}_m$ of the previous iterations:

$$\tilde{y}_m = \tilde{y}_{m-1} - F_{m-1}(\vec{x}); \tilde{y}_0 = y. \quad (4)$$

This process is repeated until either the maximum number M is reached or a sufficient accuracy has been reached. For the models $h(\vec{x}; a)$ in Eq. 3, *scikit-learn* uses decision trees calculated with a modified CART algorithm. Decision trees consist of a sequence of nodes that partitions the dataset in the feature space. This leads in the end to fitting a series of step-functions to the data which can provide highly non-linear relationships. Compared to LR, GBR is therefore more flexible in the functional form of the target function.

2.2 Uncertainty Estimation in the Bayesian Framework

In materials science, physical models are often fit to experimental and theoretical data by variation of the parameters of the model to minimize the error in the property to be predicted. Usually, a single value of the model parameters is given as the result without specifying any bounds of uncertainty. But quantification and management of the uncertainty are crucial in materials science to validate the model. Uncertainty in the model and its parameters arise as the data used to fit the model come with an error bar because the model might not be able to represent all the data

due to simplifying assumptions or because the knowledge about the system is incomplete.

Both frequentist and Bayesian inference can be used for parameter calibration and uncertainty estimation of model parameters. In the frequentist framework, the measurement is a realization of a random variable, but the model parameters are assumed to be fixed. In the Bayesian framework, the model parameters are treated as a random variable with a probability distribution. Furthermore, the Bayesian framework allows to incorporate existing knowledge about the model parameters (prior knowledge) in the inference process.

The posterior distribution $p(\theta|M,D)$ defines the probability of the parameter θ of the model M given the existing data D . The prior probability of a parameter $p(\theta|M)$ is updated to the posterior probability, given the data D using Bayes theorem [5]:

$$p(\theta|M,D) = \frac{p(\theta|M) p(D|M,\theta)}{p(D|M)}. \quad (5)$$

The prior contains already established knowledge or beliefs on the parameter values of a model M . $p(D|M)$ is called the evidence, and it is the probability to obtain the data D with the model M . $p(D|M,\theta)$, also called likelihood, is the probability that the data D is described by the model M with the parameters θ . A possibility to define the likelihood is to assume a normal distribution of the measurement around its true value with a standard deviation σ corresponding to the measurement error. The residual R_i is the deviation between data point D_i and the model $M(\theta)$. The likelihood for a single data point D_i is then calculated as:

$$p(D_i|M,\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{R_i^2}{2\sigma^2}\right). \quad (6)$$

The total likelihood is the product over the likelihoods of the single data points: $p(D|M,\theta) = \prod_i p(D_i|M,\theta)$. The closed form solution for the posterior probability in Eq. 5 is rarely possible. Markov Chain Monte Carlo (MCMC) can be used to sample the posterior, which allows for simultaneous determination of the optimal parameter set and uncertainty quantification. Starting with an initial set of parameters θ , a new parameter set θ' is proposed at each iteration. The new parameter set is accepted according to

the Metropolis Criterion, where the ratio of the posterior probabilities comes into account

$$p_{\text{accept}} = \min\left(\frac{p(\theta'|M,D)}{p(\theta|M,D)}, 1\right) = \min\left(\frac{p(\theta'|M) p(D|M,\theta')}{p(\theta|M) p(D|M,\theta)}, 1\right). \quad (7)$$

The Markov Chain converges to the optimal parameter set. After the convergence of the Markov Chain, the parameters will fluctuate around the optimum and, in this way, sample the posterior distribution [6]. The probability distribution corresponds to the uncertainty of the parameters which can be propagated to the model output. This is done by sampling the parameter space and evaluating the model with the sampled parameters to arrive at the distribution and uncertainty of the model output. As the Markov Chain already samples the parameter space after convergence, these samples can be used for this purpose.

3. Applications

3.1 Machine Learning of Grain Boundary Segregation

This section demonstrates how GPR can be used to predict grain boundary segregation energies based on local atomic environments. Grain boundary (GB) segregation describes the process of trapping of solute atoms to interfaces between different grains in a metal. GB segregation and the resulting GB excess can strongly influence different materials properties, for example intergranular brittleness or phase transformation. GB excess is related to the segregation energy (E_k^{seg}) via the McLean isotherm [7]

$$\frac{c_k}{1-c_k} = \frac{c_0}{1-c_0} \cdot e^{-\frac{E_k^{\text{seg}}}{k_b T}}, \quad (8)$$

which describes the concentration of the solutes at a GB (c_k) depending on bulk concentrations (c_0), the temperature (T). The E_k^{seg} can be provided by atomistic modelling. A routinely used and very accurate method is *density functional theory* (DFT) as it is implemented, for example, in the Vienna Ab-initio Simulation Package (VASP). Based on the

Fig. 1: Illustration of selected grain boundaries in a W-25 at.% Re alloy. Circles illustrate different atomic sites in the grain boundary and are colour-coded according to the segregation energies

Fig. 2: Predicted versus DFT-calculated segregation energies when applying GBR to cartesian coordinates (*left*), spherical coordinates (*center*) and the Steinhardt q_5 parameter (*right*)

atomic structure of a GB, the E_k^{seg} are calculated as the difference in energy between placing the solute atom at a GB site ($E_{GB,k}$) and placing it at a reference bulk site ($E_{GB,0}$) [8]:

$$E_k^{\text{seg}} = E_{GB,k} - E_{GB,0}. \quad (9)$$

This method has been used previously to calculate segregation energies for a W-25at.% Re alloy [8]. The dataset contains in total 219 segregation energies for segregation sites in 15 different GBs. Fig. 1 shows some of these GBs where the values of E_k^{seg} are indicated by different colors.

Since each calculation of a E_k^{seg} entails a DFT simulation for up to several hundreds of atoms, the associated computational effort is considerable. A way to accelerate the calculation of the E_k^{seg} is to apply ML. Based on the local atomic structure around a segregation site, the corresponding E^{seg} can be predicted. To show how the choice of different features influences the result, different descriptions of the local atomic environment are chosen including the cartesian coordinates or the spherical coordinates of the neighbouring atoms. Only the 15 atoms closest to the site are considered and the features are sorted according to the radial distance. In addition we also consider Steinhardt bond order parameters which are evaluated from the spherical harmonics Y_{lm} as [9]:

$$q_l = \left[\frac{4\pi}{2l+1} \sum_{m=-l}^l \left| \frac{1}{N_A} \sum_{n=1}^{N_A} Y_{lm}(\hat{r}_n) \right|^2 \right]^{\frac{1}{2}}. \quad (10)$$

Here the $\hat{r}_n$ are the unit vectors in direction of the neighbouring atoms and N_A is the number of atoms used in the sum. As already pointed out previously, there exists a remarkable correlation of E^{seg} and the Steinhardt q_5 parameter for the W-Re system [8]. The performance of different ML results is measured via the coefficient of determination, i.e. the R^2 -score. A score of one signifies a perfect fit, while a score of zero means that the quality of the fit is as good as the mean value of the data.

Table 1 and Fig. 2 compare the results of different ML methods for three different sets of features as mentioned before. The simplest and most straightforward feature set are the Cartesian coordinates. Applying LR clearly fails, since the corresponding R^2 score is negative indicating that

TABLE 1
Performance of ML approaches employing different features and regression methods. The scores for the models are calculated using 5-fold cross-validation

Feature types	Number of features	R^2 score, LR	R^2 score, GBR
Cartesian coordinates	45	-0.43	0.10
Spherical coordinates	45	0.70	0.80
Steinhardt q_5	1	0.80	0.78

the Cartesian coordinates are not suitable features. GBR improves the results somewhat, however the R^2 is still far from 1. These results for the Cartesian coordinates motivate the search for alternative features for the problem of GB segregation. Changing from Cartesian coordinates to spherical coordinates already greatly improves the results for LR and GBR. Since segregation energy shows a linear dependence on the Steinhardt q_5 parameter, LR produces good results which are not improved any more with GBR. In general, the results highlight that proper feature engineering is decisive to improve regression quality.

In Fig. 3 we show the benefits of training a ML model to the segregation data. The total effort of the DFT calculations, (red, empty circles in the plot) was about 1 million coreh. With the trained ML algorithm, the segregation profile can now be completed (black points) at essentially zero computational cost. Although the ML predictions do not exactly reproduce the data everywhere, the trends are very well captured.

Our results show that ML-models using spherical coordinates or the Steinhardt q_5 parameter already give a reasonable description of the GB segregation. Certainly, more elaborate features engineering can still improve this result. Furthermore, a general application of this methodology also requires extending the method to also include chemistry results. Such investigations are left to future research efforts.

Fig. 3: Profiles of the E^{seg} depending on the distance from the GBs. Red open circles mark the DFT values while black solid circles show the ML predictions with GBR using spherical coordinates as the features

3.2 Uncertainty Quantification in CALPHAD Modeling

The CALPHAD (CALculation of PHase Diagrams) method is an important part of materials science and thermodynamic calculations [10]. Starting from a thermodynamic description of the phases in a material, phase diagrams and thermodynamic properties can be estimated and also used as input in subsequent models such as diffusion simulations or phase field modeling of phase transformation.

Every experimental and theoretical data point which can be numerically related to the Gibbs energy can be used as input data for the CALPHAD assessment if a residual function can be defined. Thermochemical data, like heat capacity, enthalpy, and entropy, is used as these quantities can be calculated directly from derivatives of the Gibbs energy. Experimental evaluations of invariant points, transition temperatures, liquidus and solidus temperatures, and phase stabilities are crucial input for the determination of phase diagrams. The definition and evaluation of the residual function for this kind of data is more challenging as it requires a global minimization of the Gibbs energy.

CALPHAD assessments have mostly been deterministic, giving a single value for the parameters, without specifying any uncertainty of the parameters. Although there have been some early works, the topic of uncertainty quantification and propagation in CALPHAD has only gained momentum in recent years, often adopting the Bayesian framework [11]. As the accuracy of the thermodynamic database determines the reliability of the calculated results, the determination of the uncertainty in CALPHAD helps to understand

how strongly one can rely on results from thermodynamic simulations. The determination of uncertainty also allows to identify gaps of knowledge in the data and can lead to the identification of the experiment which will contribute the most to the accuracy of the thermodynamic database.

The open source program ESPEI [12] uses the Bayesian framework for the parameter calibration and uncertainty quantification as described in Sect. 2.2. We demonstrate the parameter calibration and uncertainty quantification in a CALPHAD optimization with the example of the binary titanium tungsten (Ti-W) system. The accepted phase diagram of Ti-W stems from an assessment in 1996 [13] based on experimental data dating back to the 50s. Using these data we performed a parameter optimization with ESPEI. The Ti-W binary system has three stable phases, the body-centered-cubic (BCC), the hexagonal (HCP) and the liquid phase. The experimental phase diagram data includes the determination of the solidus, liquidus, determination of the eutectic temperature and composition, and (rough) determination of the bcc-miscibility gap at higher temperatures.

In total 6 parameters are calibrated against the data, describing the excess terms and, thus, the mixing energy of the two components in the BCC, HCP, and liquid phase. The corner plot in Fig. 4a shows the posterior distribution of the parameters obtained from MCMC. The diagonal plots are histograms of the parameter values from the Markov Chain after convergence, resembling the posterior. The covariance of two parameters is shown by the off-diagonal images. We use parameter set samples from this parameter distribution (which we can take from the Markov Chain) to identify the uncertainty in the phase boundaries. For this

Fig. 4: Schematic illustration of the process of uncertainty propagation of CALPHAD parameters to phase boundary prediction. More information is provided in the text

we calculate, for each parameter set, the phase diagram, represented by the minimization of the Gibbs energy with the common tangent construction in Fig. 4b. For a qualitative estimation of the uncertainty of the phase boundaries, we plot the calculated phase diagrams on top of each other. When the phase boundaries lie on top of each other, as it is the case for the solidus up to 50 atomic percent tungsten, the uncertainty in the phase boundary is low. With a wider spread of the phase boundaries, there is a larger uncertainty, as we can see for the location of the miscibility gap and the solubility of titanium in BCC tungsten.

4. Conclusions

Data driven materials modeling offers powerful tools that can be applied to a variety of materials design questions. We have presented two specific applications. The first deals with machine learning of DFT-calculated segregation energies which are key to predict grain boundary chemical composition. We could demonstrate how ML can effectively complete the total segregation profile of GBs in W and save a large amount of computational resources. Further development of such methodology is foreseen to allow investigating advanced co-segregation phenomena and phase transformations at grain boundaries. The second application was about quantification and propagation of uncertainty in thermodynamic modelling of the W-Ti alloy based on Bayesian inference and Markov chain Monte Carlo sampling. The currently accepted W-Ti phase diagram was reassessed and completed with uncertainty quantification and propagation. The analysis shows that especially the

miscibility gap in the phase diagram is still associated with considerable uncertainty and additional experimental data should be supplied for higher precision of the thermodynamic model.

Acknowledgements. The authors gratefully acknowledge the financial support under the scope of the COMET program within the K2 Center “Integrated Computational Material, Process and Product Engineering (IC-MPPE)” (Project No 859480). This program is supported by the Austrian Federal Ministries for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK) and for Digital and Economic Affairs (BMDW), represented by the Austrian research funding association (FFG), and the federal states of Styria, Upper Austria, and Tyrol. This research was funded in part by the Austrian Science Fund (FWF) [P 34179-N].

Funding. Open access funding provided by Montanuniversität Leoben.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Draxl, C., Scheffler, M.: Big data-driven materials science and its FAIR data infrastructure. Handbook of materials modeling. Springer, Cham, pp 49–73 (2020)
2. Honarmandi, P., Arróyave, R.: Uncertainty quantification and propagation in computational materials science and simulation-assisted

- materials design. *Integr Mater Manuf Innov* **9**(1), 103–143 (2020). <https://doi.org/10.1007/S40192-020-00168-2>
3. Pedregosa, F., et al.: Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
 4. Friedman, J.H.: Stochastic gradient boosting. *Comput Stat Data Anal* **38**(4), 367–378 (2002). [https://doi.org/10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2)
 5. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian data analysis. Chapman and Hall/CRC, (1995)
 6. Paulson, N.H., Bocklund, B.J., Otis, R.A., Liu, Z.K., Stan, M.: Quantified uncertainty in thermodynamic modeling for materials design. *Acta Mater* **174**, 9–15 (2019). <https://doi.org/10.1016/j.actamat.2019.05.017>
 7. Sutton, A.P., Balluffi, R.W.: Interfaces in crystalline materials. Clarendon Press, Oxford (1995)
 8. Scheiber, D., Razumovskiy, V.I., Puschnig, P., Pippan, R., Romaner, L.: Ab initio description of segregation and cohesion of grain boundaries in W–25 at.% Re alloys. *Acta Mater* **88**, 180–189 (2015). <https://doi.org/10.1016/j.actamat.2014.12.053>
 9. Bartók, A.P., Kondor, R., Csányi, G.: On representing chemical environments. *Phys. Rev., B. Condens. Matter.* **87**(18), 184115 (2013). <https://doi.org/10.1103/PhysRevB.87.184115>
 10. Kattner, U.R.: The calphad method and its role in material and process development. *Tecnol Metal Mater Min* **13**(1), 3–15 (2016). <https://doi.org/10.4322/2176-1523.1059>
 11. Honarmandi, P., Paulson, N.H., Arróyave, R., Stan, M.: Uncertainty quantification and propagation in CALPHAD modeling. *Model Simul Mat Sci Eng* (2019). <https://doi.org/10.1088/1361-651X/ab08c3>
 12. Bocklund, B., Otis, R., Egorov, A., Obaied, A., Roslyakova, I., Liu, Z.-K.: ESPEI for efficient thermodynamic database development, modification, and uncertainty quantification: application to Cu-Mg. *MRS Commun* **9**(02), 1–10 (2019). <https://doi.org/10.1557/mrc.2019.59>
 13. Jonsson, S.: Reevaluation of the Ti-W system and prediction of the ti-W-N phase diagram. *Int J Mater Res* **87**(10), 784–787 (1996). <https://doi.org/10.1515/IJMR-1996-871008>

Publisher's Note. Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.