

Comparison of statistically-based methods for automated weighting of experimental data in CALPHAD-type assessment

Noah Harris Paulson^{1,*}, Setareh Zomorodpoosh², Irina Roslyakova^{2,*}, Marius Stan¹

1 – Argonne National Laboratory, 9700 S. Cass Avenue, Lemont, IL 60439, USA

2 – Interdisciplinary Centre for Advanced Materials Simulation (ICAMS), Ruhr-University Bochum, Germany

Abstract

The selection and weighting of experimental and simulated datasets is a necessary step in the development of thermodynamic property models in the calculation of phase diagrams (CALPHAD) approach. Currently, this requires painstaking and complicated evaluation of the reliability of datasets and consistency between them. In this work, we present two novel and independently developed statistical approaches to aid in this process by addressing outliers and performing automated dataset weighting. The first method, presented here for the first time, applies classical statistical techniques and commonly available optimization algorithms. The second method employs Bayesian statistics via numerical sampling techniques. In this work, we present both approaches and compare their strengths and weaknesses through an assessment of the specific heat of aluminum and hafnium metal versus temperature for several experimental datasets. We finally compare the weightings of each dataset versus a number of metrics employed by experts to evaluate the reliability of datasets.

Key words: automated weighting, statistical methods, CALPHAD-type assessment, heat capacity, enthalpy

* Corresponding author. Argonne National Laboratory, 9700 Cass Avenue, Lemont, IL 60439 USA. 630-252-1697

1. Introduction

The fidelity with which CALPHAD models represent the stability and thermodynamic properties of real materials is linked to the quality and range of the data used to calibrate them [1]. Frequently, two or more experimental or simulated datasets might conflict with each other for a property or phenomenon of interest [2]. Using the average trend between contradictory datasets is rarely acceptable, as it does not lie close to any measured or simulated material behavior. It is therefore critical to evaluate the physical reasonableness of all available datasets and somehow balance their influence on the final thermodynamic models [3].

In the traditional CALPHAD approach, practitioners assess datasets based on a variety of criteria including, but not limited to, thermodynamic consistency between related properties, agreement with other datasets, measurement methodology, sample quality, and research group reputation, with the final goal of assigning a weight to each set [3], [4]. Practitioners may even decide to remove entirely datasets from consideration, effectively identifying them as outliers and assigning them zero weight in the analysis. Furthermore, this is an iterative process, as each time the researcher performs the thermodynamic optimization they observe the quality of fit of the thermodynamic property models and phase stability to all available data and have the opportunity to modify the models and assign new weights to the datasets [1].

The current state of the art in the assessment and weighting of datasets in the development of CALPHAD models is complex and relies on the expert understanding the subtleties of the available datasets, the interactions between them, and their alignment to the calculated phase diagram [3], [4]. This is a significant barrier to entry for new CALPHAD users and requires formidable time investment on the part of the practitioner. These difficulties increase in multi-component systems where the number of datasets rises and it becomes more challenging to find thermodynamic consistency (though this difficulty tends not to increase beyond three component systems as quaternary compounds are rare). Although physics-based arguments that incorporate a deep understanding of each measurement or simulation result are indispensable, statistically-based automation techniques have the potential to aid CALPHAD practitioners in assessing data and assigning weighting factors. Specifically, these approaches assign weighting factors based on the consistency (or inconsistency) between datasets, integrating well into the CALPHAD approach.

In the past year alone, at least two research groups have independently developed statistical methods to automatically weight datasets in the assessment of thermodynamic property models – the first couched in classical regression techniques, and the second in fully Bayesian inference [5]. The classical approach, presented for the first time in this work, employs a k-fold cross-validation (KFCV) method [6], modified under the condition that each dataset contains unequal number of observations – which is a typical situation during CALPHAD assessments – to estimate the weights for each dataset. In contrast, the Bayesian methodology leverages a simple modification to the likelihood function definition in Bayes’ Theorem [7], [8] to estimate the distribution of weights for each dataset.

In this work, we briefly introduce both the classical regression-based and Bayesian automated weighting methodologies and compare them from the perspective of theory and implementation. We then present the development of thermodynamic property models for two study cases: the low temperature phases of pure hafnium (Hf) and aluminum (Al), using both approaches to automatically weight measurements from a substantial collection of experimental results. These

examples demonstrate the efficiency of automated weighting and the similarity of the results produced by the two methods. Furthermore, we highlight trends in the automatically obtained weights versus experimental details that are of significant importance to experts in the weighting of datasets in the traditional CALPHAD approach.

2. Methodology

In general, a typical CALPHAD-type assessment consists of a nonlinear regression problem where unknown parameters of a selected model are estimated based on available data. The nonlinear regression problem is solved using a numerical estimation method. The most common techniques to determine the unknown model parameters are nonlinear least-squares and Bayesian estimation methods [9].

2.1 Nonlinear regression model of thermodynamic properties

Consider a nonlinear regression model in the general form as

$$y_i = f(x_i, \vec{\theta}) + \epsilon_i, i = 1, \dots, n, \quad (1)$$

where n is the total number of observations on independent $x_i, i = 1, \dots, n$ and dependent $y_i, i = 1, \dots, n$ variables correspondingly, f is a some nonlinear in parameter regression function, and $\vec{\theta} = (\theta_1, \theta_2, \dots, \theta_p)$ is the vector of p unknown model parameters to be estimated from the available experimental or simulation data. We assume that $\epsilon_i, i = 1, \dots, n$ are independent normal distributed random variables with mean 0 and variance σ .

Usually in a thermodynamic assessment, data is collected from several literature sources, such that the total sample size n can be presented as a sum over a number of observations from k distinct references

$$n = \sum_{j=1}^k n_j, \quad (2)$$

where n_j is the number of observations for reference j . Then, following a classical CALPHAD approach, each dataset will be assigned a weight based on the knowledge and intuition of the expert. Afterwards, using the weights for each dataset, the unknown parameters models are estimated using either classical (also called frequentist) or Bayesian regression methods.

In this work, two alternative methods to determine the weights automatically for datasets involved with thermodynamic assessment are presented and compared below.

2.2 The classical regression-based technique

Following a classical approach, we estimate unknown model parameters from the experimental data using a weighted nonlinear least squares method as

$$\vec{\theta}_{bf} = \min_{\vec{\theta}} \sum_{i=1}^{n_j} \sum_{j=1}^k w_j \left(y_j^i - f(x_j^i, \vec{\theta}) \right)^2 \quad (3)$$

where we aim to find the best fit parameter vector $\vec{\theta}_{bf}$ that minimizes the weighted sum of the squared errors. Each dataset indexed by j has an associated weight w_j , and pairs of experimental inputs and outputs (x_j^i, y_j^i) . Assume that we have no prior knowledge about w_j values and thus set-up $w_j^0 = 1, j = 1, \dots, k$. Our goal is to find values of weights such that the unknown model parameters will be statistically estimated more precisely than with unweighted nonlinear regression (1).

The classical method to solve this problem is based on the k-fold cross-validation (KFCV) method [6]. In general, cross-validation (CV) is a statistical technique of evaluating and comparing the predictive models by partitioning the original data into a training set and a test set to train and validate the model, respectively. In a special case, namely KFCV, the original data points are randomly split into k equal (or nearly equal) subsets. Then, the model is built k times, each time leaving out one of the subsets (test data) and fitting the model with the remaining data (training data). Then, the error of the model is estimated by some goodness-of-fit statistic such as mean squared error (MSE) using the test data. Finally, the average of all of the MSE – known as the CV-score – gives the performance of the model. This algorithm is applied to all candidate models and at the end, the one with the lowest CV-score is regarded as the most accurate and robust model. We propose to assign weights to each dataset by employing the KFCV technique. As far as we know, this work contains the first application of KFCV for the weighting of thermodynamic datasets. Note that in regular KFCV, folds are randomly chosen and have equal sizes. In our case, we specify each fold deliberately to contain the number of observations in each reference dataset. In fact, each dataset is considered as one fold, and the number of points in each dataset are not necessarily equal. First, one fold is selected and the data from that fold are considered as test data. After fitting the model on a set of training data, the residual standard error (*RSE*) is calculated on the test fold. To include the consideration of the number of estimated parameters, the *RSE* is chosen as a goodness-of-fit estimator and is defined as follows

$$RSE_l = \sqrt{\frac{\sum_{i=1}^{n_j} \sum_{j \neq l}^k y_j^i - f(x_j^i, \vec{\theta}_j)^2}{n_j - p - 1}}, l = 1, \dots, k \quad (4)$$

where $\vec{\theta}_j$ is the best fit parameter vector when dataset j is excluded from the training data, p is the number of parameters in $\vec{\theta}_j$, and all other notation is the same as in the preceding equations.

If the number of observations is insufficient, that is $n \cdot k - p - 1 \leq 0$, then it can be inferred that specific test data is not reliable enough. In this case, the maximum RSE among the others will be assigned to this set. The goal is to investigate such weights that will increase the accuracy of the applied model by minimization of its RSE. The following formula can be used to calculate the weight for each fold, in which the dataset with the lowest RSE is deemed to be most accurate and have the greatest weight:

$$w_j = 1 - \frac{RSE_j - RSE_{min}}{RSE_{max}}, j = 1, \dots, k \quad (5)$$

where k is the number of folds, and RSE_{min} and RSE_{max} are the minimum and maximum RSEs seen among all k datasets, respectively.

2.3 The Bayesian-type approach

In the Bayesian approach, the parameters of the thermodynamic property model are calibrated using Bayes' Theorem, which is also employed to weight each dataset optimally. Bayes' Theorem describes the update of a pre-existing understanding of the distribution of a model's parameters (the prior distribution) through the observation of data to obtain an updated parameter distribution (the posterior distribution). This update is performed through the computation of the conditional probability of the observed data given the model and a specific set of model parameters (the likelihood). For a model M parameterized by a parameter vector $\vec{\theta}$ and for the observation of data \vec{D} Bayes' Theorem is given by,

$$\Pr(\vec{\theta}|\vec{D}, M) = \frac{\Pr(\vec{D}|\vec{\theta}, M)\Pr(\vec{\theta}|M)}{\Pr(\vec{D}|M)} \quad (6)$$

where $\Pr(\vec{\theta}|M)$ is the prior, $\Pr(\vec{D}|\vec{\theta}, M)$ is the likelihood, $\Pr(\vec{\theta}|\vec{D}, M)$ is the posterior, and $\Pr(\vec{D}|M)$ is the marginal likelihood, or the conditional probability of the data given a model (marginalized over the entire parameter space).

It is through careful selection of the forms of the prior and likelihood distribution that we can achieve a variety of aims, including automatic dataset weighting. Specifically, we employ the following Gaussian likelihood definition,

$$\Pr(\mathbf{D}|\Theta, \alpha, M) = \prod_i^{n_j} \prod_j^k \mathcal{N}(y_j^i | M(x_j^i, \Theta), \varepsilon_j^i / \alpha^i) \quad (7)$$

where $\mathcal{N}(x|\mu, \sigma)$ is the probability density of point x under a Gaussian distribution with mean μ and standard deviation σ , ε_j^i is the reported standard error of data point (x_j^i, y_j^i) of dataset D_i , and α^i is a hyperparameter employed to reweight the reported errors. The hyperparameter vector α is included in the Bayesian inference and therefore its posterior distribution is obtained alongside the model parameters. A small value of α^i reduces the weight of dataset D_i while a large value has the opposite effect. The Exponential distribution is typically employed as a prior for the hyperparameters as its expectation is one (corresponding to no change in the weight of a dataset). Further details are provided in the preceding work [5].

2.4 Comparison of the automated weighting methods

While the two automated weighting schemes have similar goals, they each have different strengths and weaknesses and may be more or less appropriate depending on the problem at hand. The characteristics of these methods derive from the statistical approaches employed and their specific formulations. Many of these differences are due to the characteristics of frequentist and Bayesian statistical assumptions and techniques. These differences are described in the remainder of this section and are summarized in Table 1 below.

Table 1: A comparison of the frequentist and Bayesian approaches described in this work

	Frequentist Method	Bayesian Method
range of weights	$[0, 1]$	$[0, \infty)$
method of action of weights	on squared errors in cost function	rescaling standard errors in likelihood
computational expense	low	high
computational robustness	low	high
major assumption	infinitely repeated experiment	prior parameter distributions
parameters are	fixed	random variables
data are	random variables	fixed or random variables

The major difference between the frequentist and Bayesian algorithms presented are the ranges of weights and the ways that they affect inference for the model parameters. In the frequentist approach, the weights range between 0 and 1 and directly determine the strength of influence of each observed data point on the cost function through the squared error. This cost function is minimized to obtain the best-fit model parameters, so data points with weights of 0 have no influence, while data points with weights of 1 have standard influence. In contrast, the Bayesian weights range from 0 to infinity, and for each data point rescale the standard error, which is used to describe the expected spread of the observed data around the model prediction in the likelihood function. Practically speaking, a weight of zero means that the model prediction is not constrained by the observed data point, while a weight of infinity means that the model prediction must exactly go through the observed data point. Due to these fundamental differences, we only compare the frequentist and Bayesian weights qualitatively and do not guarantee their interchangeability.

Another significant difference between the methods are their ease of application. The frequentist approach does not require the user to make any decisions beyond the selection of the model form and data. In contrast, in the Bayesian approach the user must pick appropriate prior distributions for the model parameters and weighting factors (as these are included in the set of model parameters). While prior selection can be performed in a traceable and repeatable manner, there is some degree of expertise required to select informative priors. Furthermore, in the Bayesian scheme we require the standard error to rescale in the likelihood definition whereas the frequentist approach does not utilize reported errors. Utilizing the reported errors requires more effort, but may provide better weights in the case when the number of data points per set is small. In this work, we assume $\pm 5\%$ reported standard errors in the Bayesian scheme to compare better with the frequentist weighting scheme. The frequentist approach is not computationally expensive; however, there is a risk of the calculation being trapped in local minima during the minimization

of the cost function. This minimization must be repeated with different starting points to ensure that the global minima in parameter space is found. The Bayesian approach is guaranteed to find the proper distribution of the model parameters given sufficient sampling time (assuming the use of a Monte Carlo sampling approach). However, it requires many more function evaluations than the frequentist algorithm and therefore is more computationally expensive.

One consideration that is not a major focus of this work is the simultaneous fitting of different quantities of interest (e.g. activity, specific heat, and enthalpy). This is a challenging situation in part due to the different magnitudes of measurement scales across properties. In the Bayesian approach, the reported and re-scaled error serves as the standard deviation parameter in the likelihood definition, automatically accounting for different measurement scales [10]. In the Frequentist approach, different data quantities would be rescaled to the range of [0, 1] prior to fitting.

3. Results and Discussion

The application of both automated weighing methods to the heat capacity data of pure Al and Hf described by the segmented regression model [11] is demonstrated and discussed below.

3.1 Heat capacity model

The classical regression-based and Bayesian automated weighting methodologies have been applied here to estimate the unknown model parameters and automatically define values of weights for each reference dataset involved in the assessment of the heat capacity data from 0K for pure Hf and Al. The authors have decided to test the developed methodologies on the segmented regression (SR) model [11]. This model is intended to provide an accurate description of heat capacity data from zero Kelvin up to the melting point by a decomposition of contributed physical effects into low and high temperatures. The model was successfully applied to thermodynamic properties of pure elements [11], [12], a calcium nitrate compound [13], Cr-Nb [14], Cr-Ta [15] and Cu-Mg [16] binary systems. The novel automated weighting methods described here can be applied to any alternate thermodynamic property model, for example, the model proposed by Chen and Sundman in 2001 for the description of thermodynamic properties of pure Fe [17].

According to the SR model, a temperature dependency in the heat capacity of solid stable phases is described as

$$C_p^{SR}(T) = C_V(T, \theta_D) + C_p^{bcm}(T, \beta_1, \beta_2, \alpha, \tau) + C_p^{mag}(T), \quad (8)$$

where T is the temperature in Kelvin and $\theta_D, \beta_1, \beta_2, \alpha, \tau$ are unknown model parameters to be estimated from available experimental and simulation data.

The first term in equation (6) is introduced to take into account phonon vibrations using the well-known Debye model [18],

$$C_V(T, \theta_D) = 9R \left(\frac{T}{\theta_D} \right)^3 \int_0^{\frac{\theta_D}{T}} \frac{x^4 e^{-x}}{(1-e^{-x})^2} dx, \quad (9)$$

where θ_D is the Debye temperature in Kelvin and $R = 8.3144598(48) \text{ J/mol}^{-1}\text{K}^{-1}$ is the gas constant. The Bayesian approach used Simpson’s rule with 100 samples to evaluate the Debye model, while the Frequentist approach used Gauss-Kronrod quadrature for the integration [11], [19].

The second term is used to decompose linearly dependent physical effects, which can appear in low and high temperatures,

$$C_p^{bcm}(T, \beta_1, \beta_2, \alpha, \tau) = \begin{cases} \beta_1 T, & T < \alpha - \tau \\ \beta_1 T + \beta_2 \frac{(T - \alpha + \tau)^2}{4\tau}, & \alpha - \tau \leq T \leq \alpha + \tau. \\ \beta_1 T + \beta_2 (T - \tau), & T > \alpha + \tau \end{cases} \quad (10)$$

Here, the upper index ‘bcm’ stands for the bent-cable model [20], which is a continuous segmented model with four parameters $\beta_1, \beta_2, \alpha, \tau$ going through three segments (intervals). The first linear interval (‘incoming’) has a slope of β_1 and an intercept equal to 0 and it is valid for all temperatures satisfying $T < \alpha - \tau$. The second linear segment (‘outgoing’) has a slope of $\beta_1 + \beta_2$ and an intercept of $\beta_2 \tau$ and is valid for temperatures $T > \alpha + \tau$. Both of these segments are smoothly joined at the points $\alpha - \tau$ and $\alpha + \tau$ by a quadratic bend with half-weight $\gamma > 0$.

The last term in equation (6) is used to consider magnetic contributions if necessary and it is currently described by the Inden–Hillert–Jarl model [21], [22]. Since there is no magnetic effect for heat capacity for pure Al and Hf, we set $C_p^{mag}(T)$ to zero in this work.

3.2 Comparison of model predictions for SR model using frequentist and Bayesian approaches

In this section, we compare the results of the frequentist and Bayesian approaches in the development of models for the low temperature specific heat of hafnium and aluminum metal. Experimental and simulated data for both elements was collected from the literature. In the case of hafnium metal, the specific heat and enthalpy measurements were corrected for Zirconium impurity, as this element is difficult to separate from hafnium. When possible, reported errors, measurement methods, and impurity levels were collected for the purposes of comparing these metrics against the automatically computed weights.

The collected data for the specific heat of the low temperature phase is used to calibrate the model parameters of the SR model (Eq. 6) via the frequentist and Bayesian methodologies. For both approaches, the parameters are calibrated with and without use of the automated dataset weighting schemes. This enables us to evaluate the effect of both weighting schemes versus the differences between frequentist and Bayesian schemes in calibrating the parameters of the SR model. We plot the predictions of the frequentist and Bayesian schemes both with and without weights for the specific heat and enthalpy for Al (Fig. 1) and Hf (Fig. 2). Experimental and simulated datasets are identified via markers, predictions without weights are plotted as dotted black lines, and predictions with weights are represented by solid black lines. In all cases, the model predictions align well with the trends of the available data, and differences between the predictions are minor. Furthermore, the differences between the predictions with and without

weights are smaller than the differences due to using a frequentist or Bayesian inference methodology. The differences with and without weights would certainly be larger if there were larger discrepancies between the datasets considered [5].

For the specific heat of hafnium, the frequentist method results in predictions that have higher values at low temperatures and lower values at high temperatures compared with the Bayesian method. The enthalpy predictions are nearly indistinguishable. In all cases, predictions with and without weights are coincident. The frequentist predictions for aluminum trend lower than the Bayesian predictions at high temperatures for both specific heat and enthalpy properties. The predictions with and without weights are coincident except for the frequentist specific heat models, in which case the prediction with weights trends slightly higher than the prediction without.

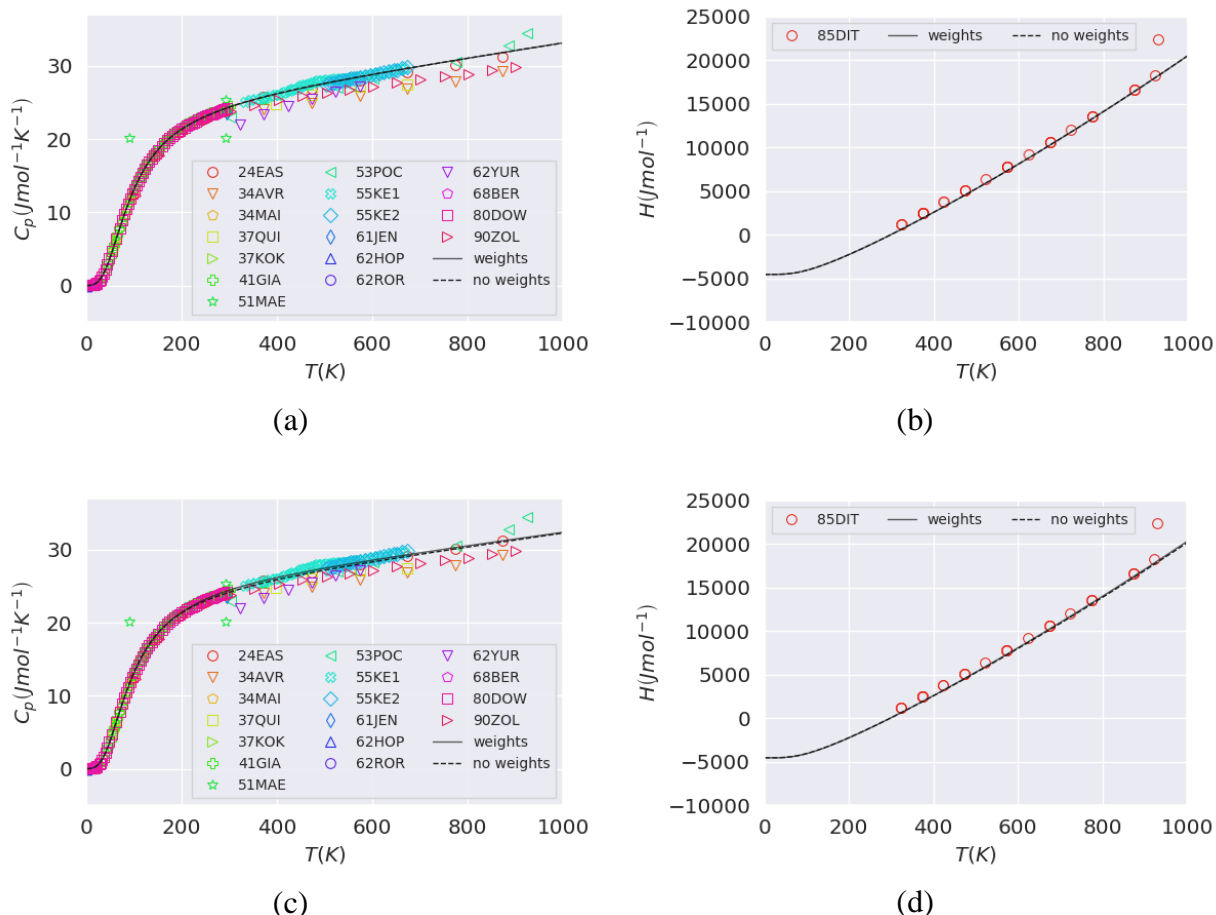


Fig. 1 The calculated (a) specific heat using Bayesian estimated, (b) enthalpy using Bayesian estimated, (c) specific heat using nonlinear regression and (d) enthalpy using nonlinear regression estimated SR parameters are compared with experiments for pure Al.

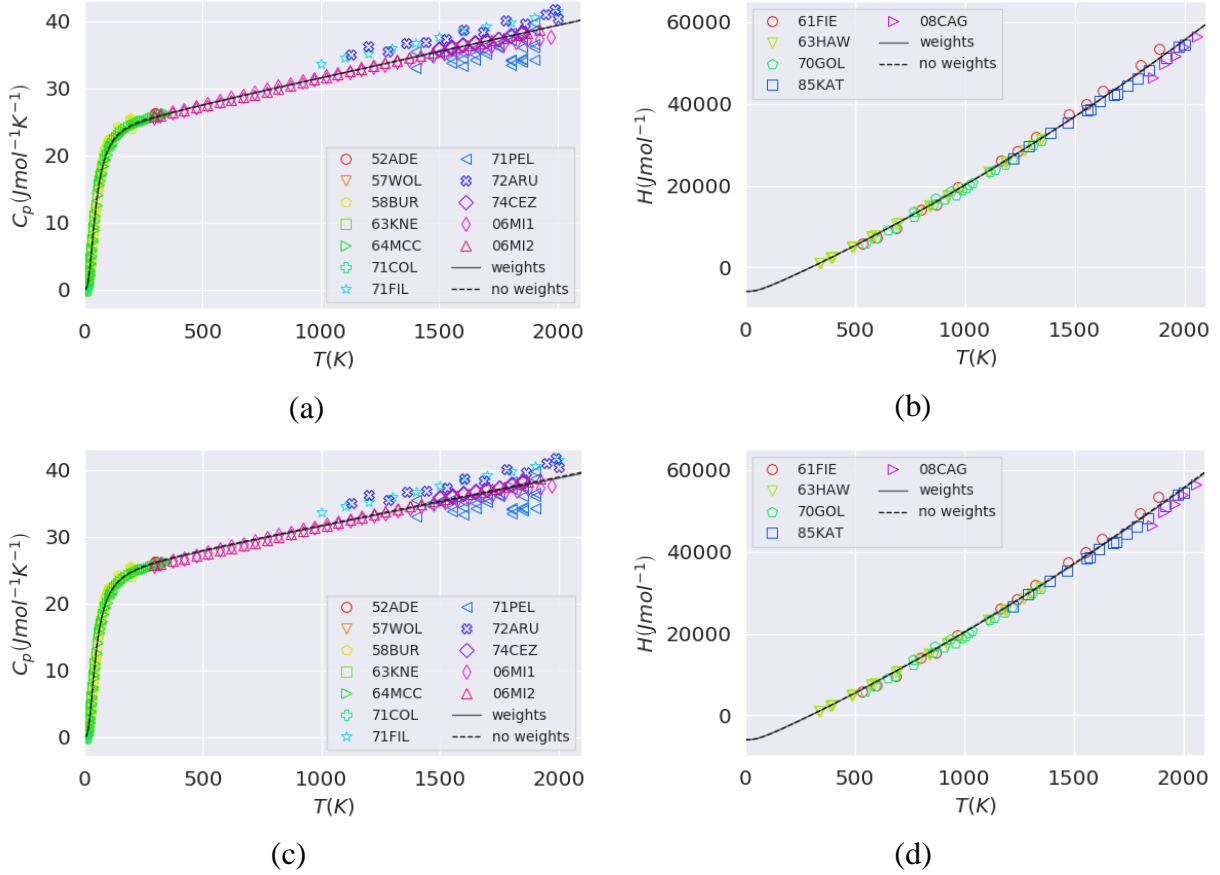


Fig. 2 The calculated (a) specific heat using Bayesian estimated, (b) enthalpy using Bayesian estimated, (c) specific heat using nonlinear regression and (d) enthalpy using nonlinear regression estimated SR parameters are compared with experiments for pure Hf.

The differences between the approaches with and without weighting is mirrored in the parameters themselves. Table 2 and Table 3 show the parameter values for both methods with and without weights for aluminum and hafnium, respectively. Out of all of the parameters, the trends in the Debye temperature parameter, θ_D , are the most consistent. For Al, θ_D is nearly identical between the frequentist and Bayesian methods without weights, and slightly increases when weights are used. For Hf, θ_D , is nearly identical for the methods, with and without weights. The remaining parameters show greater discrepancies. For example, the α parameter for Hf is systematically larger for the Bayesian model than for the frequentist model without weights; however, these values converge when weights are employed.

Table 2. Comparison of model parameters (pure Al)

Parameter	Nonlinear regression analysis		Bayesian analysis	
	Without weights	With weights	Without weights	With weights
θ_D	3.882e+02	3.923e+02	3.884e+02	3.906e+02
β_1	1.898e-03	1.819e-03	2.741e-04	1.528e-03
β_2	7.170e-03	6.917e-03	9.688e-03	8.457e-03
α	2.147e+02	1.582e+02	1.630e+02	1.880e+02
τ	1.186e+02	5.663e+01	1.020e+02	1.065e+02

Table 3. Comparison of model parameters (pure Hf)

Parameter	Nonlinear regression analysis		Bayesian analysis	
	Without weights	With weights	Without weights	With weights
θ_D	2.012e+02	2.007e+02	2.045e+02	2.065e+02
β_1	2.000e-03	2.000e-03	3.589e-03	1.441e-03
β_2	5.189e-03	5.162e-03	4.216e-03	6.326e-03
α	7.880e+01	1.007e+02	2.850e+02	1.766e+02
τ	7.000e+01	7.000e+01	2.218e+02	1.683e+02

3.3 Comparison of calculated weights

In this section, we compare the frequentist calculated weights to the weights calculated using the Bayesian method. As previously discussed, the frequentist weights range between zero to one, while the Bayesian ones range between zero and infinity. To compare the weights on an equal basis, we rescale the Bayesian weights to range between zero and one. To perform this reweighting, the 50th percentile Bayesian weight parameter for each dataset is first multiplied by the 50th percentile of the reported error. This results in a single rescaled error value for each dataset. Finally, we scale these error values between 0 and 1 based on the minimum and maximum values in the set. This re-scaling approach was not applied in the previous study and is novel to this work.

Tables 4 and 5 show the comparison of the frequentist and Bayesian weights for Al and Hf, respectively. In both tables, the weights are sorted from smallest to largest, and are color-coded according to the frequentist weight rankings (e.g. for Al, the 80DOW dataset is colored dark blue for both frequentist and Bayesian approaches). In this representation, it is clear that the weighting schemes prioritize the same datasets. For Al, the three lowest and highest weighted datasets are the same, and for Hf, the 5 lowest and 3 highest are the same. In other words, there is strong agreement about the lowest and highest weighted datasets, with moderate agreement for intermediately weighted datasets.

Table 4. Comparison of dataset weight rankings (pure Al)

Cross-validation (low to high)		Bayesian Statistics (low to high)	
51MAE [23]	0.000246	51MAE [23]	0
34AVR [24]	0.000246	34AVR [24]	0.039386
62YUR [25]	0.000246	62YUR [25]	0.226163
37QUI [26]	0.000246	90ZOL [27]	0.415104
53POC [28]	0.000246	37QUI [26]	0.438183
61JEN [29]	0.000246	53POC [28]	0.611341
24EAS [30]	0.000246	61JEN [29]	0.816454
90ZOL [27]	0.635229	24EAS [30]	0.817753
55KE1 [31]	0.799624	55KE1 [31]	0.926775
55KE2 [31]	0.844494	41GIA [32]	0.935426
41GIA [32]	0.936435	80DOW [33]	0.937149
34MAI [34]	0.946655	34MAI [34]	0.957469
80DOW [33]	0.963959	55KE2 [31]	0.962355
68BER [35]	0.997032	68BER [35]	0.998937
37KOK [36]	0.999105	62ROR [37]	0.999659
62HOP [38]	0.999859	37KOK [36]	0.999837
62ROR [37]	1	62HOP [38]	0.999997

Table 5. Comparison of dataset weight rankings (pure Hf)

Cross-validation (low to high)		Bayesian Statistics (low to high)	
71FIL [39]	0.004783507	71FIL [39]	0
52ADE [40]	0.004783507	71PEL [41]	0.247404
72ARU [42]	0.057835811	72ARU [42]	0.34173
71PEL [41]	0.248235519	58BUR [43]	0.638464
58BUR [43]	0.811417897	52ADE [40]	0.724224
74CEZ [44]	0.847799897	06MI1 [45]	0.904238
06MI1 [45]	0.855777484	64MCC [46]	0.915183
64MCC [46]	0.865621441	74CEZ [44]	0.924214
06MI2 [45]	0.894204799	06MI2 [45]	0.946867
57WOL [47]	0.978538078	57WOL [47]	0.992915
71COL [48]	0.999056293	63KNE [49]	0.999246
63KNE [49]	1	71COL [48]	0.999303

Figure 3 displays the frequentist and Bayesian weights of each dataset for Al and Hf, ordered by the frequentist weights. For Al, the lowest and highest weights are very close, and the largest discrepancies are in the middle. This can be expected, as the frequentist weights are generally close to 0 or 1, while the rescaled Bayesian weights span the range. In contrast, the weights for Hf are remarkably close throughout the range, with the largest discrepancy for the 52ADE dataset. This

discrepancy is not surprising as the 52ADE dataset contains only three data points, meaning its associated weight has little impact on the overall prediction. For both elements, the agreement between the weights of the frequentist and Bayesian methods is good. This indicates that the weighting schemes, even though they are algorithmically quite different, prioritize the same datasets.

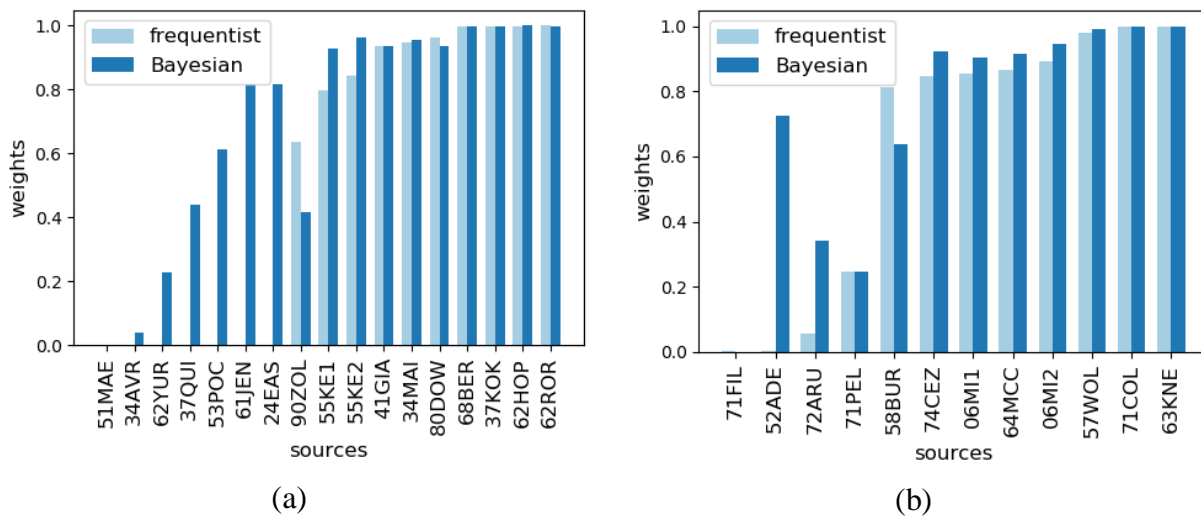


Figure 3: Frequentist and Bayesian weights are plotted for each data source for (a) Al, and (b) Hf.

Finally, we attempt to evaluate the degree of correlation between the weights for Al and Hf versus common metrics that might be considered in the manual weighting of datasets. For this investigation, we consider the dataset publication year, the reported percent error and the total impurity level. For the first three metrics, we simply evaluate the Pearson correlation and p-value versus the calculated weights from the frequentist or Bayesian method. We further hypothesize that higher correlations might be observed after separating the data by the methods used to obtain the data, out of which only calorimetry and pulse heating have sufficient examples. The conclusion of this analysis is that the p-values for the Pearson correlations are all greater than the standard 0.05 cutoff. In other words, there is a greater than 5% chance that an uncorrelated system could result in a greater or equal Pearson correlation. Consequently, one cannot satisfactorily evaluate the correlation of the weights to common metrics of data reliability. This is an interesting result because it reinforces an understanding of the assessment of thermodynamic datasets as an extremely complex and difficult process. It is not sufficient to use simple metrics of data reliability for the manual assessment of thermodynamic data – it may require a combination of these alongside a deeper understanding of the method used to obtain the data, the consistency with trusted datasets, and expert intuition. This result also strengthens the argument for the use of automated weighting schemes, as they are able to find consistency between datasets without human intervention. Such methods have the potential to reduce the effort required to weight datasets in thermodynamic analysis.

4. Conclusions

In this study, we compare two automated statistical approaches to weight datasets for the construction of models for the thermodynamic properties of materials. The two approaches, the first based in frequentist statistics (and presented for the first time in this work) and the second in Bayesian statistics, are employed to construct models for the specific heat and enthalpy of elemental aluminum and hafnium metals. Both methods are successful in fitting the segmented regression models to the data, and result in weights that indicate a similar prioritization of the available datasets. Consequently, the choice to use one method over another comes down to the needs of the researcher. The Bayesian approach requires more initial setup and incurs greater computational cost as compared to the classical method; however, it is more robust in inference whereas the classical approach require multiple initializations to avoid local minima in the cost function.

While these methods have different statistical foundations, they are both adept at finding consistency between the available datasets and lessening the effect of outliers. Such automated statistical approaches have the opportunity to reduce the burden of constructing thermodynamic models. Firstly, these schemes automatically judge the consistency between datasets. This is a significant advantage when attempting to ensure the consistency between datasets of different, but related properties, such as specific heat and enthalpy [5], or when the data are high dimensional and difficult to visualize. Additionally, a low weight might serve as a flag that indicates that a specific dataset might require additional scrutiny. Challenges remain in the automated weighting of datasets. Most notably, the weighting schemes prioritize datasets with many points as well as groups of datasets that show mutual consistency whether or not this is due to some systematic error (e.g. improper device calibration or poor sample purity). Future improvements should target these issues, but there will still always be a need to use physics-informed judgements to resolve ambiguity. These tools cannot replace the expert in the assessment of thermodynamic data, but should be viewed powerful allies in enhancing the quality and efficiency of the assessment process.

Acknowledgements

N.H.P and M.S acknowledge financial support from awards 70NANB14H012 and 70NANB19H005 from U.S. Department of Commerce, National Institute of Standards and Technology as part of the Center for Hierarchical Materials Design (CHiMaD), and Laboratory Directed Research and Development (LDRD) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-06CH11357. IR and SZ acknowledge financial support from the Collaborative Research Center “Superalloys Single Crystal” (TR-103 project C6) of the German Research Foundation (DFG) (Grant No. DFG TRR103).

Appendix

Table 6. Calculated weights using CV method (pure Al)

Reference	No. of Obs.	RSE (Eq. 2)	Weight (Eq. 3)
24EAS [30]	6	4.222533	0.000246
34AVR [24]	6	4.222533	0.000246
34MAI [34]	37	0.226294	0.946655
37KOK [36]	45	0.004818	0.999105
37QUI [26]	4	4.222533	0.000246
41GIA [32]	66	0.269446	0.936435
51MAE [23]	4	4.222533	0.000246
53POC [28]	7	4.222533	0.000246
55KE1 [31]	23	0.657672	0.844494
55KE2 [31]	46	0.847137	0.799624
61JEN [29]	1	4.222533	0.000246
62HOP [38]	20	0.001636	0.999859
62ROR [37]	121	0.001041	1
62YUR [25]	6	4.222533	0.000246
68BER [35]	22	0.013573	0.997032
80DOW [33]	68	0.153227	0.963956
90ZOL [27]	17	1.541300	0.635229

Table 7. Calculated weights using CV method (pure Hf)

Reference	No. of Obs.	RSE (Eq. 2)	Weight (Eq. 3)
52ADE [40]	3	3.657471	0.004783
57WOL [47]	56	0.095992	0.978538
58BUR [43]	74	0.707229	0.811418
63KNE [49]	43	0.017496	1
64MCC [46]	82	0.508981	0.865621
71COL [48]	27	0.020947	0.999056
71FIL [39]	11	3.657471	0.004783
71PEL [41]	37	2.767053	0.248236
72ARU [42]	13	3.463434	0.057836
74CEZ [44]	24	0.574163	0.847800
06MI1 [45]	35	0.544985	0.855777
06MI2 [45]	34	0.404438	0.894205

Table 8: Calculated weights using Bayesian method (pure A1)

dataset	50%ile hyperparameter	50%ile error	rescaled error	normalized weights
24EAS [30]	0.67048	0.285352	0.425593	0.817753
34AVR [24]	0.117587	0.26378	2.243276	0.039386
34MAI [34]	1.637876	0.162674	0.09932	0.957469
34QUI [26]	0.202992	0.266322	1.311985	0.438183
37KOK [36]	0.111336	4.25E-05	0.000382	0.999837
41GIA [32]	1.207238	0.182048	0.150797	0.935426
51MAE [23]	0.097476	0.227631	2.335253	0
53POC [28]	0.312349	0.283493	0.907617	0.611341
55KE1 [31]	1.637447	0.280003	0.170999	0.926775
55KE2 [31]	3.22483	0.283493	0.08791	0.962355
61JEN [29]	0.553472	0.237233	0.428627	0.816454
62HOP [38]	3.016577	1.83E-05	6.05E-06	0.999997
62ROR [37]	0.033703	2.68E-05	0.000795	0.999659
62YUR [25]	0.138467	0.250224	1.807105	0.226163
68BER [35]	0.048155	0.00012	0.002482	0.998937
80DOW [33]	1.103579	0.161975	0.146772	0.937149
90ZOL [27]	0.192469	0.26289	1.36588	0.415104

Table 9. Calculated weights using Bayesian method (pure Hf)

dataset	50%ile hyperparameter	50%ile error	rescaled error	normalized weights
52ADE [40]	0.466994	0.262141	0.561337	0.724224
57WOL [47]	0.048744	0.000703	0.014421	0.992915
58BUR [43]	0.235161	0.173055	0.735899	0.638464
63KNE [49]	0.066493	0.000102	0.001536	0.999246
64MCC [46]	0.861516	0.148735	0.172643	0.915183
71COL [48]	0.073487	0.000104	0.001419	0.999303
71FIL [39]	0.184614	0.375779	2.035479	0
71PEL [41]	0.230607	0.353266	1.531894	0.247404
72ARU [42]	0.195356	0.384055	1.965921	0.034173
74CEZ [44]	2.393903	0.369286	0.154261	0.924214
06MI1 [45]	1.664750	0.324495	0.194921	0.904238
06MI2 [45]	2.986378	0.322978	0.10815	0.946867

Data Availability Statement

The raw and processed data required to reproduce these findings are available to download from github.com/npaulson/UnaryBayes [50].

References

- [1] H. L. Lukas, E. T. Henig, and B. Zimmermann, "Optimization of phase diagrams by a least squares method using simultaneously different types of data," *Calphad*, vol. 1, no. 3, pp. 225–236, 1977.
- [2] A. V Davydov *et al.*, "Determination of the CoTi congruent melting point and thermodynamic reassessment of the Co-Ti system," *Metall. Mater. Trans. A*, vol. 32, no. 9, pp. 2175–2186, Sep. 2001.
- [3] R. Schmid-Fetzer *et al.*, "Assessment techniques, database design and software facilities for thermodynamics and diffusion," *Calphad*, vol. 31, no. 1, pp. 38–52, 2007.
- [4] H. Lukas, S. G. Fries, and B. Sundman, *Computational thermodynamics: the Calphad method*. Cambridge university press, 2007.
- [5] N. H. Paulson, E. Jennings, and M. Stan, "Bayesian strategies for uncertainty quantification of the thermodynamic properties of materials," *Int. J. Eng. Sci.*, vol. 142, pp. 74–93, 2019.
- [6] R. R. Picard and R. D. Cook, "Cross-Validation of Regression Models," *J. Am. Stat. Assoc.*, vol. 79, no. 387, pp. 575–583, 1984.
- [7] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. CRC press, 2013.
- [8] Y.-Z. Ma and A. Berndsen, "How to combine correlated data sets—A Bayesian hyperparameter matrix method," *Astron. Comput.*, vol. 5, pp. 45–56, 2014.
- [9] G. A. F. Seber and C. J. Wild, "Nonlinear Regression. Hoboken," *New Jersey John Wiley Sons*, vol. 62, p. 63, 2003.
- [10] N. H. Paulson, E. Jennings, and M. Stan, "Bayesian strategies for uncertainty quantification of the thermodynamic properties of materials," *arXiv Prepr. arXiv1809.07365*, 2018.
- [11] I. Roslyakova, B. Sundman, H. Dette, L. Zhang, and I. Steinbach, "Modeling of Gibbs energies of pure elements down to 0K using segmented regression," *Calphad*, vol. 55, pp. 165–180, 2016.
- [12] I. Roslyakova *et al.*, "Third generation CALPHAD databases: New unary database and its application for re-assessment of binary systems," in *CALPHAD XLVII Conference*, 2018.
- [13] D. Sergeev *et al.*, "Comprehensive analysis of thermodynamic properties of calcium nitrate," *J. Chem. Thermodyn.*, vol. 134, pp. 187–194, 2019.
- [14] Y. Jiang, S. Zomorodpoosh, I. Roslyakova, and L. Zhang, "Thermodynamic re-assessment of binary Cr-Nb system down to 0 K," *Calphad*, vol. 62, pp. 109–118, 2018.
- [15] Y. Jiang, S. Zomorodpoosh, I. Roslyakova, and L. Zhang, "Thermodynamic re-assessment of binary Cr-Ta system down to 0 K," *Int. J. Mater. Res.*, 2019.
- [16] B. Bocklund, R. Otis, A. Egorov, A. Obaied, I. Roslyakova, and Z.-K. Liu, "ESPEI for efficient thermodynamic database development, modification, and uncertainty quantification: application to Cu-Mg," *arXiv Prepr. arXiv1902.01269*, 2019.
- [17] Q. Chen and B. Sundman, "Modeling of thermodynamic properties for Bcc, Fcc, liquid, and amorphous iron," *J. Phase Equilibria*, vol. 22, no. 6, pp. 631–644, Nov. 2001.
- [18] P. Debye, "Zur theorie der spezifischen wärmen," *Ann. Phys.*, vol. 344, no. 14, pp. 789–839, 1912.
- [19] R. Piessens, E. de Doncker-Kapenga, C. W. Überhuber, and D. K. Kahaner, *Quadpack: a*

- subroutine package for automatic integration*, vol. 1. Springer Science & Business Media, 2012.
- [20] G. Chiu, R. Lockhart, and R. Routledge, “Bent-Cable Regression Theory and Applications,” *J. Am. Stat. Assoc.*, vol. 101, no. 474, pp. 542–553, 2006.
- [21] G. Inden, “Approximate description of the configurational specific heat during a magnetic order-disorder transformation,” *Proc. CALPHAD V Dusseldorf, Ger. Max Planck Inst. fuer Eisenforsch.*, pp. 1–13, 1976.
- [22] M. Hillert and M. Jarl, “A model for alloying in ferromagnetic metals,” *Calphad*, vol. 2, no. 3, pp. 227–238, 1978.
- [23] H. Mäder, “Mechanical and Physical Properties of Pure Aluminum and a Few Aluminum Alloys at the Temperature of Liquid Oxygen,” *Metall.*, pp. 1–5, 1951.
- [24] A. Avramescu, “Temperaturabhängigkeit der wahren spezifischen Wärme von Leitungskupfer und Leitungsaluminium bis zum Schmelzpunkt,” *Zeitschrift fur Tech. Phys.*, vol. 20, no. 7, pp. 213–217, 1939.
- [25] Y. A. Yurkov and L. A. Ivoninskaya, “Thermodynamic Properties of Aluminum,” *News High. Educ. Institutions Phys.*, vol. 1, pp. 138–143, 1962.
- [26] H. Quinney and G. I. Taylor, “The emission of the latent energy due to previous cold working when a metal is heated,” *Proc. R. Soc. London. Ser. A - Math. Phys. Sci.*, vol. 163, no. 913, pp. 157–181, 1937.
- [27] M. Zoli and V. Bortolani, “Thermodynamic properties of FCC metals: Cu and Al,” *J. Phys. Condens. Matter*, vol. 2, no. 3, pp. 525–539, Jan. 1990.
- [28] T. E. Pochapsky, “Heat capacity and resistance measurements for aluminum and lead wires,” *Acta Metall.*, vol. 1, no. 6, pp. 747–751, 1953.
- [29] W. J. Parker, R. J. Jenkins, C. P. Butler, and G. L. Abbott, “Flash Method of Determining Thermal Diffusivity, Heat Capacity, and Thermal Conductivity,” *J. Appl. Phys.*, vol. 32, no. 9, pp. 1679–1684, 1961.
- [30] E. D. Eastman, A. M. Williams, and T. F. Young, “The specific heats of magnesium, calcium, zinc, aluminum and silver at high temperatures,” *J. Am. Chem. Soc.*, vol. 46, no. 5, pp. 1178–1183, 1924.
- [31] K. Hirano, H. Maniwa, and Y. Takagi, “Specific-Heat Measurements on Quench-Annealed Al, Cu and alpha-phase Alloys of Cu,” *J. Phys. Soc. Japan*, vol. 10, no. 10, pp. 909–910, 1955.
- [32] W. F. Giauque and P. F. Meads, “The Heat Capacities and Entropies of Aluminum and Copper from 15 to 300° K.,” *J. Am. Chem. Soc.*, vol. 63, no. 7, pp. 1897–1901, 1941.
- [33] D. B. Downie and J. F. Martin, “An adiabatic calorimeter for heat-capacity measurements between 6 and 300 K. The molar heat capacity of aluminium,” *J. Chem. Thermodyn.*, vol. 12, no. 8, pp. 779–786, 1980.
- [34] C. G. Maier and C. T. Anderson, “The Disposition of Work Energy Applied to Crystals,” *J. Chem. Phys.*, vol. 2, no. 8, pp. 513–527, 1934.
- [35] W. T. Berg, “Heat Capacity of Aluminum between 2.7 and 20°K,” *Phys. Rev.*, vol. 167, no. 3, pp. 583–586, Mar. 1968.
- [36] J. A. Kok and W. H. Keesom, “Measurements of the atomic heat of aluminium from 1.1 to 20° K,” *Physica*, vol. 4, no. 9, pp. 835–842, 1937.
- [37] D. C. Rorer, M. Horst, and R. C. Richardson, “Specific Heat of Aluminum Near its Superconductive Transition Point,” *Zeitschrift für Naturforsch. A*, vol. 18, pp. 130–140, 1963.

- [38] D. C. Hopkins, "Thermodynamic properties of aluminum near its superconducting critical temperature," 1962.
- [39] L. P. Filippov and R. P. Yurchak, "High-temperature investigations of the thermal properties of solids," *J. Eng. Phys.*, vol. 21, no. 3, pp. 1209–1220, 1971.
- [40] H. K. Adenstedt, "Physical, thermal and electrical properties of hafnium and high purity zirconium," *Trans. Am. Soc. Met.*, vol. 44, pp. 949–973, 1952.
- [41] V. E. Peletskii and V. P. Druzhinin, "Experimental study of some physical properties of hafnium at high temperatures," *Teplofiz. Vysok. Temp.*, vol. 9, no. 3, pp. 539–545, 1971.
- [42] A. V Arutyunov, S. N. Banchila, and L. P. Filippov, "Thermal, Electrical and Emissive Properties of Hf in the High-Temperature Region," *High Temp.*, vol. 10, no. 2, pp. 375–377, 1972.
- [43] D. L. Burk, I. Estermann, and S. A. Friedberg, "The low temperature specific heats of titanium, zirconium and hafnium," *Z. Phys. Chem.(Munich)*, vol. 16, pp. 183–193, 1958.
- [44] A. Cezairliyan and J. L. McClure, "Simultaneous measurements of specific heat, electrical resistivity, and hemispherical total emittance by a pulse heating technique: hafnium--3 (wt.%) zirconium, 1500 to 2400 K," *J. Res. Natl. Bur. Stand., A*, vol. 79, no. 2, pp. 431–436, 1975.
- [45] N. D. Milošević and K. D. Maglić, "Thermophysical Properties of Solid Phase Hafnium at High Temperatures," *Int. J. Thermophys.*, vol. 27, no. 2, pp. 530–553, Mar. 2006.
- [46] T. A. McClaine, "Thermodynamic and kinetic studies for a refractory materials program," 1964.
- [47] N. M. Wolcott, "The atomic heats of titanium, zirconium and hafnium," *Philos. Mag. A J. Theor. Exp. Appl. Phys.*, vol. 2, no. 22, pp. 1246–1254, 1957.
- [48] E. W. Collings and J. C. Ho, "Magnetic-Susceptibility and Low-Temperature Specific-Heat Studies of Ti, Zr, and Hf," *Phys. Rev. B*, vol. 4, no. 2, pp. 349–356, Jul. 1971.
- [49] G. D. Kneip Jr, J. O. Betterton Jr, and J. O. Scarbrough, "Low-temperature specific heats of titanium, zirconium, and hafnium," *Phys. Rev.*, vol. 130, no. 5, p. 1687, 1963.
- [50] N. H. Paulson, "Aluminum and Hafnium Data," 2019. [Online]. Available: github.com/npaulson/UnaryBayes.