A Comprehensive Methodology for Development, Parameter Estimation, and Uncertainty Analysis of Group Contribution Based Property Models -An Application to the Heat of Combustion

Frutiger, Jerome; Marcarie, Camille; Abildskov, Jens; Sin, Gürkan

Published in:
Journal of Chemical and Engineering Data

Link to article, DOI:
10.1021/acs.jced.5b00750

Publication date:
2016

Document Version
Peer reviewed version

A comprehensive methodology for development, parameter estimation and uncertainty analysis of group contribution based property models – an application to heat of combustion

*Jérôme Frutiger, Camille Marcarie, Jens Abildskov, Gürkan Sin**

Department of Chemical and Biochemical Engineering, Technical University of Denmark (DTU), Building 229, DK-2800 Lyngby, Denmark

ABSTRACT: A rigorous methodology is developed that addresses numerical and statistical issues when developing group contribution (GC) based property models such as regression methods, optimization algorithms, performance statistics, outlier treatment, parameter identifiability and uncertainty of the prediction. The methodology is evaluated through development of a GC method for prediction of the heat of combustion (ΔH_c°) for pure components. The results showed that robust regression lead to best performance statistics for parameter estimation. Bootstrap method is found a valid alternative to calculate parameter estimation errors when underlying distribution of residuals is unknown. Many parameters (first, second, third order groups contributions) are found unidentifiable from the typically available data, with large estimation error bounds and significant correlation. Due to this poor parameter

identifiability issues, reporting of the 95%-confidence intervals of the predicted property values should be mandatory as opposed to reporting only single value prediction, currently the norm in literature. Moreover, inclusion of higher order groups (additional parameters) does not always lead to improved prediction accuracy for the GC-models, in some cases it may even increase the prediction error (hence worse prediction accuracy). However, additional parameters do not affect calculated 95%-confidence interval. Last but not least, the newly developed GC model of the heat of combustion (ΔH_c°) shows predictions of great accuracy and quality (the most data falling within the 95% confidence intervals) and provides additional information on the uncertainty of each prediction compared to other ΔH_c° models reported in literature.

INTRODUCTION

When experimental values are unavailable due to cost or time constraints, there is a strong demand for generating accurate and reliable data by predictions. In the early stage of process development, when a large number of alternative processes are evaluated and ranked, property data are often estimated, especially when new or alternative products or processes are analysed[1]. Thus, property prediction models are critically important to process systems engineering, e.g. process simulation, analysis and optimization as well as computer-aided molecular design (CAMD). Three main types of property prediction models are widely employed: group contribution (GC)[2], quantitative structure-property relationship (QSPR)[3] and *ab initio* quantum mechanics based methods[4].

GC based prediction of pure component properties uses a function of structurally dependent parameters. The best known GC methods are those of Joback and Reid[5], Lydersen[6], Klincewicz

and Reid[7], Constantinou/Gani[8] and Marrero/Gani[2]. Compared to *ab initio* procedures, GC methods have a simpler model structure, a wider application range and are computationally less demanding. The advantage of the GC approach compared to quantitative structure property relationship (QSPR) or prediction based on artificial neural networks (ANN) is that the model structure does not depend on the data set[9]. This means that GC models are likely to be more reliable for predicting properties of compounds not included in the original data set used for model building. The idea of a property function common to all species is in line with Pitzer´s corresponding states principle[10], often shown to be nearly valid for fluid properties.

In GC model development, the key task is estimation of group contributions using experimental data. Systematic reporting of uncertainty for experimental values is widely used[11]. Hence, assessing uncertainty of both estimated parameters and predicted properties is appropriate, but this issue has nevertheless traditionally not been systematically reported. While the importance of uncertainty analysis has been recognized in the literature (Whiting[12], Larsen[13], Klotz and Mathias[14], Hajipourt and Satyro[15], Maranas[16], Yan[17], Verevkin[18]), the quantification of the source of uncertainties itself (e.g. property prediction errors associated with any property models) has not received much attention. For example, Whiting[12] investigated the effects of uncertainties in thermodynamic data and models on process calculations, Larsen[13] suggested methods to analyse the data quality for chemical process design and Klotz and Mathias[14] compared van der Waals (vdW) equations of state (EOS) for specific properties. Furthermore, Hajipour and Satyro[15] illustrated the effect of uncertainty of models for critical constants and acentric factor and Maranas[16] performed an uncertainty analysis on optimization calculations involved in computer aided molecular design studies. Yan[17] compared the reliability of a variety of group contribution methods in predicting critical temperatures of organic compounds by

analysing the respective average absolute deviation. Verevkin et al.[18] proposed a new group-contribution approach involving systematic corrections for 1,4-nonbonded carbon-carbon and carbon-oxygen interactions. The authors considered uncertainties of predicted values. However, their modification of the covariance matrix calculation seems non-standard, as it is not based on known statistical theories for parameter estimation[19], and its generalization may not be straightforward.

Recently, the Marrero/Gani group contribution method (MG method) was used by Hukkerikar et al.[20] to estimate thermo-physical properties (e.g. flash point) of pure components. Hukkerikar et al. performed a GC parameter estimation based on maximum likelihood theory, an uncertainty analysis based on the parameter covariance matrix and performance criteria to assess the quality. In addition to Hukkerikar et al. there is a need for a comprehensive methodology that includes

- Formulation of parameter estimation problem (e.g. weighted least squares, ordinary least squares, robust regression)
- Performance of optimization algorithms used to locate minima of the objective function used for parameter estimation
- Additional alternative uncertainty analysis method
- Assessment of parameter estimation errors and of property model prediction errors
- Method to identify outliers and data pre-treatment
- Analysis of the source of uncertainty
- Effects of additional GC-factors on prediction and uncertainty

We aim at a methodology to perform a comprehensive and step-by-step assessment and solution of the above mentioned challenges involved in developing GC-based property models. We

demonstrate the methodology by developing a new GC model for the heat of combustion (ΔH_c°) based on the MG groups, employing molecular structural information at different levels.

The heat of combustion ΔH_c° provides important information in risk assessment in order to quantify the stabilities of chemical compounds. Furthermore the values are required when considering the thermal efficiency of process equipment in particular where either heat or power is produced. ΔH_c° is defined as the enthalpy increase of a chemical compound while undergoing an oxidation to defined combustion products at a temperature of 298.15 K and pressure of 1 atm[21].

There are a number of GC-based methods for the estimation of ΔH_c° in the literature. Cardozo[22] estimated enthalpies of combustion by developing correction factors from an equivalent alkane chain length and then utilized these factors along with simple relations developed for n-alkanes. Seaton and Harrison[23] proposed a method based on the original Benson's methods that had been used for the prediction of enthalpy of formation. Both Cardozo as well as Seaton and Harrison did not provide information on accuracy and uncertainty of their respective models. Hshieh et al.[24] developed an empirical model to estimate the heat of combustion. However, the application range is limited due to a small number of compounds taken into account for the parameter estimation.

Gharagheizi[25] developed a simple three-parameter quantitative structure-property relationship (QSPR). Cao et al.[26] suggested a model to estimate the heat of combustion based on an artificial neural network (ANN). Furthermore, Pan et al.[27] developed a four-parameter QSPR method. Recently Gharagheizi et al.[21] developed new GC model for the heat of combustion based on ANN. The latter four mentioned models showed all a high squared correlation coefficient

between the experimental and the predicted data (>0.99). However, none of the studies includes a thorough uncertainty analysis of model predictions including for example the 95%-confidence interval of the prediction or the covariance matrix of the parameters. As a case study to highlight the application of rigorous methodology developed in this study, we develop a novel GC-based model for estimation of ΔH_c° as well as provide comprehensive assessment of uncertainties and model prediction accuracy including 95% confidence interval demonstrating the added value of using the systematic methodology for the development of GC –based property models.

The paper is organized as follows: (i) the overall methodology is outlined; (ii) the property model for ΔH_c° is developed; (iii) results of parameter estimation, using different regression methods, combined with outlier detection and uncertainty analysis, are presented, and; (iv) the new ΔH_c° model performance is compared with that of existing models.

2. Method

An overview of the methodology including the workflow, the data and techniques used at each step is shown in Figure 1.

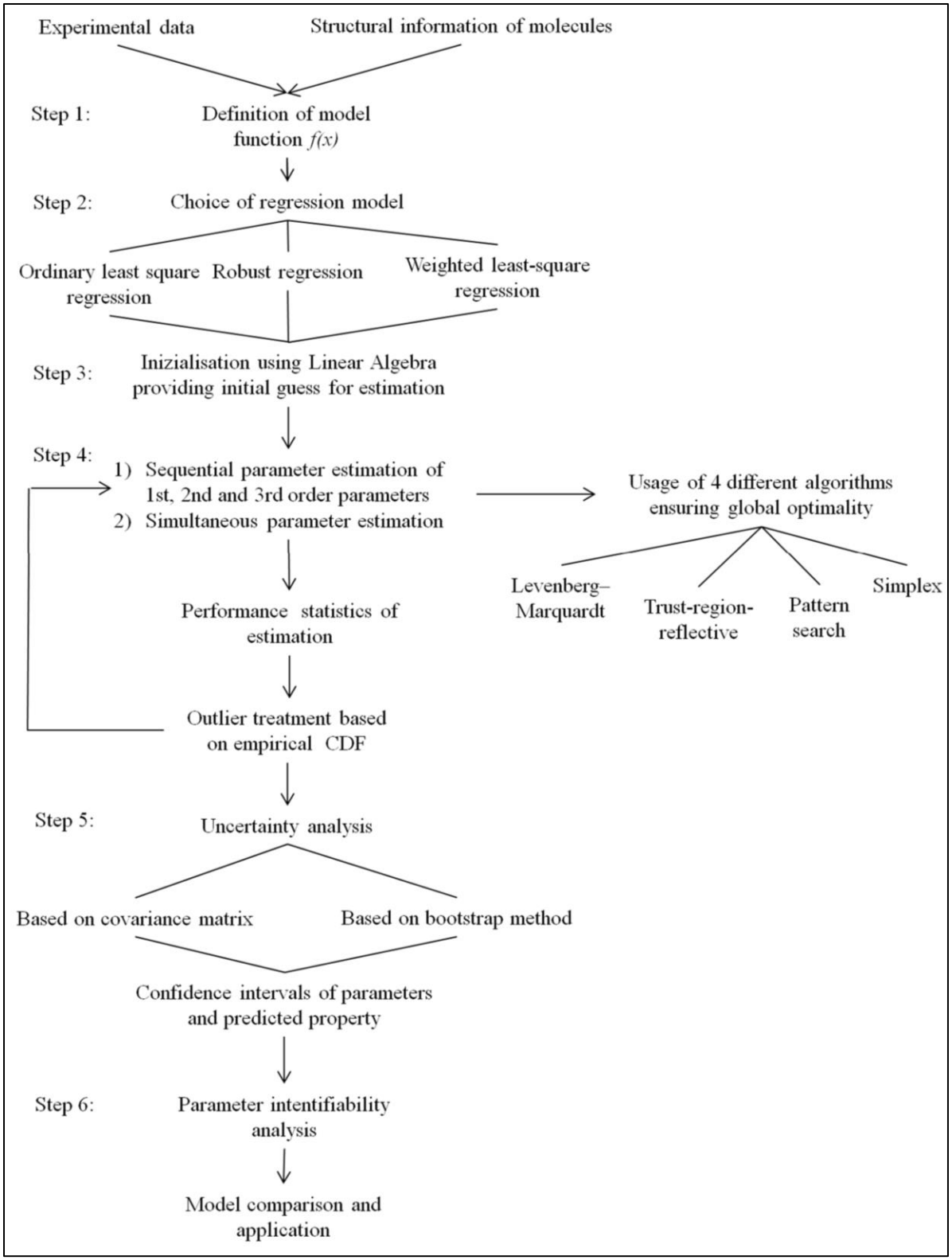

Figure 1. Overview of the methodology for development, parameter estimation and uncertainty analysis of group contribution based property models.

Detailed explanation of the tasks to perform when following the methodology is described in the following.

Property model structure definition and experimental data collection. Here the Marrero/Gani (MG)[2] method is selected for development. This method combines the contributions from a specific functional group (1st order parameters), from polyfunctional (2nd order parameters) as well as from structural groups (3rd order parameters). By using higher order parameters (2nd and 3rd), additional structural information about molecular fragments is provided. This may be useful, if the description given by 1st order groups is insufficient. The general form of the MG method is,

$$f_i(X) = \sum_j N_j C_j + \sum_k M_k D_k + \sum_l E_l O_l \tag{1}$$

$$f(X) = T \cdot \theta \tag{2}$$

In Eq. (1) C_j is the contribution of the 1st order group of type *j* that occurs N_j times whereas D_k is the contribution of the 2nd order group of type *k* that occurs M_k times in the molecular structure of a pure component. E_l is the contribution of the 3rd order group of type *l* that has O_l occurrences. The function *f(X)* is specific for a certain property *X*. The parameters can be collected in the vector θ and the occurrences of the groups can be depicted in the matrix *T* as shown in Eq. (2). As an example, the different GC-factors of 1,2-Dichloro-4-nitrobenzene and Adiponitrile are visualized in Figure 2.

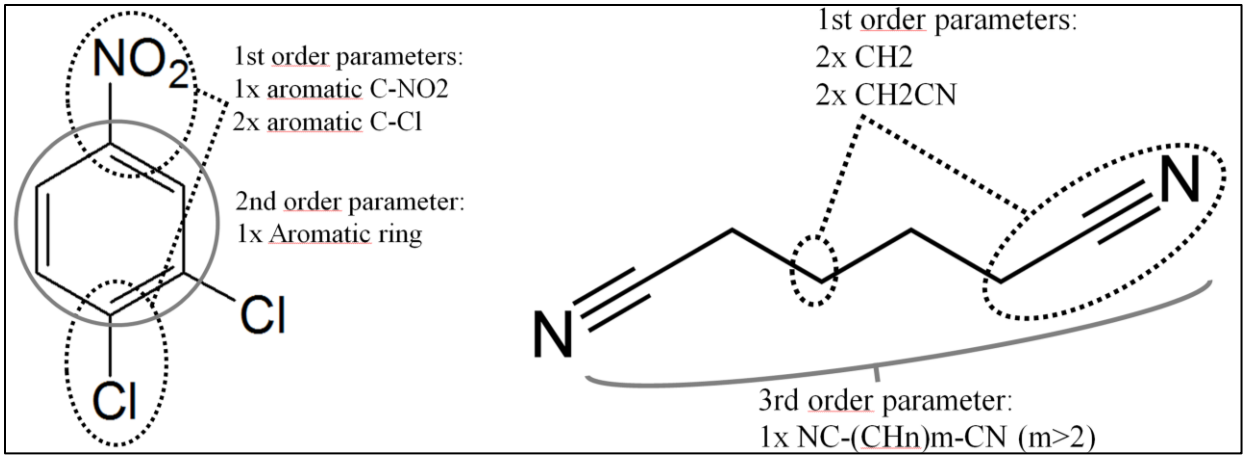

Figure 2. Example of Marrero/Gani group contribution factors of 1,2-Dichloro-4-nitrobenzene and Adiponitrile.

Experimental ΔH_c° data of 794 compounds are obtained from AIChE DIPPR 801 Database[28]. A high number of experimental data points is a prerequisite in order to obtain an accurate model with a wide application range. The heat of combustion of each compound is provided in kJ/mol.

After assigning the different 1st, 2nd and 3rd order groups to the respective molecules, it is necessary to determine a model function. We seek a function of the property which is linear in the group contributions. Hence, a suggestion for the property function is obtained by generating plots of various classes of pure components versus their increasing carbon number in homologue series as already shown by Pierotti et al.[29]. A selection of classes of compounds is shown in figure 3. From these plots, a linear function is deemed as appropriate model function for the ΔH_c° property and shown in Eq. (3), where $\Delta H_{c\,const}^\circ$ is a universal constant.

$$f(X) = f(\Delta H_c^\circ) = \Delta H_c^\circ - \Delta H_{c\,const}^\circ \tag{3}$$

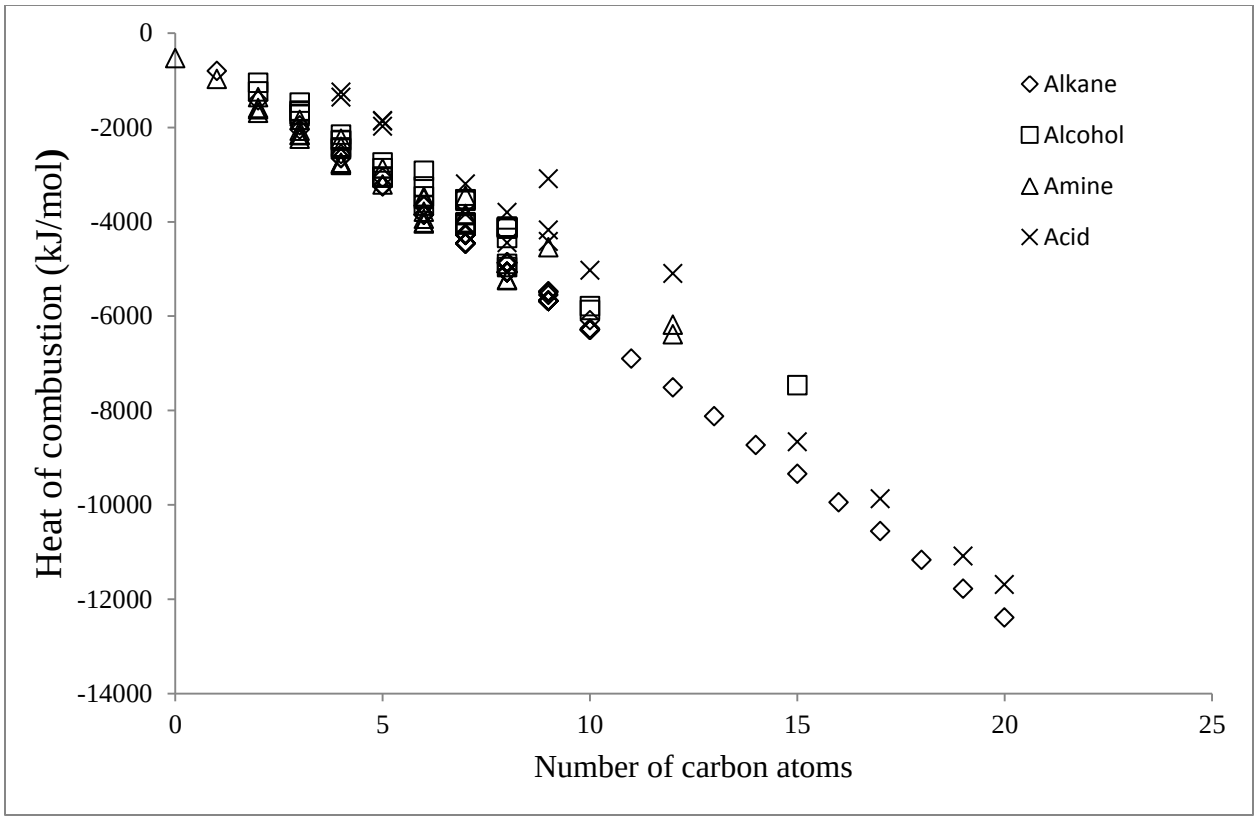

Figure 3. Graphical analysis of number of carbon atoms versus property to infer about a proper model function: (y-axis) heat of combustion ΔH_c° of a selection of pure components, (x-axis) carbon number of pure components in increasing order.

Choice of regression method. Three regression models are investigated for the use in parameter estimation in group contribution model development.

- Ordinary nonlinear least squares regression

- Robust regression

- Weighted nonlinear least squares regression

Ordinary nonlinear least squares regression is the most commonly used method for parameter estimation. The ordinary least squares regression minimizes the squares of the difference

between the experimental property value y^{exp} and the predicted property value y^{pred}, i.e. the residuals, in order to get the parameter estimates θ^*,

$$\theta^* = arg\,min \sum_i (y_i^{exp} - y_i^{pred})^2 \tag{4}$$

For the case of ΔH_c° y^{pred} is defined by combining Eq. (1) and (3), see Eq. (5). Each data point has equal weight (unity)[19],

$$y^{pred} = \Delta H^\circ_{c\,const} + \sum_j N_j C_j + \sum_k M_k D_k + \sum_l O_l E_l \tag{5}$$

Ordinary least squares regression assumes that the errors are ideally independently distributed and uncorrelated, following a Gaussian distribution with a mean value of zero and a constant variance[19]. While these assumptions are made, in practice their validity is rarely checked. This is the motivation for using a bootstrap method as outlined below.

In robust regression each residual is weighted by a certain factor w_i[19]. Here the Cauchy weight is used, placing high weights on small residuals and small weights on large residuals (see Eq. (6) and (7)). The weights are updated recursively. In this way the influence of data points producing large residuals (not following the model), i.e. potential outliers, is decreased. Another intrinsic property of robust regression is that a common variance of all data points is not assumed[30].

$$\theta^* = arg\,min \sum_i w_i \cdot (y_i^{exp} - y_i^{pred})^2 \tag{6}$$

$$w_i = \frac{1}{1 + (y_i^{exp} - y_i^{pred})^2} \tag{7}$$

Weighted non-linear least squares regression uses the variance V_i of the measurement error to weight the data as shown in Eq. (8)[31]. Data points with a high variance are considered to be less

reliable and hence their influence on the objective function is reduced. The variance of errors of the present experimental ΔH_c° measurements are obtained from the AIChE DIPPR 801 Database[28].

$$\theta^* = arg\,min \quad \sum_i V_i^{-1} \cdot (y^{exp} - y^{pred})^2 \tag{8}$$

$$V_i = \sigma_i^2 \tag{9}$$

In Eq. (9) σ_i is the standard deviation of the respective measurement error.

Initialization using linear algebra and sequential parameter estimation. The universal constant as well as the GC factors are (a priori) unknown. A first guess $\hat{\theta}$ for the parameter estimate is provided using linear algebra according to Eq. (10),

$$\hat{\theta} = (T^{tr} \cdot T)^{-1} \cdot T^{tr} \cdot f(X) \tag{10}$$

A value for the constant $\Delta H_{c\,const}^\circ$ is assumed in order to calculate the first guess for the parameters from ΔH_c° data and the occurrence matrix T. This offers a unique solution existing without iterations.

Sequential and simultaneous parameter estimation and verification of global optimality. Afterwards the universal constants as well as the 1st, 2nd and 3rd order parameters are estimated separately and sequentially applying the non-linear regression model from the previous step. The solution of Eq. (10) is used as input for the sequential parameter estimation in the next step.

The result of the sequential estimation serves as initial guess for the simultaneous parameter estimation algorithm, where all parameters are estimated together for the chosen regression problem. The purpose of this step is twofold: (a) integrated solution of the parameter estimation problem and (b) practical verification of global optimality of the parameter estimation solution. In order to test that the global minimum of the least-squares regression has been achieved, a practical approach is followed, in which 4 different optimization algorithms are applied.

Derivative based: - Levenberg–Marquardt algorithm[32]

- Trust-region reflective algorithm[33]

Non-derivative based: - Simplex algorithm[34]

- Pattern search optimization[35]

The Levenberg-Marquardt as well as the Trust-region reflective algorithm are based on the method of steepest descent and the line search approach. They differ in the solution of the quadratic subproblems[36]. Both algorithms are commonly known as computationally very fast compared to non-derivative based algorithms. However, if the parameter number is high and the parameters are a priori unknown (as in developing GC models), it is suggested to additionally use robust non-derivative based algorithm such as simplex and pattern search[34].

Statistical performance indicators for parameter estimation. Performance of the parameter estimates is quantified by a variety of statistics in order to obtain a broad set of measures. Hukkerikar et al.[20] adopted the following statistics:

- Sum of squared errors between the experimental and predicted data,

$$SSE = \sum_i (y_i^{exp} - y_i^{pred})^2 \tag{11}$$

- Standard deviation, *SD*, measures the spread of the data about the mean value μ_y,

$$SD = \frac{1}{N} \cdot \sqrt{\sum_j (y_j^{pred} - y_j^{exp})^2} \tag{12}$$

- R^2 between the experimental and the predicted values suggests the quality of the model fit by assessing linear correlation,

$$R^2 = \frac{\sum_j (y_j^{exp} - y_j^{pred})^2}{\sum_j (y_j^{exp} - \mu_y)^2} \tag{13}$$

R^2 close to 1 indicates that the experimental data used in the regression have been fitted to a good accuracy.

- Average absolute deviation (AAD) is the measure of deviation of predicted property values from the experimentally measured property values,

$$AAD = \frac{1}{N} \sum_j |y_j^{exp} - y_j^{pred}| \tag{14}$$

- Average relative error *ARE* provides an average of relative error calculated with respect to the experimentally measured property values,

$$ARE = \frac{1}{N} \sum_j \frac{(y_j^{exp} - y_j^{pred})}{y_j^{exp}} \tag{15}$$

- The percentage of the experimental data-points P_{rc} represents the fraction of data found within ± 25% relative error range respectively.

In addition to the above suggested performance statistics, the rank correlation coefficient ρ^2, is proposed,

$$\rho^2 = (1 - \frac{6\sum_i^n (y_i^{pred} - y_i^{exp})^2}{n(n^2-1)})^2 \quad (16)$$

Similar to the R^2, the rank correlation coefficient ρ^2 measures the quality of the model fit. A value near unity is desired. An advantage of ρ^2 is that it is more suitable to assess monotonically increasing nonlinear functions which is the nature of ranked property[37].

The classical parameter estimation problem assumes that the error of the data is normally distributed. In addition to the above statistical performance indicators suggested by Hukkerikar et al.[20], different probability plots of the residual errors are considered to test if the underlying assumptions are valid:

1. Normal probability plot: Illustrates sequential departure from Gaussian normality, hence how closely the errors follow normal distribution.
2. Cauchy probability plot: Illustrates how well the errors follow a potential Cauchy distribution, which is better suited to describe residual distributions deviating from normal distribution due to in particular long tails (residuals distribution obtained from property prediction models mostly have long tails as observed in Hukkerikar et al.[20]). The Cauchy distribution is defined as in Eq. (17)[38],

$$F_{cauchy}(x) = \frac{1}{\pi(1+x^2)}. \quad (17)$$

Outlier treatment based on empirical cumulative distribution. The GC parameter estimation can be strongly influenced by outliers from the model structure. Although principles for their detection and deletion are well known, in property modeling literature it is uncommon to see an explicit account of a systematic treatment of outliers. In engineering applications usually a normal distribution of data is assumed to be followed and residuals beyond 2-3 standard deviations are considered to be outliers. Here outliers are detected based on the empirical cumulative distribution function (CDF) of the residuals between experimental and predicted values. This methodology was suggested by Frutiger et al.[39] for the identification of outliers in group contribution models, exemplified for the upper flammability limit UFL and compared to outlier detection based on Cook's distance and normal cumulative distribution.

The empirical CDF is a step function that increases by 1/n in every data point, where n is the number of data points. In this way, it seeks to estimate the true underlying distribution function of residuals and thereby improve the detection of outliers. It does not assume that residuals follow a normal distribution (or any other distribution function a priori), as e.g. the approach suggested by Ferguson[40]. This can be an advantage if the probability plots show great deviations from Gaussian normality. Data points that lie below the 2.5% or above the 97.5% probability levels which corresponds to 2-sigma deviation in normal distribution, are taken to be outliers.

Figure 4 shows the empirical CDF of the parameter estimation using ordinary non-linear least squares regression and Levenberg-Marquardt algorithm.

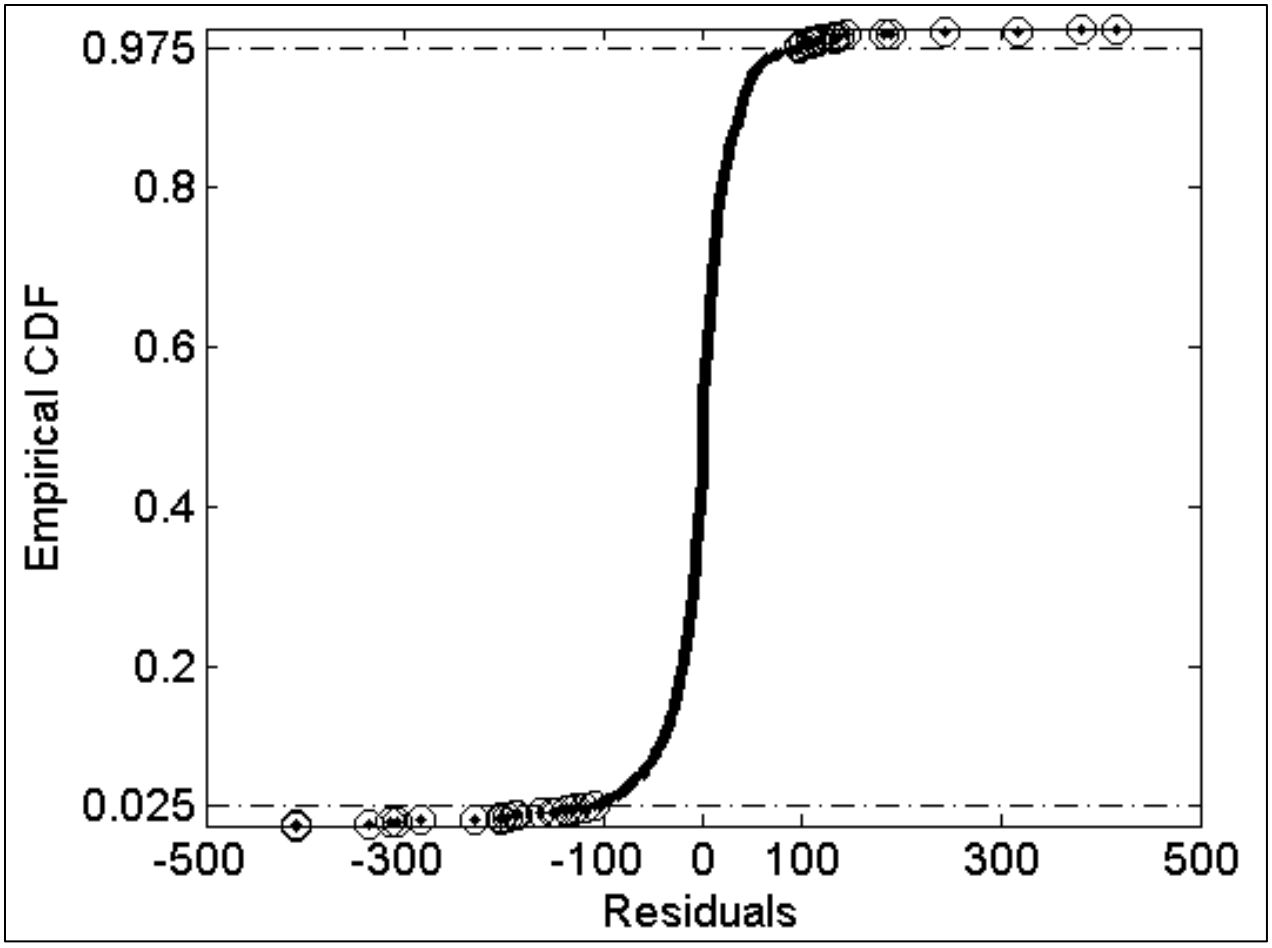

Figure 4. Empirical CDF of the residuals obtained from the parameter estimation using ordinary non-linear least squares regression and Levenberg-Marquardt algorithm. Below a probability of 0.025 and above 0.975 the data points are considered to be outliers.

UNCERTAINTY OF PARAMETER ESTIMATION AND PROPERTY PREDICTION

Uncertainty analysis based on linear error propagation using parameter covariance matrix. The underlying assumption of this method for uncertainty analysis method is that the measurement errors are ideally and independently distributed and defined by a Gaussian distribution white noise (normal distribution with zero mean and unit standard deviation).

The uncertainty of the parameter estimates is based on the asymptotic approximation of the covariance matrix, $COV(\theta^*)$ of parameter estimators[19,41]

$$COV(\theta^*) = \frac{SSE}{n-p}(J(\theta^*)^T J(\theta^*))^{-1} \tag{18}$$

In Eq.(18) *SSE* is the minimum sum of squared errors obtained from the least-squares parameter estimation method, *n* is the number of data points and *p* the number of parameters. The Jacobian *J* is the local sensitivity of the property model *f* with respect to the parameter values θ^*. The corresponding elements of the parameter correlation matrix can be obtained by

$$Corr(\theta_i^*, \theta_j^*) = \frac{COV(\theta_i^*, \theta_j^*)}{Var(\theta_i^*)Var(\theta_j^*)} \tag{19}$$

In Eq. (19) $COV(\theta_i^*, \theta_j^*)$ is the respective element of θ_i^* and θ_j^* of the covariance matrix and $Var(\theta_i^*)$ and $Var(\theta_j^*)$ are the variances of the respective parameters. The error on property predictions are estimated using linear error propagation in which the covariance matrix of the predictions $COV(y^{pred})$ is approximated using the Jacobian and the covariance of the parameter estimates as shown in Eq. (6),

$$COV(y^{pred}) = J(\theta^*)COV(\theta^*)J(\theta^*)^T \tag{20}$$

If the assumptions behind the model are satisfied (as ensured in previous steps) the parameter estimates will follow a student *t*-distribution, so

$$\theta_{1-\alpha}^* = \theta \pm \sqrt{diag(COV(\theta^*))} \cdot t(n-p, \alpha_t/2) \tag{21}$$

Similarly, the confidence intervals of the property predictions are given by:

$$y_{1-\alpha}^{pred} = y^{pred} \pm \sqrt{diag(J(\theta^*)COV(\theta^*)J(\theta^*)^T)} \cdot t(n-p, \alpha_t/2) \quad (22)$$

In Eq. (21) and (22) $t(n-p, \alpha_t/2)$ is the *t*-distribution value corresponding to the $\alpha_t/2$ percentile of Students t-distribution, $diag(COV(\theta^*))$ represents the diagonal elements of $COV(\theta^*)$ and $diag(J(\theta^*)COV(\theta^*)J(\theta^*)^T))$ the corresponding diagonal elements of $COV(J(\theta^*)COV(\theta^*)J(\theta^*)^T))$.

Uncertainty analysis based on bootstrap method. Using the parameter covariance matrix as described, assumes that the residuals are independent and follow normal distribution with zero mean[19]. However in practice this is rarely such (see e.g. the residual plots in Hukkerikar et al.[20]). The bootstrap method is an attempt to calculate the distributions of the errors from the data, and to use these to calculate the errors on the parameter estimation[42]. In a certain sense, the bootstrap method aims to relax the restriction to independent and identically distributed measurement errors, which is a central assumption in nonlinear least squares theory. In order to perform bootstrap method[42], first a reference parameter estimation is made, giving

$$\theta^* = arg\,min \quad \sum_i (y_i^{exp} - y_i^{pred}(\theta^*)^2) \quad (23)$$

The bootstrap method defines $\hat{F}$ as the sample probability distribution of the errors $\hat{\epsilon}$:

$$\hat{F} = mass\frac{1}{n} \quad at \quad \hat{\epsilon}_i = (y_i^{exp} - y_i^{pred}(\theta^*)) \quad (24)$$

From this the new set of errors can be obtained. The residuals are assumed to be uniformly distributed $\hat{F}$ i.e. each residual has equal probability of realization. In the next steps, new synthetic data sets are produced. The bootstrap method generates any number of synthetic data

sets *(y*(1); y*(2),…, y*(k))* also with *n* data points (*n* being here the total number of observations, and *k* being the total number of bootstrap samples) by using random sampling with replacement from the residuals $\hat{\epsilon}$. The procedure is simply to add the *n* bootstrap samples of residuals to the model predictions obtained using the estimated parameters in the reference step above as follows:

$$y_i^* = y_i^{pred}(\theta) + \hat{\epsilon}_i \quad where \quad \hat{\epsilon}_i \in \hat{F} \tag{25}$$

Parameter estimation is repeated using each synthesis data set *y*(k),* which results in a new set of estimated parameters $\theta^*(k)$ and a new predicted value of $y_i^{pred*}(k)$ solving the minimization problem as formulated above. The resulting sample of estimated parameter values are plotted to graphically visualize the uncertainty in the estimated parameter values. In addition, inference statistics can be used to estimate the mean μ_{θ^*} and standard deviation $\sigma_{\theta^*}$of the distribution of the estimated parameter values. The mean value and the standard deviation of all the estimated parameter sets can be used to calculate the confidence intervals:

$$\mu_{\theta^*} = \frac{1}{n}\sum_{k=1}^{n}\theta^*(k) \tag{26}$$

$$\sigma_{\theta^*} = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(\theta^*(k) - \mu_{\theta^*})^2} \tag{27}$$

In Eq. (26) and (27) *n* is the number of data points and $\theta^*(k)$ is the estimated parameter using the *k-th* synthetic data set.

Parameter identifiability. Parameter identifiability is a common issue in nonlinear regression[19] with important implications for model validation and application. Parameter identifiability is

basically the issue, can the model parameters be estimated *uniquely* from a certain data set? We use the following diagnostic measures to analyze parameter identifiability in GC models:

a) The parameter estimates *must not be linearly dependent*, so the linear correlation coefficients between parameter estimates should be sufficiently low, e.g. less than 0.7 [43,44], and

b) Parameter estimation errors (i.e. 95% confidence intervals) should be sufficiently low[45]. One obvious indication of poor parameter identifiability is a large confidence interval, e.g. relative parameter estimation error being larger than 50% [46,45].

RESULTS AND DISCUSSION

Regression models and practical global optimality of parameter estimation. The performance of the applied regression models for the ΔH_c° GC method is shown in table 1. The results are depicted before and after outlier deletion, where N_{out} is the number of outliers removed. R^2, ρ^2, *SD, ARE, SSE* and *AAD* are defined above. P_{rc} represents the percentage of the experimental data points found within ± 25% relative error range respectively [20]. Figure 5 shows the prediction of ΔH_c° versus the experimental value for Robust regression and Weighted least squares regression *after* outlier deletion.

Table 1. Regression model performance statistics, the best value of the respective column is highlighted.

	R^2 *Pearson*	ρ^2 *Spearman*	N_{out}	*SD*	*AAD*	*ARE [%]*	*SSE*	P_{rc} *25%*
Ordinary least-squares *before* outlier deletion	0.99	0.99	0	76.63	30.35	1.10	$4.66 \cdot 10^6$	99.75
Robust regression *before* outlier deletion	0.99	0.99	0	87.33	21.80	0.75	$3.29 \cdot 10^5$	99.62
Weighted least squares *before* outlier deletion	0.99	0.99	0	134.39	61.25	1.82	$6.89 \cdot 10^4$	99.62
Ordinary least-squares *after* outlier deletion	0.99	0.99	40	**22.14**	14.18	0.52	$3.70 \cdot 10^5$	**100**
Robust regression *after* outlier deletion	0.99	0.99	40	23.30	**13.09**	**0.50**	$1.42 \cdot 10^5$	99.87
Weighted least squares *after* outlier deletion	0.99	0.99	40	29.64	18.37	0.57	$\mathbf{2.84 \cdot 10^3}$	**100**

Outlier deletion improves the regression performance. After outlier deletion, the results are relatively close for the three models. The best fit according to *ARE* and *AAD* was achieved by robust regression after outlier deletion. However, for robust regression *SD* is slightly higher than

ordinary least squares and *SSE* is slightly higher than weighted-least squares and P_{rc} is slightly lower compared to both of them. The regression models performed an very good fit (see Figure 5).

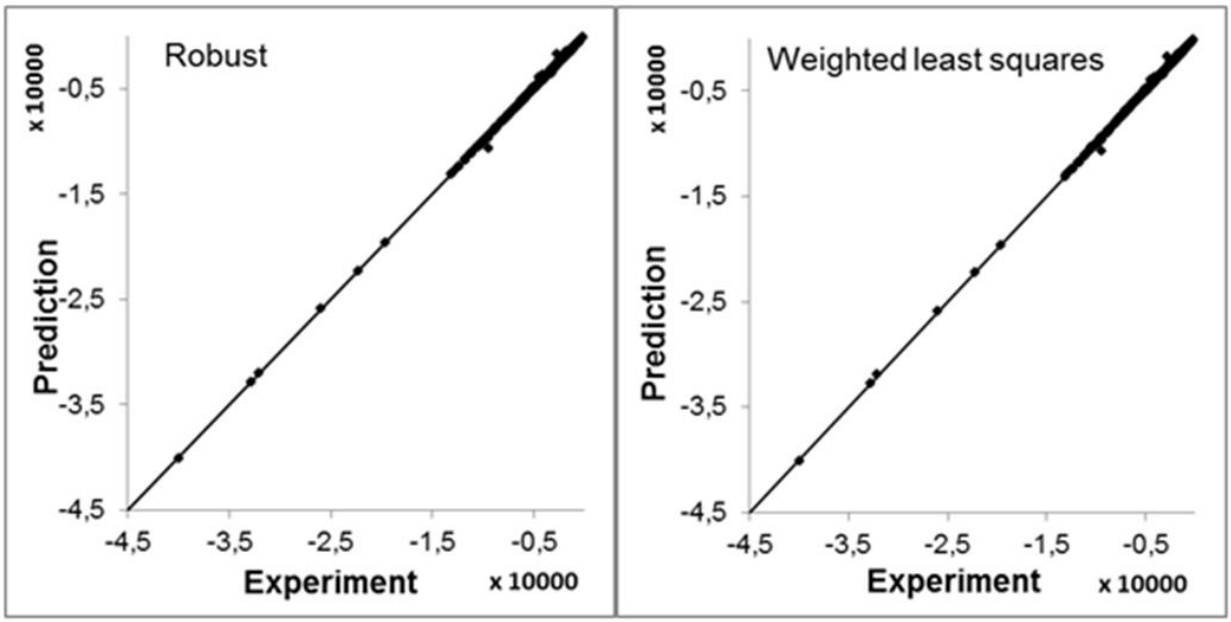

Figure 5. Prediction versus the experimental value of ΔH_c° after outlier removal for a) robust (left) and b) weighted least squares regression (right).

The reason why weighted least square performs slightly worse in terms of *SD*, *ARE* and *AAD* than robust regression can be explained as follows: The measurement error, which is the basis for the variance in the regression model, is given in percentage. Hence, large data points are often assigned a large variance and are therefore weighted less, such that the minimization of the residuals of large data points has a lower influence on the optimization. As a consequence, weighted least squares regression fits small property values much better than the large property values, whereas robust regression has no such bias. In that sense overall robust regression seems slightly favorable model for the GC parameter estimation of ΔH_c° property data.

Four separate search algorithms were used to cross-check and validate the global minimum of the solution. Table 2 shows the sum of squares errors *SSE* after the corresponding sequential and the simultaneous parameter estimation. A higher amount of parameters increases the goodness of the fit.

When comparing the final performance of the different optimization algorithms (see Table 2, final *SSE*), it can be seen that the Simplex and Trust-region reflective-algorithm lead to the best solutions, whereas *SSE* for pattern search algorithm and Levenberg-Marquart-algorithm was terminated at a higher *SSE* value. The solution found by the Simplex and Trust-region reflective-algorithms can be considered as *practically* (considering the four different search algorithms) globally optimal solution. The Levenberg-Marquart and Trust-region reflective-algorithm are strongly depending on the initial guess, since they are local search algorithms. The initial guess might have been suitable for Trust-region reflective, but not for Levenberg-Marquart. A possible explanation why pattern search did not find the same minimum as the others could be the nature of the search algorithm. It is known to be powerful for specific classes of functions[47].

Table 2. Sum of squares errors *SSE* of the parameter estimation for different optimization algorithms using sequential (sequ.) and simultaneous (sim.) estimation.

	SSE (sequ.) 1st order	*SSE (sequ.) 1st and 2nd order*	*SSE (sequ.) 1st, 2nd and 3rd order*	***Final SSE (sim.)***
Simplex	$9.65 \cdot 10^6$	$6.32 \cdot 10^5$	$3.33 \cdot 10^6$	$\mathbf{3.93 \cdot 10^5}$
Pattern search	$1.95 \cdot 10^7$	$1.69 \cdot 10^7$	$1.30 \cdot 10^7$	$\mathbf{1.26 \cdot 10^7}$
Levenberg-Marquart	$5.95 \cdot 10^6$	$5.39 \cdot 10^6$	$5.26 \cdot 10^6$	$\mathbf{4.66 \cdot 10^6}$
Trust-region reflective	$5.31 \cdot 10^5$	$4.82 \cdot 10^5$	$4.52 \cdot 10^5$	$\mathbf{3.70 \cdot 10^5}$

Uncertainty analysis property prediction errors. Figures 6 and 7 show the experimental and the predicted values of the heat of combustion with the respective 95%-confidence interval of the prediction for every substance both for covariance-based uncertainty analysis bootstrap sampling-based methods. As an example the prediction based on parameter estimates obtained using the robust regression is shown. The compounds are ordered from lowest to highest value and given an index number respectively. The confidence intervals are individual for each compound. The trend is a narrow band along with the experimental values.

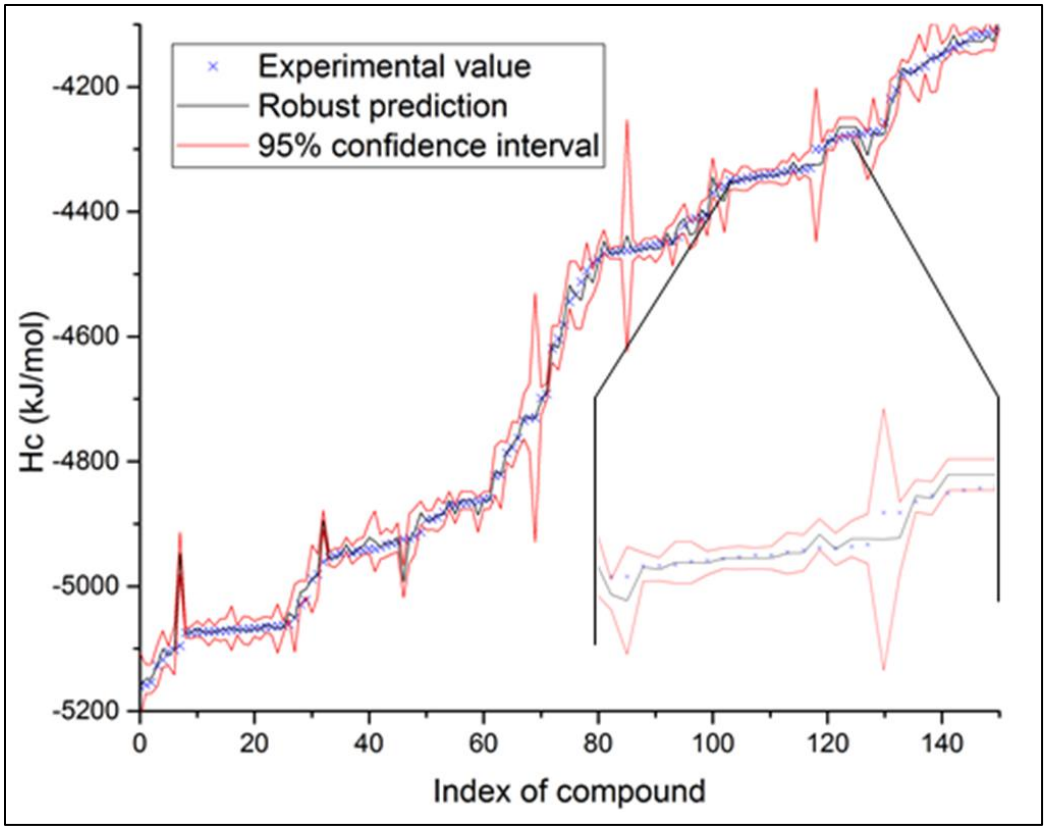

Figure 6. Experimental as well as predicted value of ΔH_c° for every compound with 95%-confidence intervals generated by covariance-based uncertainty analysis (robust regression without outliers). A section of the plot is enlarged to show the distribution of the experimental values around the prediction.

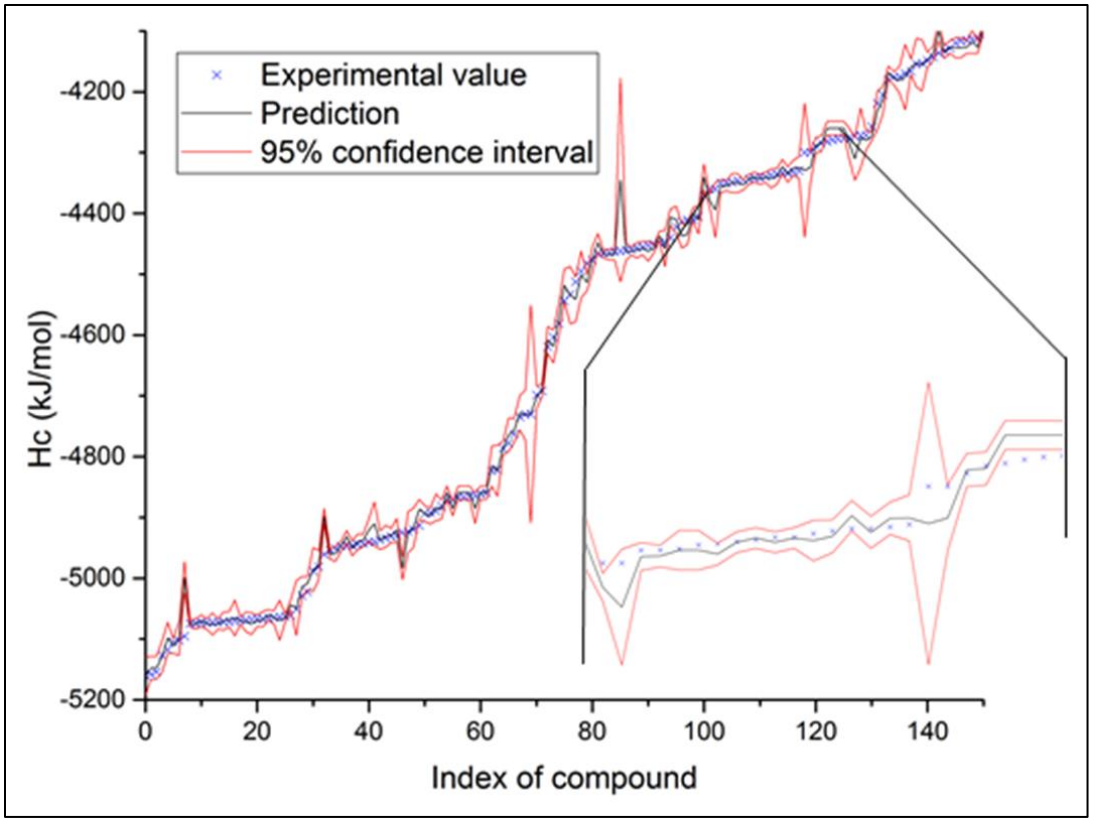

Figure 7. Experimental as well as predicted value of ΔH_c° for every compound with 95%-confidence intervals generated by bootstrap sampling-based uncertainty analysis.

Both methods (linear error propagation versus bootstrap) used for the calculation of the uncertainty of the prediction of the corresponding experimental value show a similar result, i.e. - in both methods the experimental value lies within the calculated 95% confidence intervals. Although bootstrap technique requires more model evaluations and computations compared to the linear error propagation (where only one model evaluation is needed), it has the advantage of being sampling- based, which allows non-linear error propagation.

Parameter identifiability analysis. The consideration of the 95%-confidence interval of the parameter estimates (see appendix), allows evaluating the practical identifiability of the GC factors. Although for all regression models the parameter fit was satisfying (see 3.1), there is a

large number of parameters that have a large confidence interval corresponding to a relative parameter estimation error $\sigma_{\theta^*}/\mu_{\theta^*}$ being larger than 50%. For the use of ordinary least-squares regression 96 out of 235 parameters are not practically identifiable, whereas for robust regression it is 95 out of 235. 83 out of 235 parameters fail practical identifiability for weighted least-squares. However, the universal parameter $\Delta H^{\circ}_{c\,const}$ is identifiable. Furthermore, almost all of the 1st order parameters (beside 3) could be identified practically compared to 2nd and 3rd order parameters where a larger part is not practically identifiable.

The practical identifiability depends on two main issues: The amount of data for the parameter estimation and the correlation between parameters.

If there is sufficient information (i.e. enough data points) to calculate the parameter estimates, the confidence interval gets smaller and hence, the parameters are practically identifiable. However, in GC parameter estimation there might be several functional groups that only occur in very few compounds. For some 3rd order parameters, there was only one compound available with a certain functional group. Hence, the 95%-confidence interval is very high and the parameters get non-identifiable

The second major source of parameter identifiability problems is high correlation (>0.7) between parameters, which can be observed in the parameter correlation matrix given in the supplementary material. The elements of the correlation matrix are directly linked to the covariance of two parameters, which is subsequently obtained from the Jacobian (see Eq. (18) and (19)). This means, if two parameters have a similar or identical sensitivity to the model output, they are highly correlated. In GC methods, correlation is intrinsically often the case, because certain functional groups can occur frequently together (depending on the data set)[48].

In many property modeling studies, practical identifiability of parameters has either not been considered or neglected. The diagnostic measures mentioned above indicate clearly that not all of the model parameters are uniquely identifiable. The first implication of this is that the estimated parameter values should not be attributed physical meaning since their values are not unique. Second, for practical application purposes, it is desirable to keep the parameters in the model, despite their identifiability issues, because in this way the application range of the GC model is higher (the more first, second and third order group contribution parameters in the model, the more chemicals property can be predicted).

However in that case, i.e. using a model with poorly identifiable parameters, the uncertainty of the prediction (i.e. perform propagation of parameter estimation errors to the property prediction) as shown in figures 6 and 7 becomes critical. The confidence interval of property prediction provides a measure of the prediction quality (accuracy) of the model developed, which the end user can use to judge if the prediction accuracy is fit for the intended application or else a more accurate measurement needs to be done instead of using a model prediction.

Effect of addition of higher order groups on property value and uncertainty. It is valuable to analyze, what the influence of correlated parameters is on the prediction and on the uncertainty of the prediction. The results obtained in this study showed that high correlation influences the mean prediction but not the uncertainty bounds (the upper and lower 95% confidence interval). In 155 out of 794 molecules the introduction of 2nd or 3rd order groups increased the relative error between experimental and predictive values for more than 10%. This particularity is exemplified and investigated by using two compounds namely cis,trans-2,4-Hexadiene and Acrolein. The parameter correlation matrix given in Table 3, shows that the GC factors of cis,trans-2,4-Hexadiene are highly correlated in comparison to the GC factors of Acrolein. The

prediction and 95%-confidence interval for the two selected substances are shown in Table 4 considers 1st order only, 1st and 2nd order as well as 1st, 2nd and 3rd order GC factors. These two examples shows that while adding more groups increases the relative error of prediction for cis,trans-2,4-Hexadiene compound (worse case), however it leads to a lower relative prediction error for Acrolein (better case). However, it does not affect the calculation of the 95%-confidence interval of the property prediction (reliable case). To understand this, we need to look back at the non-linear regression theory and parameter identifiability issues in detail.

Table 3. Parameter correlation matrices. The red color indicates a positive correlation of higher than 0.7 and the orange color indicates a negative correlation lower than -0.7.

GC-factors	$\Delta H^{\circ}_{c\,const}$	"CH3	'CH=CH'	'CHn=CHm-CHp=CHk'	'CH3-CHm=CHn'
$\Delta H^{\circ}_{c\,const}$	1.00				
'CH3'	-0.96	1.00			
'CH=CH'	-0.02	0.03	1.00		
'CHn=CHm-CHp=CHk'	0.02	-0.03	-0.94	1.00	
'CH3-CHm=CHn'	0.02	-0.07	-0.96	0.86	1.00

cis,trans-2,4-Hexadiene

GC-factors	$\Delta H^{\circ}_{c\,const}$	'CH2=CH'	'CHO'	'CHm=CHn-CHO'
$\Delta H^{\circ}_{c\,const}$	1.0			
'CH2=CH'	-0.45	1.0		
'CHO'	0.61	-0.26	1.0	
'CHm=CHn-CHO'	0.01	-0.55	0.30	1.0

Acrolein

Table 4. Prediction and 95%-confidence interval for a selection of substances comparing the usage of only 1st order GC-factors with the usage of 1st and 2nd as well as 1st, 2nd 3rd order groups.

	Relative error between prediction and experimental value			Boundary of 95%-confidence interval		
Used GC-factors	1st	1st, 2nd	1st, 2nd, 3rd	1st	1st, 2nd	1st, 2nd, 3rd
cis,trans-2,4-Hexadiene	0.024	0.038	0.029	±13.96	±13.93	±13.14
Acrolein	0.0094	0.0051	0.0051	±32.68	±32.67	±32.67

Table 5. Comparison of sample variance, s^2, as a function of increasing GC model parameters: comparison between a GC model containing only 1st order, 1st and 2nd as well as 1st, 2nd and 3rd order groups.

	Levenberg-Marquart algorithm		
Used GC-factors	1st	1st, 2nd	1st, 2nd, 3rd
SSE	531121	481983	452126
$n-p$	627	555	523
$S=\frac{SSE}{n-p}$	847	868	864

SSE is the sum of squared errors, *n* is the number of compounds for which experimental data is available and *p* the number of parameters.

In the case of Acrolein, most of the parameters are not significantly correlated and the relative error between experimental and predicted value as well as the 95%-confidence interval gets smaller by the introduction of 2nd order group. This outcome is observed for the majority (80%)

of the estimated compounds considered in this work. However, cis,trans-2,4-Hexadiene shows high correlation between the parameters, in particular negative correlation. The negative correlation between the universal constant and the 1st order parameters and 1st and 2nd as well as between 1st and 3rd order groups has an influence on the prediction. The relative error increases for cis,trans-2,4-Hexadiene by the introduction of the 2nd order group. However, the uncertainty (i.e. the 95%-confidence interval) is not enlarged by the introduction of higher order group. This particularity can be understood by looking at Eq. (18) and (20) (see above). The first reason lies in the negative correlation. If two parameters are negatively correlated and have similar sensitivity to the model output (corresponding to the Jacobian $J(\theta^*)$), their uncertainties will tend to cancel [48]. The second cause is the nature of the calculation of mean sum of squared error *S=SEE/(n-p)*. Table 5 shows that this normalization factor for the covariance matrix remains constant, because the relative decrease of SSE is compensated by the corresponding increase in the number of parameters used for its estimation.

As a result, one can conclude that definition and inclusion of higher groups for a GC model may not always lead to a more accurate property prediction. At least for some chemical compounds relative prediction error will become worse due to parameter identifiability issues. This can be for GC models that have a large amount of factors to ensure a brought applicability. However, the 95%-confidence interval does not enlarge due to poor parameter identifiability. We suggest therefore that developers and users of GC models in general always state the 95%-confidence interval, which includes information on the parameter correlation structure associated with poor parameter identifiability issues.

Comparison of different classes of compound classes. The average relative error *ARE*, the average absolute deviation and the number of compounds included for some selected classes of chemicals are shown in Table 6. The data is ordered according to the number of data points.

Table 7. Comparison of performance of different classes of chemicals.

Class	*ARE* (ΔH_c°) in %	*AAD* (ΔH_c°) in kJ/mol	No. of ΔH_c° compounds
Aromatic Compounds	0.18	10.97	104
Alkanes	0.14	7.09	103
Alkenes	0.24	8.70	65
Acids	1.04	17.29	60
Alcohols	0.41	12.70	56
Sulfur containing Compounds	0.46	11.55	44
Amines	0.70	17.91	37
Halogen containing Compounds	1.27	10.93	33
Ketones	0.52	14.72	30
Nitro-Compounds	0.51	11.13	26
Carboxylates	0.66	24.55	25
Esters	0.70	16.31	24
Ethers	0.49	13.43	20
Nitriles	0.92	13.12	18
Aldehydes	0.39	6.16	13
Pyridines	0.46	18.66	12

Overall all the major classes of chemicals, except for the Halogen containing compounds which has an ARE of 1.27%, have an ARE below 1%. In particular the model performs best for Alkanes, Aromatic Compounds and Alkenes with an *ARE* below 0.3%. This demonstrates the accuracy of the model over a great variety of chemical compounds. The classes not included in the Table 6 consist of 10 or less compounds and the corresponding results can be found in the supporting material.

Comparison of the new GC model with other property estimation models. The squared Pearson correlation coefficient R^2, average relative error *ARE*, the average absolute deviation and the number of data included of this study for the model using robust regression are compared to 5 other property prediction models in Table 6: Another group contribution (GC), quantitative structure-property relationship (QSPR), as well as artificial neural networks (ANN) for the calculation of ΔH_c°.

Table 7. Comparison of present model with existing models.

	Current study	Hshieh et al.[24] 2003	Gharageizi[25] 2008	Cao et al.[26] 2009	Pan et al.[27] 2011	Gharagheizi et al.[21], 2011
Model structure	**MG GC (robust reg.)**	Empirical Atomic Indices.	QSPR	QSPR with ANN	QSPR	ANN
R^2 Pearson	**0.99**	0.99	0.99	0.99	0.99	0.99
ARE (ΔH_c°) in %	**0.51**	3.90	3.45	-	-	0.16
AAD (ΔH_c°) in kJ/mol	**13.03**	-	-	155.32	104.13	-
No. of ΔH_c° data	**794**	75	1714	1496*	1650*	4590

*included experimental and predicted data hence it is biased.

Considering the average relative error *ARE* of ΔH_c°, the model developed in this study performs better than Hshieh et al.. Furthermore, the amount of data that is taken into account is much higher for the present model. This increases the application range of the model, since more substance from different classes of molecules have been used. In terms of *ARE* the model shows increased performance compared to Gharagheizi (2008), although the number of data points are lower. This is an indication that the parameter estimation methodology is very efficient. Cao et al. and Pan et al. have a higher absolute average error *AAE* than the new model. Furthermore, the amount of data consists of all experimental and predicted data available in the DIPPR database which is not a proper way to perform model development and performance statistics (which should solely be based on experimental data points only). The ANN model of Gharagheizi (2011) has a lower *ARE* and more data points. ANN is a fundamentally different approach to GC models. As regards the comparison of two different approaches for heat of combustion modelling, it is important to note that in ANN approach the aim is to build the best possible model structure (i.e. how many variables, descriptors, to include). In GC-based approach, the model structure is fixed. Therefore, the aim is instead on identifying and estimating in the best possible way the parameters of the fixed model given a certain available set of measurements. Therefore, the structure of the MG GC model is much simpler compared to ANN and much easier to work with and apply in industrial applications. Furthermore the reliability of the GC model predictions have been statistically demonstrated and verified against application in practice. However, establishing the reliability and confidence of parameter estimation in ANN remains to be demonstrated. Furthermore, due to the fact that the model is predefined, new experimental values can be added to the parameter estimation without changing the model

structure in GC models, while in QSPR and ANN model building need to be performed all over again.

CONCLUSION

In this study, a systematic methodology for the development, parameter estimation and uncertainty analysis of GC models was developed. The methodology was successfully applied for the development of new GC-based model with improved prediction performance statistics for the heat of combustion (ΔH_c°). In particular, the systematically developed new model has a higher accuracy than existing GC models and is much simpler to apply than ANN models.

The following are the main conclusions from the systematic development of GC-based models:

- Concerning the regression models, robust regression showed best performance statistics.
- The bootstrap method can be considered as a valid alternative to classical uncertainty analysis (linear approximation of covariance matrix of parameter estimators) when the underlying distribution of errors is considered to be unknown or not normally distributed.
- Although GC-based models have severe parameter identifiability issues characterized by significant correlation between estimated parameters and large confidence interval, the GC-based models still can be used successfully provided that 95% confidence interval of model predictions (prediction accuracy) are also calculated and reported.
- Addition of higher order groups (additional parameters) may in certain cases increase the prediction error, but does not enlarge the uncertainty (95%-confidence interval), due to parameter correlation associated with poor parameter identifiability.
- The use of different optimization algorithms for the parameter estimation is suggested as a simple method to ensure that the practically globally optimal solution was found.

GC-based property models are highly valuable and effective tools of property predictors. To ensure accurate and reliable estimation of properties of interest, comprehensive uncertainty analysis in particular 95% confidence interval of model predictions must be performed using systematic methods as presented in this work.

ASSOCIATED CONTENT

Supporting Information Available: Marrero/Gani group contribution factors and the universal constant from robust regression fit without outliers in tabular form. Example of model application. This material is available free of charge via the Internet at http://pubs.acs.org

AUTHOR INFORMATION

Corresponding Author

*Tel.: +45 45252806, E-mail address: gsi@kt.dtu.dk

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding Sources

This work was funded by the Innovation Fund Denmark under the THERMCYC project.

REFERENCES

(1) Gani, R.; O'Connell, J. P. Properties and CAPE: From Present Uses to Future Challenges. *Comput. Chem. Eng.* **2001**, *25*, 3–14.

(2) Marrero, J.; Gani, R. Group-Contribution Based Estimation of Pure Component Properties. *Fluid Phase Equilib.* **2001**, *183-184*, 183–208.

(3) Bünz, A. .; Braun, B.; Janowsky, R. Quantitative Structure–property Relationships and Neural Networks: Correlation and Prediction of Physical Properties of Pure Components and Mixtures from Molecular Structure. *Fluid Phase Equilib.* **1999**, *158-160*, 367–374.

(4) Peterson, K. a.; Feller, D.; Dixon, D. a. Chemical Accuracy in Ab Initio Thermochemistry and Spectroscopy: Current Strategies and Future Challenges. *Theor. Chem. Acc.* **2012**, *131*, 1–20.

(5) Joback, K. Estimation of Pure-Component Properties from Group-Contribution. *Chem. Eng. Commun.* **1987**, *57*, 233–243.

(6) Lydersen, A. L. Estimation of Critical Properties of Organic Compounds. *Coll. Eng. Univ. Wisconsin Eng. Exp. Stn. Rep. 3, Madison, WI* **1955**.

(7) Klincewicz, K. Estimation of Critical Properties with Group Contribution Methods. *AICHE J.* **1984**, *30*, 137–142

(8) Constantinou, L.; Gani, R. New Group Contribution Method for Estimating Properties of Pure Compounds. *AIChE J.* **1994**, *40*, 1697–1709.

(9) Hukkerikar, A. S.; Meier, R. J.; Sin, G.; Gani, R. A Method to Estimate the Enthalpy of Formation of Organic Compounds with Chemical Accuracy. *Fluid Phase Equilib.* **2013**, *348*, 23–32.

(10) Poling, B. E.; Prausnitz, J. M.; O'Connell, J. P. The Estimation of Physical Properties. In *The Properties of Gases and Liquids*; McGraw-Hill: New York, 2004; pp 1–9.

(11) Dong, Q.; Chirico, R. D.; Yan, X.; Hong, X.; Frenkel, M. Uncertainty Reporting for Experimental Thermodynamic Properties. *J. Chem. Eng. Data* **2005**, *50*, 546–550.

(12) Whiting, W. B. Effects of Uncertainties in Thermodynamic Data and Models on Process Calculations †. *J. Chem. Eng. Data* **1996**, *41*, 935–941.

(13) Larsen, A. H. Data Quality for Process Design. *Fluid Phase Equilib.* **1986**, *29*, 47–58.

(14) Mathias, P.; Klotz, H. Take a Closer Look at Thermodynamic Property Models. *Chem. Eng. Prog.* **1994**, *90*, 67–75.

(15) Hajipour, S.; Satyro, M. A. Uncertainty Analysis Applied to Thermodynamic Models and Process Design – 1. Pure Components. *Fluid Phase Equilib.* **2011**, *307*, 78–94.

(16) Maranas, C. Optimal Molecular Design under Property Prediction Uncertainty. *AICHE J.* **1997**, *43*, 1250–1264.

(17) Yan, X.; Dong, Q.; Hong, X. Reliability Analysis of Group-Contribution Methods in Predicting Critical Temperatures of Organic Compounds. *J. Chem. Eng. Data* **2003**, *48*, 374–380.

(18) Verevkin, S. P.; Emel'yanenko, V. N.; Diky, V.; Muzny, C. D.; Chirico, R. D.; Frenkel, M. New Group-Contribution Approach to Thermochemical Properties of Organic Compounds: Hydrocarbons and Oxygen-Containing Compounds. *J. Phys. Chem. Ref. Data* **2013**, *42*, 1–48.

(19) Seber, G.; Wild, C. *Nonlinear Regression*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1989.

(20) Hukkerikar, A. S.; Sarup, B.; Ten Kate, A.; Abildskov, J.; Sin, G.; Gani, R. Group-Contribution+ (GC+) Based Estimation of Properties of Pure Components: Improved Property Estimation and Uncertainty Analysis. *Fluid Phase Equilib.* **2012**, *321*, 25–43.
(21) Gharagheizi, F.; Mirkhani, S. A.; Tofangchi Mahyari, A. R. Prediction of Standard Enthalpy of Combustion of Pure Compounds Using a Very Accurate Group-Contribution-Based Method. *Energy Fuels* **2011**, *25*, 2651–2654.
(22) Cardozo, R. L. Prediction of the Enthalpy of Combustion of Organic Compounds. *AIChE J.* **1986**, *32*, 844–848.
(23) Seaton, W. H.; Harrison, B. K. A New General Method for Estimation of Heats of Combustion for Hazard Evaluation. *J. Loss Prev. Process Ind.* **1990**, *3*, 311–320.
(24) Hshieh, F. Y.; Hirsch, D. B.; Beeson, H. D. Predicting Heats of Combustion of Polymers Using an Empirical Approach. *Fire Mater.* **2003**, *27*, 9–17.
(25) Gharagheizi, F. A Simple Equation for Prediction of Net Heat of Combustion of Pure Chemicals. *Chemom. Intell. Lab. Syst.* **2008**, *91*, 177–180.
(26) Cao, H. Y.; Jiang, J. C.; Pan, Y.; Wang, R.; Cui, Y. Prediction of the Net Heat of Combustion of Organic Compounds Based on Atom-Type Electrotopological State Indices. *J. Loss Prev. Process Ind.* **2009**, *22*, 222–227.
(27) Pan, Y.; Jiang, J. C.; Wang, R.; Jiang, J. J. Predicting the Net Heat of Combustion of Organic Compounds from Molecular Structures Based on Ant Colony Optimization. *J. Loss Prev. Process Ind.* **2011**, *24*, 85–89.
(28) Project 801, Evaluated Process Design Data, Public Release Documentation, Design Institute for Physical Properties (DIPPR), American Institute of Chemical Engineers (AIChE), 2014.
(29) Pierotti, G. J.; Deal, C. H.; Derr, E. L. Activity Coefficients and Molecular Structure. *Ind. Eng. Chem.* **1959**, *51*, 95–102.
(30) Huber, P. J. Robust Estimation of a Location Parameter. *Ann. Math. Stat.* **1964**, *35*, 73–101.
(31) Wassermann, L. *All of Nonparametric Statistics*; Springer: Berlin, 2006.
(32) Marquardt, D. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *J. Soc. Ind. Appl. Math.* **1963**, *11*, 431–441.
(33) Coleman, T. F. On the Convergence of Interior-Reflective Newton Methods for Nonlinear Minimization Subject to Bounds ". *Math. Program.* **1994**, *67*, 189–224.
(34) Lagarias, J. C.; Reeds, J. A.; Wright, M. H.; Wright, P. E. Convergence Properties of the Nelder–mead Simplex Method in Low Dimensions. *SIAM J. Optim.* **1998**, *9*, 112–147.
(35) Audet, C.; Dennis, J. E. Analysis of Generalized Pattern Searches. *SIAM J. Optim.* **2003**, *13*, 889–903.
(36) Byrd, R. H.; Schnabel, R. B.; Shultz, G. A. A Trust Region Algorithm for Nonlinearly Constrained Optimization. *SIAM J. Numer. Anal.* **1987**, *24*, 1152–1170.
(37) Louangrath, P. I. Correlation Coefficient According to Data Classification. *SSRN Electron. J.* **2014**, 1–28.
(38) Ferguson, T. S. Maximum Likelihood Estimates of the Parameters of the Cauchy Distribution for Samples of Size 3 and 4. *J. Am. Stat. Assoc.* **1978**, *73*, 211–213.
(39) Frutiger, J.; Abildskov, J.; Sin, G. Outlier Treatment for Improving Parameter Estimation of Group Contribution Based Models for Upper Flammability Limit. In *12th International Symposium on Process Systems Engineering and 25th European Symposium on Computer Aided Process Engineering*; Gernaey, K. V., Huusom, J. K., Gani, R., Eds.; Copenhagen,

2015.
(40) Ferguson, T. S. On the Rejection of Outliers. *Proc. Berkeley Symp. Math. Stat. Probab.* **1961**, *1*, 253–287.
(41) Sin, G.; Gernaey, K. V.; Neumann, M. B.; van Loosdrecht, M. C. M.; Gujer, W. Global Sensitivity Analysis in Wastewater Treatment Plant Model Applications: Prioritizing Sources of Uncertainty. *Water Res.* **2011**, *45*, 639–651.
(42) Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1–26.
(43) Brun, R.; Kühni, M.; Siegrist, H.; Gujer, W.; Reichert, P. Practical Identifiability of ASM2d Parameters - Systematic Selection and Tuning of Parameter Subsets. *Water Res.* **2002**, *36*, 4113–4127.
(44) Sin, G.; Vanrolleghem, P. a. Extensions to Modeling Aerobic Carbon Degradation Using Combined Respirometric-Titrimetric Measurements in View of Activated Sludge Model Calibration. *Water Res.* **2007**, *41*, 3345–3358.
(45) Homberg, A. On the Practical Identifiability of Microbial Growth Models Incorporating Michaelis-Menten Type Nonlinearities. *Math. Biosci.* **1982**, *62*, 23–43.
(46) Baltes, M.; Schneider, R.; Reuss, M. Optimal Experimental Design for Parameter Estimation in Unstructured Growth Models. *Biotechnolgoy Prog.* **1994**, *10*, 480–488.
(47) Powell, M. J. D. On Search Directions for Minimization Algorithms. *Math. Program. 4* **1973**, *4*, 193–201.
(48) Kirchner, J. W. *Uncertainty Analysis and Error Propagation*; University of California, Berkeley, 2001.