Group-contribution+ (GC+) based estimation of properties of pure components: Improved property estimation and uncertainty analysis

Amol Shivajirao Hukkerikar[a], Bent Sarup[b], Antoon Ten Kate[c], Jens Abildskov[a], Gürkan Sin[a], Rafiqul Gani[a,*]

[a] *Computer Aided Process-Product Engineering Center (CAPEC), Department of Chemical and Biochemical Engineering, Technical University of Denmark, DK-2800 kgs. Lyngby, Denmark*
[b] *Vegetable Oil Technology Business Unit, Alfa Laval Copenhagen A/S, Maskinvej 5, DK-2860 Soeborg, Denmark*
[c] *Akzonobel Research, Development and Innovation, Expert Capability Group – Process Technology, Velperweg 76, 6824 BM Arnhem, The Netherlands*

ABSTRACT

The aim of this work is to present revised and improved model parameters for group-contribution+ (GC+) models (combined group-contribution (GC) method and atom connectivity index (CI) method) employed for the estimation of pure component properties, together with covariance matrices to quantify uncertainties in the estimated property values. For this purpose, a systematic methodology for property modeling and uncertainty analysis of GC models and CI models using maximum-likelihood estimation theory is developed. For parameter estimation, large data-sets of experimentally measured property values of pure components of various classes (hydrocarbons, oxygenated components, nitrogenated components, poly-functional components, etc.) taken from the CAPEC database are used. In total 18 pure component properties are analyzed, namely normal boiling point, critical temperature, critical pressure, critical volume, normal melting point, standard Gibbs energy of formation, standard enthalpy of formation, normal enthalpy of fusion, enthalpy of vaporization at 298 K, enthalpy of vaporization at the normal boiling point, entropy of vaporization at the normal boiling point, flash point, auto ignition temperature, Hansen solubility parameters, Hildebrand solubility parameter, octanol/water partition coefficient, acentric factor, and liquid molar volume at 298 K. Important issues related to property modeling such as reliability and predictive capability of the property prediction models, and thermodynamic consistency of the predicted properties (such as, relation of normal boiling point versus critical temperature) are also analyzed and discussed. The developed methodology is simple, yet sound and effective and provides not only the estimated pure component property values but also the uncertainties (e.g. prediction errors in terms of 95% confidence intervals) in the estimated property values. This feature allows one to evaluate the effects of these uncertainties on product-process design, simulation and optimization calculations, contributing to better-informed and more reliable engineering solutions.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Physical and thermodynamic property data and property prediction models for pure components are vital pre-requisites for performing tasks such as, process design, simulation and optimization, and, computer aided molecular/mixture (product) design. While use of experimentally measured values of the needed pure component properties is desirable in these tasks, the experimental data of the properties of interest may not be available in many cases. Also, property prediction models employed in computer aided molecular/mixture design need to be predictive. In such cases, group-contribution (GC) methods such as those reported by Joback and Reid [1], Lydersen [2], Klincewicz and Reid [3], Constantinou and Gani [4], and Marrero and Gani [5] are generally suitable to obtain the needed property values since these methods provide the advantage of quick estimates without requiring substantial computational work. In GC methods, the property of a component is a function of structurally dependent parameters which are determined as a function of the frequency of the groups representing the molecules and their contributions. The application range and reliability of these methods depends on a number of factors:

- the group definitions used to represent the molecular structure of the pure components,
- the property model,

---

* Corresponding author. Tel.: +45 45252882; fax: +45 45932906.
 E-mail address: rag@kt.dtu.dk (R. Gani).

- the quantity and quality (information-wise) of the experimental data-set used in the regression to estimate the model parameters.

One other challenge with the GC methods is that the selected property model may not have all the needed parameters, such as groups and/or their contributions for a specific property. For such special cases, where the molecular structure of a given component is not completely described by any of the available groups, atom connectivity index (CI) method is employed together with the GC method to create the missing groups and to predict their contributions [6]. This combined approach leads to the development of a group-contribution⁺ (GC⁺) method of wider application range than before since the missing groups and their contributions can now be easily predicted through the regressed contributions of connectivity indices.

Currently available GC methods, such as those listed above, provide estimates of properties of pure components and the performance of these methods can be evaluated in terms of statistical performance indicators such as coefficient of determination (R^2), standard deviation (SD), average absolute error (AAE), and average relative error (ARE). However, for assessing the quality and reliability of the selected property prediction method, it is also necessary to know the uncertainties in the estimated property values. With this information, it is possible to perform better-informed design and simulation calculations by taking into account these uncertainties. Several studies in literature [7–10] have reported the impact of uncertainties in physical and thermodynamic property data and in property model parameters on the design and operation of many unit operations such as distillation, liquid–liquid extraction, flash process etc. Recently, Hajipour and Satyro [11] have discussed uncertainty analysis of models for critical constants and acentric factor using maximum-likelihood estimation and evaluated the effects of physical property uncertainties in process simulation using Monte Carlo technique. Maranas [12] illustrated the effects of uncertainties in property estimates on the optimization calculations involved in computer aided molecular design studies.

Hence given the importance of accurate and reliable property prediction and uncertainties of property estimates in the engineering design and operation tasks, this study aims to revise and improve GC⁺ models for estimation of pure component properties together with the confidence intervals of estimated property values. To this end, a systematic methodology for property modeling and uncertainty analysis is developed and its performance analyzed. For property modeling with a GC method, the Marrero and Gani (MG) method has been considered [5,13–15]. The MG method allows estimation of properties of pure components based exclusively on the molecular structure of the component and exhibits a good accuracy and a wide range of applicability covering chemical, biochemical and environmental-related components. For property modeling with a CI method, the models proposed by Gani et al. [6] have been considered. Since the publication of the original work on the property modeling by Gani and co-workers [5,6,13–15], significant amount of new experimental data of pure components (especially poly-functional, polycyclic, and complex components) has been added to the CAPEC database [16] developed and extended at CAPEC-DTU. The extended CAPEC database of pure component properties is used to develop new model parameters of the GC⁺ property models with the objective of providing more accurate and reliable estimation of pure component properties. In addition, the application range of the analyzed property models is increased by estimating the model parameters (group/atom contributions) whose values were not determined and published in the previous works due to the lack of necessary experimental data. In this work, two pure component properties – acentric factor and liquid molar volume at 298 K which were modelled earlier by Constantinou and Gani [4] are modelled using MG method in order to provide improved property estimations for a wide range of pure components. Finally, an uncertainty analysis has been added to the model parameter estimation so that the performance of the models can be evaluated also based on the confidence interval of their predictions giving additional insights into quality and credibility of pure component property estimation.

The following 18 pure component properties were considered for the analysis: normal boiling point (T_b), critical temperature (T_c), critical pressure (P_c), critical volume (V_c), normal melting point (T_m), standard Gibbs energy of formation (G_f), standard enthalpy of formation (H_f), normal enthalpy of fusion (H_{fus}), enthalpy of vaporization at 298 K (H_v), enthalpy of vaporization at the normal boiling point (H_{vb}), entropy of vaporization at the normal boiling point (S_{vb}), flash point (F_p), auto ignition temperature (Ait), Hansen solubility parameters (δ_D, δ_P, and δ_H), Hildebrand solubility parameter (δ), octanol/water partition coefficient ($Logk_{ow}$), acentric factor (ω), and liquid molar volume at 298 K (V_m).

The paper first gives a brief overview of the methods and tools employed in property modeling and uncertainty analysis; followed by the results of parameter estimation and uncertainty analysis; model performance and model fits; and finally a detailed discussion on reliability, predictive capability and thermodynamic consistency of the property prediction models. Tables containing list of model parameters together with parameter values and other related information, due to their large size, are provided as supplementary material.

2. Methods and tools for property modeling and uncertainty analysis

2.1. Methodology

As illustrated in Fig. 1, the developed methodology for property modeling and uncertainty analysis includes:

- a parameter estimation step to develop new and improved model parameters (group/atom contributions, adjustable parameters, and universal parameter),
- an uncertainty analysis step to establish statistical information about the quality of parameter estimation, such as the parameter covariance, the standard errors in predicted properties, and the confidence intervals.

The methodology is applied for property modeling and uncertainty analysis of 18 pure component properties listed above. The optimization algorithm used for the parameter estimation is the Levenberg–Marquardt technique [17] and this algorithm was implemented in MatLab (The Mathworks, Natick, Massachusetts). The experimental data of pure component properties used in the regression is taken from the CAPEC database [16].

2.2. MG group-contribution method

The property prediction model to estimate the properties of pure components employing MG method [5] has the form,

$$f(X) = \sum_i N_i C_i + w \sum_j M_j D_j + z \sum_k E_k O_k \qquad (1)$$

The function $f(X)$ is a function of property X and it may contain additional adjustable model parameters (universal constants) depending on the property involved. In Eq. (1), C_i is the contribution of the first-order group of type-i that occurs N_i times. D_j is the contribution of the second-order group of type-j that occurs M_j times. E_k is the contribution of the third-order group of type-k that has O_k occurrences in a component. Note that, Eq. (1) is a general model for
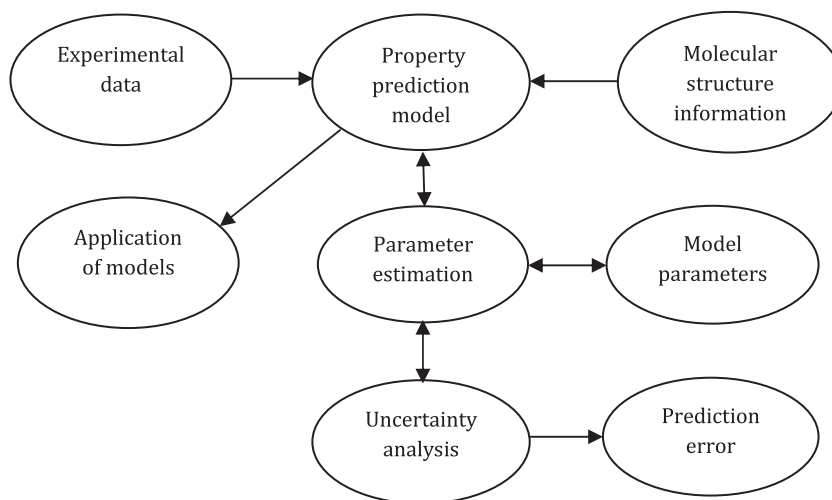
Fig. 1. Methodology for property modeling and uncertainty analysis.

all the properties and the definition of $f(X)$ is specific for each property X (see Table 2 for definition of $f(X)$ for each property model). For determination of the contributions, C_i, D_j, and E_k, a three-step regression procedure is followed.

Step 1: In this step, the constants w and z are assigned zero values because only contributions of the first-order groups are estimated, that is, the first-order groups, C_i and the additional adjustable parameters of the model.

$$f(X) = \sum_i N_i C_i \qquad (2)$$

Step 2: In this step, the constants w and z are assigned unity and zero values, respectively, because only first and second-order groups are considered. The regression is performed (by keeping fixed the C_i and the adjustable parameters obtained from step 1) to determine the contributions of the second-order groups, D_j.

$$f(X) = \sum_i N_i C_i + w \sum_j M_j D_j \qquad (3)$$

Step 3: In this step, both w and z are set to unity and regression is performed (by keeping fixed the obtained C_i, D_j, and the adjustable parameters obtained from steps 1 and 2) to determine the contributions of the third-order groups, E_k (see Eq. (1)).

In this way, the contributions of the higher levels act as corrections to the approximations of the lower levels. While this step-wise regression has advantages of a more flexible model, the disadvantage is that higher-order terms absorb the errors of the lower-order terms. As an alternative to the step-wise regression method, a simultaneous regression method is developed in which regression is performed by considering all the terms of Eq. (1) to obtain contributions of first, second and third-order groups in a single regression step.

2.3. Atom connectivity index (CI) method

In this method, the following pure-component property model is employed [6]:

$$f(X) = \sum_i a_i A_i + b(^v\chi^0) + 2c(^v\chi^1) + d \qquad (4)$$

where a_i is the contribution of the atom of type-i that occurs A_i times in the molecular structure, $^v\chi^0$ is the zeroth-order (atom

connectivity index, $^v\chi^1$ is the first-order (bond) connectivity index, b and c are adjustable parameters, and d is a universal parameter. Please note that $f(X)$ of the models in the MG method and in the CI method, (i.e. left hand side of Eqs. (1) and (4)) has the same functional form for a particular pure component property X and the values of universal constants for the CI models are the same as those for the GC models.

2.4. Database

For estimation of property model parameters, large experimental data-sets of pure components of various classes (hydrocarbons, oxygenated components, nitrogenated components, poly-functional components, etc.) taken from the CAPEC database [16] are used. The details of data-set of each property in terms of number of components belonging to various classes are given in Table 1.

2.5. Parameter estimation and uncertainty analysis (maximum-likelihood estimation theory)

For notational convenience, the property prediction model is represented by f and the model parameters (group/atom contributions, adjustable parameters, and universal parameter) by $\boldsymbol{P}$. The minimization of a cost function, $S(\boldsymbol{P})$, which is the sum of the squares of the difference between the experimental value, X^{exp} and evaluated property value, X^{pred} yields the values of unknown parameters $\boldsymbol{P}^*$. Note that $\boldsymbol{P}^*$ is the set of model parameter values obtained at the minimum value of the cost function value.

$$S(\boldsymbol{P}) = \min \sum_{j=1}^{N} (X_j^{exp} - X_j^{pred})^2 \qquad (5)$$

The subscript j indicates the pure component evaluated and N is the total number of pure components included in the evaluation. After estimation of the model parameters, uncertainty analysis is performed to quantify the model prediction errors as explained below.

For non-linear least-squares, the covariance matrix of the estimated model parameters can be approximated from a linear approximation of the model [18]. The linear approximation of the

Table 1
Class-wise description of the data-sets used for the regression purpose.

Class of pure components	T_b	T_c	P_c	V_c	T_m	G_f	H_f	H_{fus}	$LogK_{ow}$	F_p	δ_D	δ_P	δ_H	H_v	H_{vb}	S_{vb}	δ	Ait	ω	V_m
Hydrocarbons	662	277	288	281	492	279	272	270	233	168	73	57	64	189	112	112	386	191	430	324
Oxygenated	1187	297	314	280	1493	266	245	237	1500	229	443	444	444	229	185	185	498	243	659	352
Nitogenated	369	91	86	76	374	72	149	73	785	42	76	76	75	91	67	67	125	45	140	76
Chlorinated	202	38	30	29	149	29	28	18	271	21	75	74	74	34	30	30	62	28	65	49
Fluorinated	46	21	11	8	41	6	5	10	27	–	18	15	12	15	17	17	26	3	41	29
Brominated	99	8	8	8	89	8	10	11	41	5	29	29	29	23	12	12	20	6	19	15
Iodinated	30	4	5	5	27	5	5	2	10	–	9	9	9	9	7	7	8	–	8	8
Phosphorous containing	2	2	–	1	2	–	1	1	2	–	–	–	–	–	–	–	1	–	1	1
Sulfonated	109	34	33	32	83	31	29	36	65	2	23	23	23	51	39	39	35	2	73	45
Silicon containing	14	2	2	–	4	–	–	–	6	2	–	–	–	–	–	–	4	4	2	2
Multifunctional	790	84	75	77	2429	53	138	103	9253	43	291	290	286	64	43	43	219	48	285	155
Total number of components	3510	858	852	797	5183	749	882	761	12193	512	1037	1017	1016	705	512	512	1384	570	1723	1056

covariance matrix, $COV(\boldsymbol{P}^*)$, for the estimated parameters is given by,

$$COV(\boldsymbol{P}^*) = \frac{SSE}{\nu}(J(\boldsymbol{P}^*)^T J(\boldsymbol{P}^*))^{-1} \tag{6}$$

where SSE is the minimum sum of squared errors obtained from the least-squares parameter estimation method, ν is the degree of freedom (that is, the total number of measurements, n minus the number of unknown parameters, m). The Jacobian matrix $J(\boldsymbol{P}^*)$ calculated using $\partial f/\partial \boldsymbol{P}^*$ represents the local sensitivity of the property model f to variations in the estimated parameter values $\boldsymbol{P}^*$. The covariance matrix computed using Eq. (6) is used for assessing the quality of parameter estimation. The diagonal elements of this matrix are the variances of the errors of the parameter estimates and the off-diagonal elements are the covariances between the parameter estimation errors.

For linear least squares, the covariance matrix of the estimated model parameters is given by,

$$COV(\boldsymbol{P}^*) = \frac{SSE}{\nu}(\boldsymbol{A}^T \boldsymbol{A})^{-1} \tag{7}$$

For the GC-model with linear form of $f(X)$, $\boldsymbol{A}$ is the matrix containing frequencies of groups used to represent the components in the data-set used for the regression. For the CI model with linear form of $f(X)$, $\boldsymbol{A}$ is the matrix containing frequencies of atoms, and zeroth-order and first-order connectivity index for each component included in the data-set.

The error on the estimated property values can be calculated via linear error propagation [18] as follows:

$$COV(X^{pred}) = (J(\boldsymbol{P}^*)COV(\boldsymbol{P}^*)J(\boldsymbol{P}^*)^T) \tag{8}$$

The confidence interval of the parameters, $\boldsymbol{P}^*$, at α_t significance level is given as,

$$\boldsymbol{P}^*_{1-\alpha_t} = \boldsymbol{P}^* \pm \sqrt{diag(COV(\boldsymbol{P}^*))}. \quad t(\nu, \alpha_t/2) \tag{9}$$

In Eq. (9), $t(\nu, \alpha_t/2)$ is the t-distribution value corresponding to the $\alpha_t/2$ percentile (α_t is usually a value of 0.05) and $diag(COV(\boldsymbol{P}^*))$ represents the diagonal elements of $COV(\boldsymbol{P}^*)$. The t-distribution value is obtained from the probability distribution function of Students t-distribution [19], $P_r(t, \nu)$, and is given as,

$$0 = \sqrt{\nu}B\left(\frac{1}{2}, \frac{\nu}{2}\right)^{-1} \int_{-t}^{t}(1 + x^2/\nu)^{-1/2(\nu+1)}dx - P_r(t, \nu) \tag{10}$$

where $x = \nu/(\nu+t^2)$ and $B(1/2, \nu/2)$ is the beta function. For 95% confidence interval calculation, the value of $P_r(t, \nu)$ is 0.95. The t-distribution value can also be obtained using the "tinv function command" available in MatLab.

The confidence interval of the predicted property value, X^{pred}, at α_t significance level is given as,

$$X^{pred}_{1-\alpha_t} = X^{pred} \pm \sqrt{diag(J(\boldsymbol{P}^*)COV(\boldsymbol{P}^*)J(\boldsymbol{P}^*)^T)} \cdot t(\nu, \alpha_t/2) \tag{11}$$

The $X^{pred}_{1-\alpha_t}$ calculated from Eq. (11) provides the confidence interval of the predicted property value at a specified confidence level (usually at 95%) and it can be used to assess the reliability of the prediction. For instance, when experimental data is available for the property of a given component, one can compare the measurement with the confidence interval. If the experimental value of the property is within the calculated confidence interval, then one can conclude that the predicted property value (and hence the prediction method) is reliable. For the case when no experimental data is available, the calculated confidence interval provides a measure of the likely prediction error (hence an uncertainty) of the predicted property value (the larger the confidence interval, the higher the

Table 2
Performance of MG method based property models analyzed using step-wise regression method.
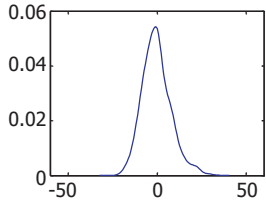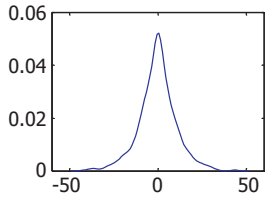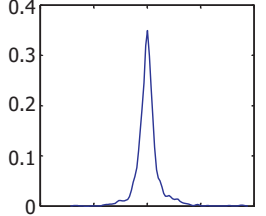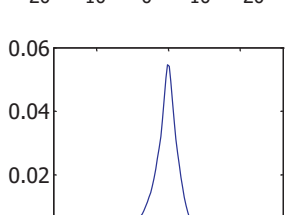
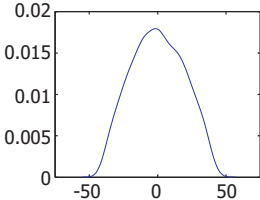| Sl. No.	Property	$f(X)$	N	ν	R^2	Residual distribution plot	P_{rc} ($\pm 1\%$)	P_{rc} ($\pm 5\%$)	P_{rc} ($\pm 10\%$)	SD	AAE	ARE [a]	AE_{max}	RE_{max}
						MG group-contribution model $f(X) = \sum_i N_i C_i + w \sum_j M_j D_j + z \sum_k E_k O_k$								
1	Normal boiling point [K]	$\exp\left(\frac{T_b}{T_{bo}}\right)$	3510	3179	0.9980		56.95	99.44	100.0	7.90	6.17	1.44	35.88	11.46
2	Critical temperature [K]	$\exp\left(\frac{T_c}{T_{co}}\right)$	858	607	0.9983		66.03	98.87	100.0	10.77	7.72	1.23	44.34	7.21
3	Critical pressure [bar]	$(P_c - P_{c1})^{-0.5} - P_{c2}$	852	608	0.9690		31.87	80.18	92.34	2.38	1.40	3.90	17.87	25.79
4	Critical volume [cc/mol]	$V_c - V_{c0}$	797	565	0.9956		50.31	93.22	99.00	11.65	7.97	2.05	61.72	21.73

Table 2 (*Continued*)

Sl. No.	Property	$f(X)$	N	ν	R^2	Residual distribution plot	MG group-contribution model $f(X) = \sum_i N_i C_i + w \sum_j M_j D_j + z \sum_k E_k O_k$							
							P_{rc} ($\pm 1\%$)	P_{rc} ($\pm 5\%$)	P_{rc} ($\pm 10\%$)	SD	AAE	ARE [a]	AE_{max}	RE_{max}
5	Normal melting point [K]	$\exp\left(\frac{T_m}{T_{mo}}\right)$	5183	4803	0.9456		17.60	73.28	93.90	19.16	15.99	5.07	44.43	29.59
6	Gibbs free energy [kJ/mol]	$G_f - G_{f0}$	749	521	0.9984		37.12	72.23	85.31	8.36	5.24	–	40.50	–
7	Enthalpy of formation [kJ/mol]	$H_f - H_{f0}$	882	649	0.9992		42.86	80.05	90.02	7.74	5.03	–	45.02	–
8	Enthalpy of fusion [kJ/mol]	$H_{fus} - H_{fus0}$	764	516	0.8324		13.48	32.98	53.40	5.16	2.79	–	68.01	–

Table 2 (*Continued*)

Sl. No.	Property	$f(X)$	N	ν	R^2	Residual distribution plot	MG group-contribution model $f(X) = \sum_i N_i C_i + w \sum_j M_j D_j + z \sum_k E_k O_k$							
							P_{rc} (±1%)	P_{rc} (±5%)	P_{rc} (±10%)	SD	AAE	ARE [a]	AE_{max}	RE_{max}
9	Octanol/water partition coefficient	$LogK_{ow} - K_{ow0}$	12193	11817	0.874		5.45	25.44	43.85	0.64	0.48	–	4.35	–
10	Flash point [K]	$F_p - F_{po}$	512	340	0.9671		37.50	89.65	98.44	12.10	8.97	2.8	56.79	12.86
11	Hansen solubility parameter [MPa$^{1/2}$]	Dispersion (δ_D)	1037	769	0.73	(shown here for δ_D parameter)	39.05	86.89	96.53	1.08	0.60	–	9.52	–
		Polar (δ_P)	1017	754	0.75		9.34	25.66	46.02	2.20	1.81	–	9.68	–
		H$_2$-bond (δ_H)	1016	754	0.87		12.99	34.06	56.00	2.79	1.28	–	9.24	–

Table 2 (*Continued*)

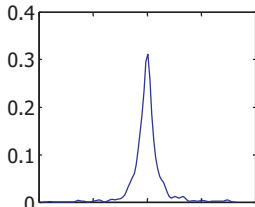| Sl. No.	Property	$f(X)$	N	ν	R^2	Residual distribution plot	MG group-contribution model $f(X) = \sum_i N_i C_i + w \sum_j M_j D_j + z \sum_k E_k O_k$							
							P_{rc} (±1%)	P_{rc} (±5%)	P_{rc} (±10%)	SD	AAE	ARE [a]	AE_{max}	RE_{max}
12	Enthalpy of vaporization (298 K) [kJ/mol]	$H_v - H_{v0}$	705	509	0.9716		40.99	90.07	96.88	2.34	1.29	3.24	26.92	133.96
13	Enthalpy of vaporization (T_b) [kJ/mol]	$H_{vb} - H_{vb0}$	512	346	0.9606		41.41	90.04	98.05	1.42	0.95	2.66	8.16	22.21
14	Entropy of vaporization (T_b) [J/mol K]	$S_{vb} - S_{vb0}$	512	346	0.8539		58.20	95.90	98.63	3.0	1.72	1.84	17.98	19.16
15	Hildebrand solubility parameter [MPa$^{1/2}$]	$\delta - \delta_0$	1384	1089	0.8290		34.25	70.45	90.46	1.63	1.08	5.61	12.17	44.49

Table 2 (*Continued*)

Sl. No.	Property						MG group-contribution model $f(X) = \sum_i N_i C_i + w \sum_j M_j D_j + z \sum_k E_k O_k$							
		$f(X)$	N	ν	R^2	Residual distribution plot	$P_{rc}(\pm 1\%)$	$P_{rc}(\pm 5\%)$	$P_{rc}(\pm 10\%)$	SD	AAE	ARE [a]	AE_{max}	RE_{max}
16	Auto ignition temperature[b] [K]	T_{AiT}				This property is modeled using simultaneous regression method only and hence model performance statistics for this property are provided in Table S1 (given in the supplementary material).								
17	Acentric factor	$exp\left(\frac{\omega}{\omega_a}\right)^{\omega_b} - \omega_c$	1723	1422	0.8921		41.56	56.36	78.58	0.1002	0.0534	11.09	1.68	77.55
18	Liquid molar volume [cc/kmol]	$V_m - V_{m0}$	1056	800	0.9967		69.60	92.90	98.77	0.0045	0.0024	2.03	0.0401	37.05

[a] For G_f, H_f, H_{fus}, and $LogK_{ow}$, ARE is not reported since these properties contain both positive and negative values. For Hansen solubility parameters δ_D, δ_P, and δ_H, ARE is not reported as these properties contain very small experimental values.

[b] For auto ignition temperature, the right hand side of the MG model is different and it is given as, $T_{Ait} = Ait_1 \, 10^{\left(-\sum N_i C_{ai} + \sum M_j D_{aj} + \sum E_k O_{ak}\right)} + Ait_2 + \sum N_i C_{bi} + \sum M_j D_{bj} + \sum E_k O_{bk}$.

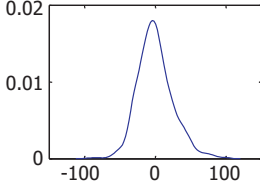**Table 3**
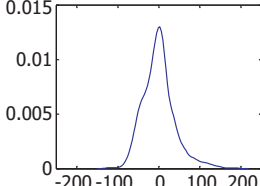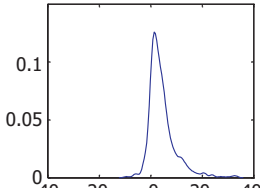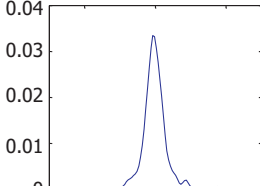Performance of CI method based property models.

Sl. No.	Property	$f(X)$	N	v	R^2	Residual distribution plot	CI method based property prediction model $f(X) = \sum_i a_i A_i + b(^v\chi^0) + 2c(^v\chi^1) + d$					
							P_{rc} (±1%)	P_{rc} (±5%)	P_{rc} (±10%)	SD	AAE	ARE [a]
1	Normal boiling point [K]	$\exp\left(\frac{T_b}{T_{bo}}\right)$	3510	3496	0.8865		23.85	76.10	95.75	25.54	19.51	4.48
2	Critical temperature [K]	$\exp\left(\frac{T_c}{T_{co}}\right)$	858	844	0.8592		27.74	72.26	93.47	39.82	29.24	4.70
3	Critical pressure [bar]	$(P_c - P_{c1})^{-0.5} - P_{c2}$	852	838	0.8342		7.86	47.77	84.15	6.94	4.74	11.35
4	Critical volume [cc/mol]	$V_c - V_{c0}$	797	783	0.9908		34.63	88.96	97.87	18.48	12.16	3.12

Table 3 (*Continued*)

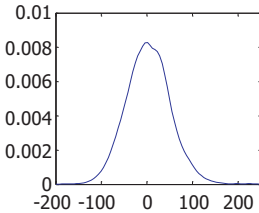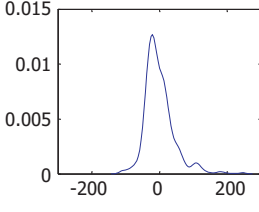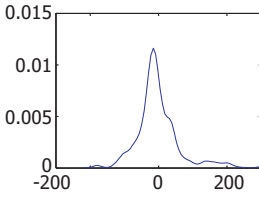| Sl. No.	Property	$f(X)$	N	ν	R^2	Residual distribution plot	CI method based property prediction model $f(X) = \sum_i a_i A_i + b(^{\nu}\chi^0) + 2c(^{\nu}\chi^1) + d$					
							$P_{rc}\,(\pm1\%)$	$P_{rc}\,(\pm5\%)$	$P_{rc}\,(\pm10\%)$	SD	AAE	ARE [a]
5	Normal melting point [K]	$\exp\left(\frac{T_m}{T_{mo}}\right)$	5183	5169	0.7135		8.47	41.75	71.77	51.31	38.68	12.32
6	Gibbs free energy [kJ/mol]	$G_f - G_{f0}$	749	737	0.9555		2.94	18.69	41.52	44.90	32.23	–
7	Enthalpy of formation [kJ/mol]	$H_f - H_{f0}$	882	869	0.9271		4.65	27.21	49.32	67.39	44.70	–

Table 3 (*Continued*)
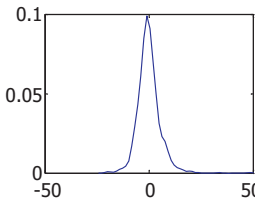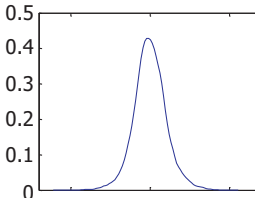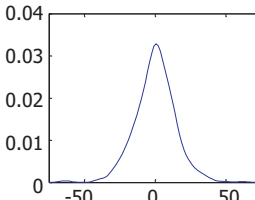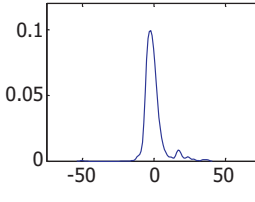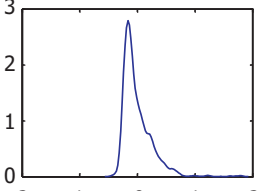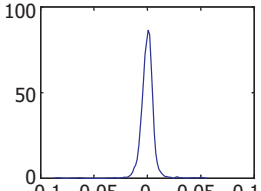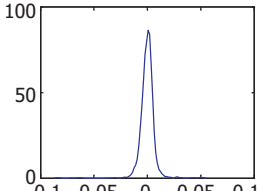
Sl. No.	Property	CI method based property prediction model $f(X) = \sum_i a_i A_i + b(^v\chi^0) + 2c(^v\chi^1) + d$										
		$f(X)$	N	ν	R^2	Residual distribution plot	P_{rc} (±1%)	P_{rc} (±5%)	P_{rc} (±10%)	SD	AAE	ARE [a]
8	Enthalpy of fusion [kJ/mol]	$H_{fus} - H_{fus0}$	764	751	0.6913		4.32	18.72	35.34	7.00	4.04	–
9	Octanol/Water partition coefficient	$Log K_{ow} - K_{ow0}$	12193	12179	0.6705		2.61	14.68	28.76	1.03	0.78	–
10	Flash point [K]	$F_p - F_{po}$	512	500	0.7951		31.45	86.91	97.46	14.30	10.66	3.27
11	Enthalpy of vaporization (298 K) [kJ/mol]	$H_v - H_{v0}$	705	692	0.7232		8.51	43.12	66.52	7.31	4.63	10.47

Table 3 (*Continued*)

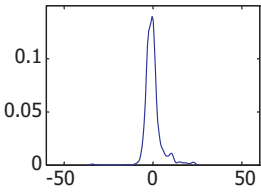| Sl. No.	Property	$f(X)$	N	ν	R^2	Residual distribution plot	CI method based property prediction model $f(X) = \sum_i a_i A_i + b(^\nu\chi^0) + 2c(^\nu\chi^1) + d$					
							$P_{rc}\,(\pm1\%)$	$P_{rc}\,(\pm5\%)$	$P_{rc}\,(\pm10\%)$	SD	AAE	ARE [a]
12	Enthalpy of vaporization (T_b) [kJ/mol]	$H_{vb} - H_{vb0}$	512	499	0.5668		17.19	51.56	77.73	4.72	2.98	8.14
13	Acentric factor	$exp\left(\frac{\omega}{\omega_a}\right)^{\omega_b} - \omega_c$	1723	1709	0.6243		4.24	23.33	40.57	0.1837	0.1207	25.74
14	Liquid molar volume [cc/kmol]	$V_m - V_{m0}$	1056	1042	0.9918		35.13	85.51	94.98	0.007	0.004	3.66

[a] For G_f, H_f, H_{fus}, and $LogK_{ow}$, ARE is not reported since these properties contain both positive and negative values.

uncertainty). This information, for example, can be used in product-process design to take into account the effect of uncertainties of predicted property values on the quality of the design.

2.6. Statistical performance indicators

The statistical significance of the developed correlations discussed in this paper is based on the following performance indicators [5,20].

- Standard deviation (SD): This parameter measures the spread of the data about its mean value μ and is given by,

$$SD = \frac{\sqrt{\sum_j (X_j^{exp} - X_j^{pred})^2}}{N} \quad (12)$$

- Average absolute error (AAE): This is the measure of deviation of predicted property values from the experimentally measured property values and is given by,

$$AAE = \frac{1}{N} \sum_j |(X_j^{exp} - X_j^{pred})| \quad (13)$$

- Average relative error (ARE): This provides an average of relative error calculated with respect to the experimentally measured property values and is given by,

$$ARE = \frac{1}{N} \sum_j \frac{|(X_j^{exp} - X_j^{pred})|}{X_j^{exp}} \times 100 \quad (14)$$

- Coefficient of determination (R^2): This parameter provides information about the goodness of model fit. An R^2 close to 1.0 indicates that the experimental data used in the regression have been fitted to a good accuracy. It is calculated using,

$$R^2 = 1 - \left[\frac{\sum_j (X_j^{exp} - X_j^{pred})^2}{\sum_j (X_j^{exp} - \mu)^2} \right] \quad (15)$$

3. Results

The results of property modeling and uncertainty analysis are provided in this section for the following pure component property prediction models.

- MG method based property models analyzed using step-wise regression method.
- MG method based property models analyzed using simultaneous regression method.
- CI method based property models.

3.1.1. Model performance

The model performance statistics for the property models analyzed using the step-wise regression method and the CI method is summarized in Tables 2 and 3 respectively. The model fits (Figs. S1–S11) to the data (parity plots) for pure component properties analyzed using the step-wise regression method, the simultaneous regression method and CI method are given as a supplementary material and can be downloaded from http://www.capec.kt.dtu.dk/documents/property_modeling_and_uncertainty_analysis/Supplementary_material.pdf. The model performance statistics for properties analyzed using simultaneous

regression method are given in Table S1 in the supplementary material. Visual inspection of Figs. S1–S11 indicates that the goodness of the model-fits is excellent and most of the data have been fitted to a good degree of accuracy. In Table 2, N is the number of data-points considered in the regression and ν is the degree of freedom for each property and is obtained by subtracting number of estimated model parameters from N. $P_{rc}(\pm 1\%)$, $P_{rc}(\pm 5\%)$, and $P_{rc}(\pm 10\%)$ represents the percentage of the experimental data-points (N) found within $\pm 1\%$, $\pm 5\%$, and $\pm 10\%$ relative error range respectively. AE_{max} is the maximum absolute error and RE_{max} is the maximum relative error obtained in the regression analysis. For property models analyzed using step-wise regression method, the results for R^2, SD, AAE and ARE have been obtained after third-level estimation and hence they represent the global results. All property models have been improved compared to their earlier versions. The property models analyzed using simultaneous regression method have shown some improvement in the performance as compared to the performance of property models analyzed using step-wise regression method. The residuals ($X^{exp} - X^{pred}$) for data-points considered in the regression are plotted in the form of residual distribution plots and are included in Tables 2 and 3, and S1. From these distribution plots it can be seen that the residuals obtained from most of the models follow a normal distribution curve with mean zero suggesting a good fit of the experimental data used in the regression and there is no apparent bias in the predicted property values as well as the assumption (random errors follow a normal distribution) behind the approach is valid. The use of the CI method for creating the missing groups and predicting their contributions through the regressed contributions of connectivity indices as suggested by Gani et al. [6] has been implemented. This allows one to make predictions for a number of properties for which neither experimental data nor the GC-property model parameters are available. The CI method based property models for T_b, T_c, P_c, V_c, T_m, G_f, H_f, H_{fus}, $LogK_{ow}$, F_p, H_v, H_{vb}, ω, and V_m are included in this work.

It is important to note that we have considered all of the available experimental data of the pure component properties in the regression step. There are two main reasons:

- large quantity of experimental data used in the regression helps in achieving improved quality of parameter estimation (lower SD, lower AAE and lower ARE);
- since the validation set is to be formed by randomly selecting the experimental data-points, some of the group/atom contributions might not be estimated (which affects the application range of the property prediction method) due to the lack of necessary experimental data that was made unavailable during random selection process.

Our viewpoint is that by including all of the available experimental data of the property in the regression, it is possible to improve the predictive capability and application range of the property model. To illustrate this point, for standard enthalpy of formation, the effect of quantity of experimental data on the quality of the parameter estimation is shown by considering different combinations of training sets and validation sets formed by random selection of the experimental data-points (see Table 4). It can be seen that better model performance statistics (lower SD, and lower AAE) is obtained for the case in which all of the experimental data-points are considered in the regression. Also, the total number of model parameters involved in the regression (i.e. model parameters for which estimation of contribution was possible) is highest when all of the data-points are considered in the regression.

The variables T_{b0}, T_{c0}, P_{c1}, P_{c2}, V_{c0}, T_{m0}, G_{f0}, H_{f0}, H_{fus0}, K_{ow0}, F_{p0}, Ait_1, Ait_2, H_{v0}, H_{vbo}, S_{vb0}, δ_0, ω_a, ω_b, ω_c, and V_{mo} as defined in the functional forms, $f(X)$ given in Tables 2 and 3, and S1 are additional

Table 4
Effect of quantity of experimental data on quality of parameter estimation.

Distribution of experimental data	SD in kJ/mol	AAE in kJ/mol	No. of model parameters estimated [a]
50% for training purpose	9.59	5.31	194
66.67% for training purpose	8.73	4.75	217
75% for training purpose	7.09	4.34	229
All data-points for training purpose	6.60	4.15	232

[a] The total no. of model parameters (first-order, second-order and third-order group contributions) is 424.

Table 5
Values of the universal constants (additional adjustable parameters).[a]

Universal constants	Value (step-wise method)	Value (simultanous method)
T_{b0} [K]	244.5165	244.7889
T_{c0} [K]	181.6716	181.6738
P_{c1} [bar]	0.0519	0.0519
P_{c2} [bar$^{-0.5}$]	0.1347	0.1155
V_{c0} [cc/mol]	28.0018	14.6182
T_{m0} [K]	143.5706	144.0977
G_{f0} [kJ/mol]	−1.3385	8.5016
H_{f0} [kJ/mol]	35.1778	83.9657
H_{v0} [kJ/mol]	9.6127	10.4327
H_{fus0} [kJ/mol]	−1.7795	−1.2993
K_{ow0} [−]	0.4876	0.7520
F_{p0} [K]	170.7058	150.0218
Ait_1 [b] [−]	–	71.2584
Ait_2 [b] [K]	–	525.93
H_{vb0} [kJ/mol]	15.4199	15.0884
S_{vb0} [kJ/mol]	83.3097	83.7779
δ_0 [MPa$^{1/2}$]	21.6654	20.7339
ω_a [−]	0.9080	0.9132
ω_b [−]	0.1055	0.0447
ω_c [−]	1.0012	1.0039
V_{mo} [cc/kmol]	0.0160	0.0123

[a] The values of universal constants for the CI models are the same as those given for the step-wise method.
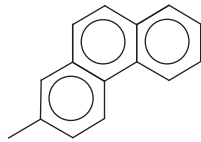
[b] *Ait* property is modeled using simultaneous regression method only.

adjustable parameters of property prediction models. The values of these parameters are listed in Table 5. The total list of groups and their contributions C_i, D_j, and E_k for the 18 pure component properties modeled in this paper are given in the supplementary material (see Tables S6–S8 for MG method based models analyzed using step-wise regression method, and Tables S2–S4 for MG method based models analyzed using simultaneous regression method). The list of atoms, their contributions a_i, adjustable model parameters (b and c), and universal parameter d for CI method based property prediction models are given in the supplementary material (see Table S8). The covariance matrix computed using Eq. (6) for each property prediction model analyzed using the MG method (for models with step-wise regression method and simultaneous regression method) and using the CI method is available upon request from the authors. A sample table (Table S9) containing list of first-order, second-order, and third-order groups along with their total occurrences in the data-set (used for the regression) of the properties T_b, T_c, P_c, V_c, and T_m is provided in the supplementary material. Such tables for all other properties are available upon request from the authors.

3.1.2. Application of the developed methodology

The application of the developed methodology to estimate properties of pure components and to quantify the uncertainty in the estimated property value is illustrated by considering predictions of normal boiling points (using the model parameters obtained from simultaneous regression method) for the component: Phenanthrene, 2-methyl- (CAS No. 2531-84-2). Table 6

Table 6
Estimation of normal boiling point of phenanthrene, 2-methyl-.

Phenanthrene, 2-methyl- CAS No. 2531-84-2	Molecular structure
Molecular formula: $C_{15}H_{12}$	

	Occurrences	Contribution
First-order groups		
aCH	9	0.7332
aC fused with aromatic ring	4	1.2531
aC-CH$_3$	1	1.2616
Second-order groups		
No second-order groups are involved		
Third-order groups		
AROM.FUSED[2]	1	-0.1599
AROM.FUSED[2]s^2	1	-0.1829
AROM.FUSED[4p]	1	0.0119

$$T_b^{pred} = T_{bo} \ln \left(\sum_i N_i C_i + w \sum_j M_j D_j + z \sum_k E_k O_k \right) = 619.09 \text{ K.}$$

Absolute error = 623.15 K–619.09 K = 4.05 K
Constantinou and Gani method [4] = 577.01 K (this value is estimated using first order groups only as second-order groups are not available for this component).
Joback and Reid method [1] = 617.20 K.
Marrero and Gani method [5] = 626.65 K.
The SD, AAE, ARE and R^2 for Marrero and Gani method [5] are 9.42 K, 6.74 K, 1.56%, and 0.98 respectively. The SD, AAE, ARE and R^2 for Marrero and Gani method (this work) are 7.91 K, 6.16 K, 1.43%, and 0.99 respectively. For calculation of these performance indicators, a common data-set of 3435 components that can be described by Marrero and Gani method [5] and by the present work was considered.

provides information of first-order, second-order and third-order groups used to represent Phenanthrene, 2-methyl-, their frequency and the contributions for each group (T_{b1i}, T_{b2j}, and T_{b3k}) taken from Tables S5–S7 given in the supplementary material. Using this information and the property model for normal boiling point, we estimate the normal boiling point of Phenanthrene, 2-methyl- as 619.09 K (experimental value is 623.15 K).

Note that even though the absolute error for this component with the Marrero and Gani [5] is slightly better, the SD, AAE, ARE and R^2 for the whole data-set is better for the Marrero and Gani method (this work). Also, it can be noted from Fig. 2 (plot of absolute relative error versus data-set for normal boiling point), the performance of the new model is better.

As discussed in Section 2.5, to quantify the uncertainty in the estimated normal boiling point, we use information of covariance $COV(\boldsymbol{P}^*)$ of the involved groups and universal constant T_{bo}, and local sensitivity $J(\boldsymbol{P}^*)$ of the normal boiling point model. The covariance of the listed groups as given in Table 7 was extracted from the whole covariance matrix for all the groups obtained for the case of normal boiling point model analyzed using simultaneous regression method. In Table 7, only lower triangular elements are shown since the upper triangular matrix elements are identical to the lower one.
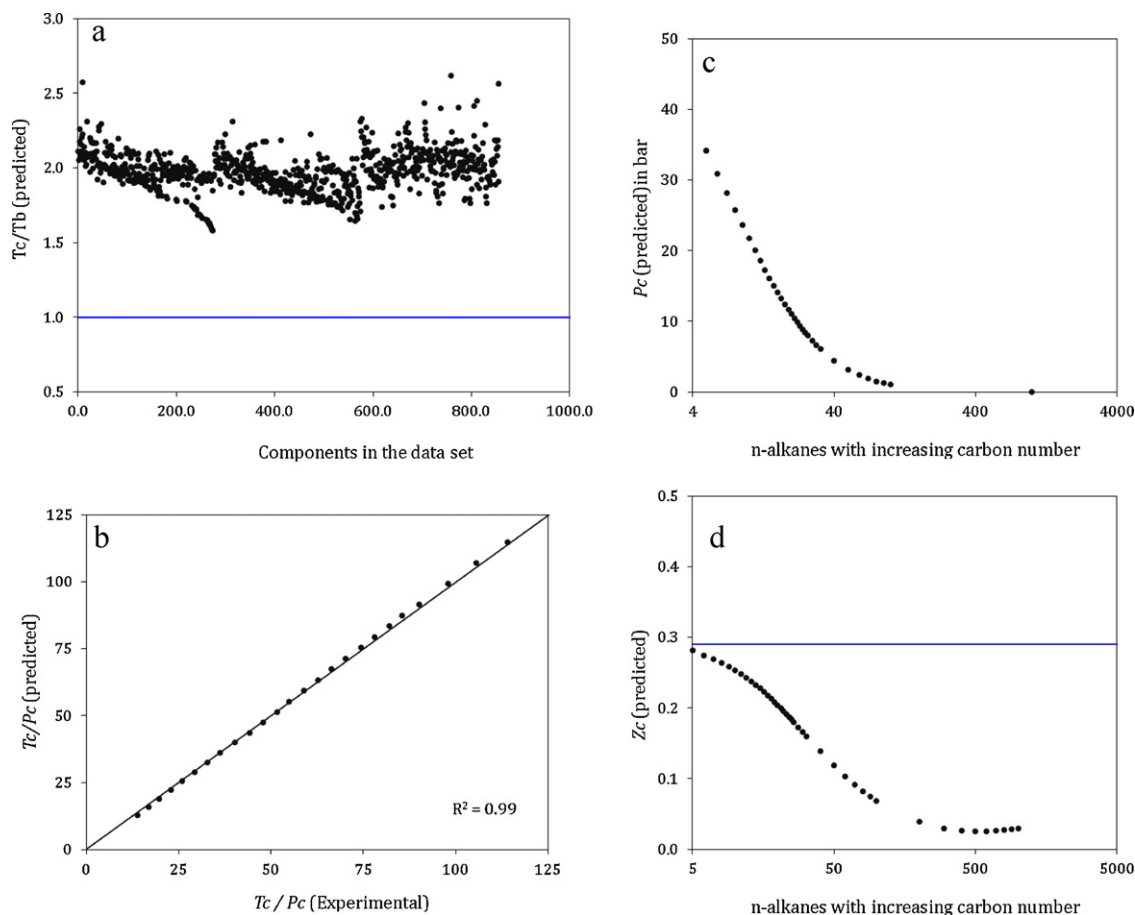
Fig. 5. Plots for: (a) T_c/T_b ratio for a wide range of components in the data set; (b) experimental values of T_c/P_c versus predicted values of T_c/P_c; (c) predicted P_c of n-alkanes; (d) predicted Z_c of n-alkanes.

4.2.5. Critical density (d_c) of n-alkanes

Critical density can be expressed as an inverse of the critical volume. An n-alkane molecule with c number of carbons can be represented by CH_3 group with 2 occurrences and by CH_2 group with c-2 occurrences. The molecular weight of n-alkane can be represented by $14c + 2$ (kg/kmol). Therefore, expression for d_c (in kg/m^3) after substituting values of group-contributions for CH_3 and CH_2 is given by,

$$d_c = \frac{1}{V_c} = \frac{(14c + 2)}{(56.5948c - 24.0164)} \times 1000 \qquad (16)$$

As c tends to an infinite value, the limiting value for the critical density is 247.37 kg/m^3. The limiting value obtained for critical density seems to be reasonable and is in excellent agreement with the theoretical foundation [24].

5. Conclusions

Property modeling and uncertainty analysis of GC$^+$ models for pure component properties has been performed to provide more reliable property predictions together with an estimate of prediction errors (uncertainties) of the predicted property values. To improve the performance, the application range, and the prediction capability of earlier version of the property models, the updated CAPEC database containing large experimental data-sets of properties is used in the parameter estimation to develop revised and improved parameters of the GC$^+$ models. The application range of the GC$^+$ models is increased by making use of the new experimental data available in the CAPEC database to estimate model parameters whose values were not determined and reported in the earlier

work due to the lack of necessary experimental data. In addition, a new approach for property modeling based on the simultaneous regression method is developed to improve the performance of the GC$^+$ models. In total 18 properties of pure components have been modeled and analyzed. The properties- the acentric factor and the liquid molar volume (298) which were modeled earlier by the Constantinou and Gani method are modeled using the Marrero and Gani method to provide improved accuracy and the application range of the models. The application of the developed methodology to estimate pure component properties and the uncertainties of the predicted property values is highlighted through an application example. The reliability of the property models has been tested and illustrated by comparing model prediction uncertainties with reported range of experimental measurement uncertainties for the properties with related available data. Important issues related to property modeling such as thermodynamic consistency of the predicted properties, and predictive capability of the models have been addressed. The developed methodology for property modeling and uncertainty analysis is simple, yet sound and effective and provides not only the estimated property values but also the uncertainties in the estimated property values. Although not shown in this paper, it can be stated that this feature would allow one to evaluate the effect of these uncertainties on the product-process design, simulation, and optimisation calculations involving use of the predictive models for obtaining the needed property values thereby contributing to better-informed and reliable engineering solutions. Motivated by the results obtained in this work, our current and future work is focused on extending the developed methodology to other pure component properties, such as transport properties (surface tension, viscosity, and

thermal conductivity), and environment-related properties (LC50, aqueous solubility, bio-concentration factor, LD50, photochemical oxidation potential, global warming potential, ozone depletion potential, acidification potential, and permissible exposure limit). As an extension of the current work on the property modeling and uncertainty analysis, we are investigating the possibility of improving the current performance of the property models by analyzing the available information for each group (such as, occurrences of the groups in the data-set, relative prediction error for each group). The work on quantification of effect of uncertainties of predicted properties on the product-process design is under investigation as well.

List of symbols

AAE	average absolute error
AE_{max}	maximum absolute error found from the regression
a_i	contribution of atom of type-i
A_i	occurrence of atom of type-i
Ait	auto ignition temperature [K]
ARE	average relative error [%]
b	adjustable parameter of Eq. (4)
$B(1/2, \nu/2)$	beta function
c	adjustable parameter of Eq. (4)
CI	atom connectivity index
C_i	contribution of first-order group of type-i
$COV(\boldsymbol{P}^*)$	covariance matrix
d	universal parameter of Eq. (4)
d_c	Critical density [kg/m^3]
D_j	contribution of second-order group of type-j
E_k	contribution of third-order group of type-k
$f(X)$	function of property X
F_p	flash point [K]
GC	group-contribution
GC$^+$	group-contribution$^+$
G_f	standard Gibbs energy of formation [kJ/mol]
H_f	standard enthalpy of formation [kJ/mol]
H_{fus}	normal enthalpy of fusion [kJ/mol]
H_v	enthalpy of vaporization at 298 K [kJ/mol]
H_{vb}	enthalpy of vaporization at the normal boiling point [kJ/mol]
$J(\boldsymbol{P}^*)$	local sensitivity of the model to variations in estimated model parameters
$Logk_{ow}$	octanol/water partition coefficient
MG	Marrero and Gani
M_j	occurrence of second-order group of type-j
N	number of experimental data-points used in the regression
N_i	occurrence of first-order group of type-i
O_k	occurrence of third-order group of type-k
$\boldsymbol{P}$	model parameters
$\boldsymbol{P}^*$	estimated values of model parameters
P_c	critical pressure [bar]
$P_{c\infty}$	limiting value for critical pressure as carbon number goes to a high number [bar]
$P_r(t, \nu)$	Students t-distribution function
P_{rc}	percentage of the experimental data-points [%]
R	universal gas constant [cc-bar/mol K]
R^2	coefficient of determination
RE_{max}	maximum relative error found from the regression [%]
$S(\boldsymbol{P})$	cost function in Eq. (5)
SD	standard deviation
SSE	minimum sum of squared errors
S_{vb}	entropy of vaporization at the normal boiling point [J/mol K]

$t(\nu, \alpha_t/2)$	t-distribution value corresponding to the $\alpha_t/2$ percentile
T_b	normal boiling point [K]
T_c	critical temperature [K]
T_m	normal melting point [K]
V_c	critical volume [cc/mol]
V_m	liquid molar volume at 298 K [cc/kmol]
X^{exp}	experimental property value
X^{pred}	predicted property value
Z_c	Critical compressibility factor

Greek symbols

ω	acentric factor
$^\nu\chi^0$	zeroth-order (atom) connectivity index
$^\nu\chi^1$	first-order (bond) connectivity index
ν	degrees of freedom
δ	Hildebrand solubility parameter [MPa$^{1/2}$]
δ_D	Hansen solubility parameter- dispersion [MPa$^{1/2}$]
δ_H	Hansen solubility parameter-hydrogen bond [MPa$^{1/2}$]
δ_P	Hansen solubility parameter- polar [MPa$^{1/2}$]

Acknowledgments

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fluid.2012.02.010.

References

[1] K.G. Joback, R.C. Reid, Chem. Eng. Commun. 57 (1987) 233–243.
[2] A.L. Lydersen, Estimation of critical properties of organic compounds, College Engineering University Wisconsin Engineering Experimental Station Report 3, Madison, WI, April, 1955.
[3] K.M. Klincewicz, R.C. Reid, AIChE J. 30 (1984) 137–142.
[4] L. Constantinou, R. Gani, AIChE J. 40 (1994) 1697–1710.
[5] J. Marrero, R. Gani, Fluid Phase Equilibr. 183-184 (2001) 183–208.
[6] R. Gani, P.M. Harper, M. Hostrup, Ind. Eng. Chem. Res. 44 (2005) 7262–7269.
[7] W.B. Whiting, T.M. Tong, E.M. Reed, Ind. Eng. Chem. Res. 32 (1993) 1367–1371.
[8] R. Dohrn, O. Pfohl, Fluid Phase Equilibr. 194–197 (2002) 15–29.
[9] W.B. Whiting, J. Chem. Eng. Data 41 (1996) 935–941.
[10] S. Macchietto, G. Maduabeuke, R. Szczepanski, Fluid Phase Equilibr. 29 (1986) 59–67.
[11] S. Hajipour, M. Satyro, Fluid Phase Equilibr. 307 (2011) 78–94.
[12] C. Maranas, Comp. Chem. Eng. 21 (1997) 1019–1024.
[13] J. Marrero, R. Gani, Ind. Eng. Chem. Res. 41 (2002) 6623–6633.
[14] Z. Kolska, V. Ruzicka, R. Gani, Ind. Eng. Chem. Res. 44 (2005) 8436–8454.
[15] H. Modarresi, E. Conte, J. Abildskov, R. Gani, P. Crafts, Ind. Eng. Chem. Res. 47 (2008) 5234–5242.
[16] T. Nielsen, J. Abildskov, P. Harper, I. Papaeconomou, R. Gani, J. Chem. Eng. Data 46 (2001) 1041–1044.
[17] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C – The Art of Scientific Computing, 2nd ed., Cambridge University, Cambridge, UK, 1992.
[18] G. Seber, C. Wild, Nonlinear Regression, Wiley, New York, 1989.
[19] M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, Dover, New York, 1972.
[20] E. Belia, Y. Amerlinck, L. Benedetti, B. Johnson, G. Sin, P.A. Vanrolleghem, K.V. Gernaey, S. Gillot, M.B. Neumann, L. Rieger, A. Shaw, K. Villez, Water Sci. Tech. 60 (8) (2009) 1929–1941.
[21] Design Institute for Physical Properties, Sponsored by AIChE DIPPR Project 801 – Full Version, Design Institute for Physical Property Data/AIChE.
[22] B.E. Poling, J.M. Prausnitz, J.P. O'Connell, The Properties of Gases and Liquids, 5th ed., McGraw-Hill, New York, 2000.
[23] C. Tsonopoulos, AIChE J. 33 (1987) 2080–2083.
[24] C. Tsonopoulos, Z. Tan, Fluid Phase Equilibr. 83 (1993) 127–138.
[25] R.A. McFarlane, M.R. Gray, J.M. Shaw, Fluid Phase Equilibr. 293 (2010) 87–100.