Article

Group Contribution Revisited: The Enthalpy of Formation of Organic Compounds with "Chemical Accuracy" Part II

Robert J. Meier [ID]

Pro-Deo Consultant, 52525 Heinsberg, North Rhine-Westphalia, Germany; r.meier@planet.nl

Abstract: Group contribution (GC) methods to predict thermochemical properties are eminently important to process design. We present a group contribution parametrization for the heat of formation of organic molecules exhibiting chemical accuracy, maximum 1 kcal/mol (4.2 kJ/mol) difference between experiment and model values while minimizing the number of parameters avoiding overfitting and therewith avoiding reduced predictability. Compared to the contemporary literature, this was successfully achieved by employing available literature high-quality and consistent experimental data, optimizing parameters group by group, and introducing additional parameters when chemical understanding was obtained supporting these. A further important result is the observation that the applicability of the group contribution approach breaks down with increasing substitution levels, i.e., more heavily alkyl-substituted molecules, the reason being a serious influence of substitution on the conformation of the flexible part of the entire molecule within particular valence angles and torsional angles affected, which cannot be accounted for by additional GC parameters with fixed numerical values.

Keywords: enthalpy of formation; thermodynamics; physico-chemical property prediction; molecular modeling; group contribution method; chemical accuracy; quantum chemistry

1. Introduction

For the purpose of the evaluation of the enthalpy of formation ΔH_f of organic molecules from their molecular structure, the group contribution (GC) approach is one of the most important and applied methods. For further background and references, we refer to the work of [1]. In this work, the ΔH_f is the value for the gaseous phase at 298.15 K. The original GC method is based on the assumption that a molecule can be decomposed in molecular fragments, which are in essence mutually independent, and the molecular property of interest is the sum of the individual properties of the molecular fragments j:

$$\Delta H_f = \sum_{j=1,N} N_j \, \Delta H_f \, (j) \tag{1}$$

Equation (1) is the general formulas for a GC method with chemical groups present and where only the first-order term $\sum_I N_i C_i$ has been retained [1–3]. ΔH_f (j) is the contribution to the heat of formation associated with group j, and N_i is the number of times this group is present in the molecule. We hereby basically follow the original van Krevelen-Chermin approach [2]. In Scheme 1, Structure (1) this is illustrated for four distinct groups, CH_3 (3 times), CH_2 (4 times), CH, and phenyl, that constitute the molecular structure depicted. This philosophy truly works according to Equation (1) under the assumption that there are no or very minimal interactions between the groups that influence the properties of a single group, such as conjugation or steric effects due to, e.g., larger, substituents. When we look at linear and slightly branched alkanes, we know from experience that the condition of independent fragments is fulfilled, e.g., see the work by van Krevelen [2] and others, but also the observation that one obtains very suitable results (model calculated values) for the mono-methyl branched alkanes while using a single parameter for both terminal as well

as branched methyl groups. Thus, whereas a methyl branch as in Structure (2) is, on the one hand, not exactly the same as the terminal methyl groups in that structure (different local environment), the difference is negligible in the current context. However, when we look at Structure (3), we will see, in the current manuscript, that corrections are required due to steric interactions. In a Structure such as (4), the fluorine atoms have an electronic effect on the nearest and next-nearest carbon atoms as they induce measurable changes in electronic structure, e.g., the core electron energy levels (the C_{1s} atomic orbitals) of the carbon atoms [3], and thus modifying the properties of the carbon atoms. Primary and secondary alcohols already require different OH GC parameter values in order to obtain a truly suitable model, as we will see discussed in the present work.

Structure (1)

Structure (2) Structure (3) Structure (4)
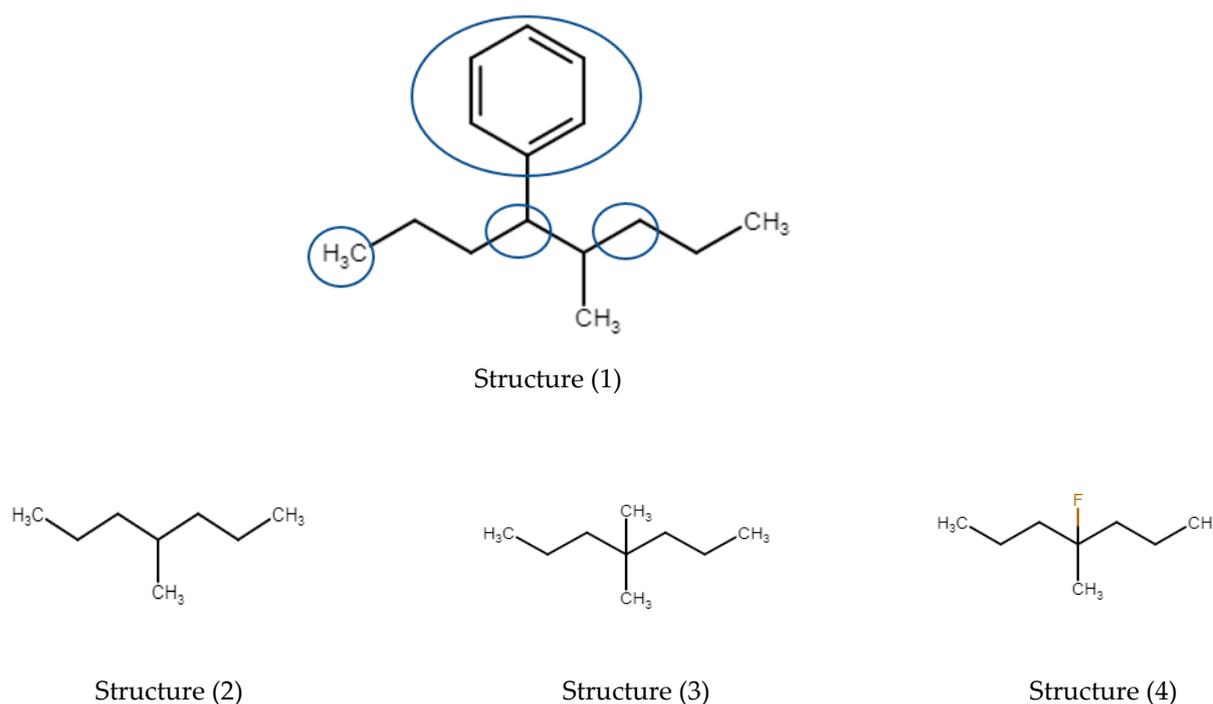
Scheme 1. Structures illustrating the Group Contribution concept (1) and steric (2,3) and electronic (4) interactions. For details see the text.

An important point is that the definition of the groups is not unique. In literature, different approaches defining the constituting groups have been employed. One approach is to define the smallest possible entities, e.g., the Gani et al. approach [4] takes an aCH, an aromatic CH group, as a group [5,6] (Note for Ref. [5]: This software module is regularly updated, so individual results may differ from those reported in papers including Refs. [4,6]), which was also practiced by Kadda et al. [7] (see in particular the Supplementary Material to reference [7]). Adopting this choice leads to benzene consisting of 6 aCH groups. Other approaches, e.g., van Krevelen and Chermin [2] and Meier [1], took larger entities as the individual groups, e.g., the phenyl entity (rather than 6 aCH groups). In specific cases, the one choice has been argued more successful, e.g., in a group contribution-based SAFT approach to describe the phase behavior of multi-component mixtures [8]. Ideally, all such approaches can potentially lead to an appropriate GC model, but it remains to be seen whether this holds in practice. The question that now comes up is what precisely is an appropriate or sufficiently suitable model? Several factors determine the quality of the model, i.e., making predictions with a pre-set reliability (can I rely on the generated number) and accuracy (difference experimental and model values). First of all, the quality of the experimental data employed to parametrize is paramount. Several groups with high-level knowledge of how to derive the thermodynamic quantities from experimental data along with the availability of experimental equipment to establish these data have

published a series of papers that are therefore also highly consistent (same equipment, same method, same lead scientists). Secondly, focusing on the heat of formation, we want to require "chemical accuracy", implying predictions should be within 4.2 kJ/mol from the correct experimental value. This is quite a challenge and mostly unrealized up till now, but the criterium is not unrealistically tight because the practical goal should be the prediction of chemical equilibrium $\Delta G = \Delta H - T\Delta S$ of a chemical reaction such as A + B → C + D, and an error of 1 kcal/mol in each of the individual heats of formation can lead to a cancellation of errors but also to a total error of up to 4 kcal/mol. As there are quite some reactions for which $\Delta G \sim 0$, we do need highly accurate predictions within chemical accuracy. It needs to be accepted; however, that experimental data also have an error, often indicated in high-quality works, and this error might go up to several kJ/mol and therewith similar to the required chemical accuracy.

Up until recently, the best agreement between model and experimental values was claimed by following the Marrero-Gani (MG) approach [4] and subsequent parametrization of the GC parameters [6,7]. This approach also involves second and third-order parameters [4] in order to obtain a suitable fit when straightforward addition of first-order groups did not lead to the required accuracy. Whereas the averaged absolute difference between model and experimental values was quite suitable, often within chemical accuracy, there were very clear outliers. More importantly, overfitting consequently reduced reliable predictivity, which was a most definite issue. Explicit numbers illustrate this, namely Kadda et al. [4] employing the MG approach [4], taking a database of 750 molecules employed just over 300 groups parameters (115 first-order, 77 second-order, 36 third-order groups, and 83 new groups) a ratio of almost 2.

Verevkin et al. [9] have reported a new GC approach with the parametrization of hydrocarbons and oxygen-containing molecules. Actual values and analysis of these were reported for the error in experimental values, the model values, as well as for the difference between experimental and model values, and this for each individual species. In a way, their approach has more in common with the approach in our previous paper [1] than with the MG approach in the sense that the chemically more intuitive 1,4- and 1,5-interactions were included to obtain better agreement between model and experimental values. Highly branched structures such as 3,3,4,4-tetraethylhexane as well as various other structures were excluded on the basis of the argumentation that properties of such compounds commonly exhibit non-additive behavior (see also Section 3.1.4).

In 2017 Mathieu [10] reported on an atom pair contribution (APC) model in which the elements constituting the APC approach are pairs of bonded and geminal atoms. The GC and APC methods seem to have a huge similarity: a molecule is decomposed into a set of constituting elements, molecular groups (GC method), or bonded atom pairs (APC model), and the physical property is evaluated by an equation equivalent to Equation (1). Mathieu argued that "Unlike GC methods, the present APC model has wide applicability with the number of adjustable parameters limited to 68". Contrary to the approach by Verevkin (GC method) who involved 1,4-interactions, Mathieu's approach "In contrast to all existing methods, it is based on the assumption that atom−atom interactions between bonded and geminal atom pairs are transferable, while 1,4-interactions are neglected". Somewhat shockingly, the root-mean-square-errors (RMSEs) for the APC model [10] was typically 20 kJ/mol, and for hydrocarbons even over 30 kJ/mol. Some of the individual absolute deviations are consequently still much larger than these values. The APC method as proposed [10] is thus way beyond a method to be used to predict heats of formation with any acceptable accuracy and reliability.

In the work of [1], we explained why we reinvestigated the GC approach for the evaluation of the heat of formation of organic molecules. A key issue was the adoption of the criterion "chemical accuracy" (1 kcal/mol) as an absolute necessity for useful application. Overfitting was avoided by carefully analyzing *individual* experimental data and only allowing for (next)nearest neighbor effects, which will lead to additional parameters but as minimally as possible. As an example, for the di-, tri-, and tetrasubstituted alkylbenzenes,

we added three additional parameters while achieving chemical accuracy for *each* individual species, whereas the MG approach required 11 (see reference [1] for further details). Furthermore, the group-related parameters were determined stepwise: first alkanes only, then subsequently only one additional group in the next class of molecules. This ensures unique and optimal parameter values for each chemical group. Furthermore, we have found it important (i) to explicitly account for common chemical knowledge, e.g., ring strain and geminal effects, (ii) for specific cases, we applied ab initio quantum calculations to verify whether certain effects are truly present rather than straightforwardly applying mathematical fitting routines to obtain GC parameters without knowing whether we have to do with true non-linearity effects, (iii) use a set of reliable, verifiable and consistent (from a single or very few sources) experimental data being crucial to achieving the desired quality of the resulting predictive tool. These ingredients have led to a method showing chemical accuracy throughout while minimizing the number of parameters so we obtain more reliable predictive behavior because we avoid the risk of overfitting.

In the work of [1], we reported group contribution parameters for the n-alkanes, monomethylalkanes, various classes of alkenes, n-alcohols, n-alkylamines, n-aldehydes, ethers, 2-alkanones, mono-, and di-carboxylic acids, alkynes, mono- and di-nitriles, alkyl-substituted benzenes, alkyl-substituted naphthalenes, and alkyl-substituted cycloalkanes. In the present paper, we will revisit the mono-substituted methylalkanes and the alcohols and extend the model with di-, tri-, tetra-, and penta-substituted alkylalkanes, alkylethers, and dienes. As we will see, these extensions compared to the work presented in reference [1] will provide us a better view of the possibilities and restrictions of the group contribution approach when considering the heat of formation of organic molecules.

2. Methods

2.1. Experimental Data and Computational Methods

The key model employed in this work is the group contribution approach. The practical use is by applying Equation (1) once the groups and individual group parameters have been determined. The latter is what the content of this paper is about. As the GC method is a data-driven model (experimental data being used to parametrize the model), our self-imposed requirement on chemical accuracy makes that the quality and accuracy of the experimental data preeminent. Therefore we prefer to use experimental data from a few sources involving high-quality expertise and always the same measuring equipment, measuring protocol, and data processing. We primarily adopted data from publications from the Rossini group and Verevkin c.s. (explicit references will be referred to in the Section 3). Apart from a few exceptions in order to have additional data points, contrary to our work in reference [1], we have avoided using the NIST database [11] and the CAPEC database [12] (the explicit values were not published in Ref. [12] and therefore not quoted in this paper explicitly). Experimental errors as commonly around 1–1.5 kJ/mol (see, e.g., the work of [11] and the papers by the Rossini group and Verevkin c.s. referenced further below), though for some species, the error is indicated larger. Therefore we need to take into account that an experimental value could be off the true value up to half the value of the chemical accuracy we want to achieve. It goes without saying this will have a certain impact on the quality of parametrization but unfortunately unavoidable.

In specific cases, we will use density functional theory (DFT)-type quantum chemical calculations to verify energy differences between similar species in order to verify whether experimental heat of formation differences can be substantiated independently. Relative energy differences between structurally very similar structures can be evaluated with suitable confidence. Quantum chemical calculations were performed using the Spartan 10 program suite [13], involving full geometry optimization within the density functional theory (DFT)-type calculations involving the B3LYP functional. This function is well known in the quantum chemistry world as one that describes structures of standard organic molecules well, and also relative energies are well-accounted for. Medium-level basis set was employed, the 6-31+G* basis set when the key interest was the structure (the

diffuse function characterized by + basis set was added to potentially account better for steric overlap effects for branched structures), and the 6-311+G** basis set where the relative energy as also of interest. The ab initio part is only used in specific cases to investigate whether certain effects seen in the experimental data set are realistic and not due to an error in single experimental data points (in essence, a consistency check). Thus the ab initio results are part of a check and do not directly influence the value of the GC parameters. Larger differences between the initial GC model and experiment, which, when turning out true as supported by ab initio calculations, gives us the reason to introduce one more specific GC parameter, e.g., to account for Me-Me neighbor interactions.

Whenever referred to, heats of formation and group contributions from the Marrero-Gani method [4] were obtained using the ICAS23 software suite [5].

As argued in reference [1], we decided to determine the numerical values of the group contributions by hand and group by group. For each subsequent class of molecules, we determined the new group parameter value on the basis of typically 50% of the molecules named in the tables. We subsequently added further experimental value for molecules belonging to this class to test how well other molecules belonging to this class can be predicted, which can lead to fine-tuning of the parameter values. Performance of the parameter estimation is verified by calculation the differences between model and experimental values, and in addition, calculating the averaged absolute differences (ADD) per class of molecules expressed by

$$\text{AAD} = (1/N) \sum\nolimits_{j\,=\,1,N} (\text{model} - \text{experiment}) \tag{2}$$

2.2. On Corrections beyond the Isolated Group Approach

When we consider a set of tri-substituted alkylbenzenes as collected in Table 1 with the experimental data taken from Rossini c.s. [14], we see that the model values using the model from reference [1], which involves a single substitution parameter of magnitude 30 kJ/mol, leads to an averaged absolute difference between model and experiment of 1.97 kJ/mol, and all individual values are also within chemical accuracy. Analogous to the approach proposed by Gani c.s. [4] involving parameters specific for each substitution pattern, we have determined such additional parameters and obtained the result shown in the last column of Table 1. This lowers the AAD to an amazing 0.30 kJ/mol with the maximum individual difference being 1.42 kJ/mol or, alternatively with slightly different substitution parameters, an ADD of 0.39 kJ/mol and all individual parameters below 0.8 kJ/mol. The ICAS23 tool [5] and taking the experimental data from the CAPEC database [12] led to an ADD of 0.81 kJ/mol and a maximum individual difference of 2.6 kJ/mol, whereas comparing ICAS23 results with the Rossini c.s. [14] experimental data (second column Table 1), we obtained 1.27 kJ/mol and a maximum individual difference of 3.6 kJ/mol.

The magnitude of the substitution pattern-dependent contributions, which are the second-order contributions in the MG approach (AROMRINGs1s2s3, AROMRINGs1s2s4, AROMRINGs1s3s5a), lies in the range of a few kJ/mol and is quite similar for the MG approach (see also the thesis from Hukkerikar [15]) and the current approach. Kadda et al. [7], who also based their analysis on the MG approach, have reported values that mutually differ up to 13 kJ/mol (AROMRINGs1s2s3–AROMRINGs1s3s5), which is far beyond chemically realistic, i.e., much larger than the differences between the corresponding experimental values, viz. Table 1.

Whereas all these results are within chemical accuracy, we see that our stepwise and non-automated approach, determining a single GC parameter at a time, leads to the best results. Still, we need to consider that the experimental error for the data set varies from 1.3 (trimethylbenzenes) to 2.5 kJ/mol (ethyl-substituted species) [14], which is of the same magnitude as the substitution-specific corrections we have determined (−2.4 to +3.6 kJ/mol). Therefore we do not consider it a priori justified to explicitly add additional parameters for the different substitution patterns even though it may sound not unlogical from a chemical point of view. As, e.g., for the tri-substituted alkylbenzenes, our

initial approach led to the desired suitable results with a single substitution parameter of 30 kJ/mol, we will refrain from additional parameters when not truly evidenced and that in addition could also lead to overfitting.

Table 1. Experimental and model values for tri-substituted alkylbenzenes. All values in kJ/mol. Experimental data were adopted from Rossini c.s. [14]. The model values (model dHf) were calculated using the GC parameters and applying the GC formula ΔH_f (tri-substituted benzene) = GC_{Phenyl} + $\sum N_{substituent}$ * $GC_{Substituent}$ + AromTrialkyl previously reported in reference [1]. The red values in the last row are the averaged absolute differences between our model using a single substitution correction [1] (one but last column) or after adopting substitution-specific, e.g., accounting for the difference between 1,2,3- and 1,3,5- substitution patterns, corrections as discussed in the present sub-section (last column). Note that all individual model values, those with as well as those without specific substitution correction, are within chemical accuracy from the experimental values.

Tri-Substituted Alkylbenzenes	Rossini	Model dHf	Model-Exp	ABS (Model-Exp)	Specific Correction 1,2,3-1,2,4-1,3,5-2,3,4-2,3,5-2,4,5 Substitution
benzene, 1,2,4-trimethyl-	−13.9	−12.58	1.32	1.32	1.42
benzene, 1,2,3-trimethyl-	−9.6	−12.58	−2.98	2.98	0.42
benzene, 1,3,5-trimethyl-	−16.08	−12.58	3.50	3.50	1.10
benzene, 1-ethyl-2,4-dimethyl- (=1,3-dimethyl-5-ethylbenzene)	−35.59	−33.21	2.38	2.38	0.02
benzene, 1-ethyl-3,5-dimethyl-	−35.6	−33.21	2.39	2.39	0.01
benzene, 2-ethyl-1,3-dimethyl-	−29.8	−33.21	−3.41	3.41	0.01
1,3·dimethyl-4-ethylbenzene	−33.12	−33.21	−0.09	0.09	0.01
benzene, 2-ethyl-1,4-dimethyl-	−33.12	−33.21	−0.09	0.09	0.01
benzene, 3-ethyl-1,2-dimethyl-	−29.77	−33.21	−3.44	3.44	0.04
benzene, 4-ethyl-1,2-dimethyl-	−33.12	−33.21	−0.09	0.09	0.01
averaged absolute difference				1.97	0.30

3. Results

3.1. Alkyl-Substituted Alkanes: Introduction

In the previous paper [1], we presented results for the monomethylalkanes, which were very satisfactory with an ADD = 1.9 kJ/mol. At that point, we did not discuss the multiple substituted alkylalkanes because it was already clear at that moment that the simple scheme of the monomethylalkanes did not apply to the higher substituted species, i.e., it did not provide satisfactory results. Going from the di- to the tri- to the tetra-methyl-substituted alkanes, the difference between model and experiment became ever larger with values up to 40 kJ/mol for specific tetra-methyl-substituted alkanes. This, however, cannot be a surprise when studying the chemical literature and, more specifically, geminal (the relationship between two atoms or functional groups that are attached to the same atom) effects and steric effects on the entire molecular conformational structure. In this context, two papers by Rüchardt and Beckhaus [16,17] and references therein to their earlier published experimental data are very worthwhile reading in this context. They also referenced back to them in this context the pioneering work of Karl Ziegler from the 1940s on factors determining C-C bond strengths. In more detail, the steric effects

of two neighboring methyl substituents, i.e., as in -C-C(CH_3)-C(CH_3)-C-C-, may initially be regarded as just a van der Waals interaction between such groups. However, what is disregarded and specifically emphasized by Rüchardt and Beckhaus [16,17] is the effect that the presence of the substituents has on the conformational structure of the backbone alkane. Whereas an unsubstituted alkane can be drawn as an all-trans C-C chain, the substituents induce changes in C-C bond lengths, valence (CCC), and torsional (CCCC) angles, the extent depending on the degree, type, and location of the substitutions. Rüchardt and Beckhaus have very nicely formulated [16] what we need to pay attention to and why: "this simple additive and hence rigid structural model rapidly reaches its limits when used to explain steric effects on reactivity. The deliberately pragmatic concept "steric effects" incorporates phenomena arising from individual structural properties of the reactants and the activated complexes. If the result of a steric effect is a change in the activation enthalpy, it can be described by the model quantity "strain enthalpy Hs". This is defined as the difference between the enthalpy of formation of a real molecule, $H_f(g)$, and a norm or standard value H_f calculated from group increments". As different arrangements of substituents induce different, and in particular, non-additive effects on the backbone carbon chain, the pure group contribution approach with its linear additive terms does not apply anymore. A model that accounts for these effects would involve a specific additional parameter for each type of different configuration of substituents. This would be more or less similar when accounting for ring strain in the cyclic alkanes. However, whereas in practice, the cyclic alkanes form a limited number of different structures, there is a shear infinite number of alkyl-substituted alkanes with potentially all individual strain energies, and, therefore, one, at least formally, needs as many additional parameters. Consequently, we will describe the heats of formation by introducing a limited set of additional nearest or next-nearest neighbor GC parameters similar as we practiced for the substituted benzenes [1].

3.1.1. Mono-Methyl Alkanes

During the investigations that led to the present paper, and while having studied the higher substituted alkanes, di-, tri-, and tetra-alkylalkanes, we found that also for the monomethylalkanes geminal effects play a role, even though limited regarding the magnitude. When one studies the optimized molecular structures for the monomethylalkanes after energy minimization employing DFT calculations, one does observe changes in CCC bond angles along the main alkyl chain, making the entire chain a little bent (such as described by Rüchardt and Beckhaus [16,17]). In the work of [1], we had the formulas

$$\Delta H_f \text{ (mono-methyl alkanes)} = 3 * GC_{CH3} + N_{CH2} * GC_{CH2} + N_{CH} * GC_{CH} \qquad (3)$$

where GC_i is the contribution of the group i to the heat of formation and N_i is the number of times this group is present in the molecule. For the 2-methylalkanes, we have a potential geminal effect due to the interaction of two methyl groups: the 2-methyl group and the terminal methyl group. When we apply a correction of −2.1 kJ/mol to the 2-methylalkanes, we obtain the results shown in Table S1, revealing that the averaged absolute difference between model and experiment decreases from 1.91 [1] to 1.08 kJ/mol. Moreover, now all individual values are within chemical accuracy, whereas beforehand [1], the value for 2-methylnonane was deviating by 5.3 kJ/mol. Actually, the correction factor (−2.1 kJ/mol) can be directly obtained from the difference between the heats of formation of the 2-methyl- and the corresponding 3-methylalkanes. As a result, we arrive at the modified model equation for the 2-methylalkanes

$$\Delta H_f \text{ (2-methylalkanes)} = 3 * GC_{CH3} + N_{CH2} * GC_{CH2} + N_{CH} * GC_{CH} - 2.1 \qquad (4a)$$

and for the other (3-,4-, …) methylalkanes what we already had in reference [1]

$$\Delta H_f \text{ (mono-methyl alkanes)} = 3 * GC_{CH3} + N_{CH2} * GC_{CH2} + N_{CH} * GC_{CH} \qquad (4b)$$

3.1.2. Dimethyl Alkanes

For the dimethylalkanes, we initially attempted to describe the heat of formation by

$$\Delta H_f \text{ (n,n'-dimethyl alkanes)} = 4 * GC_{CH3} + N_{CH2} * GC_{CH2} + 2 * GC_{CH} \quad (5a)$$

for the n,n'-dimethylalkanes, and for the n,n-dimethylalkanes

$$\Delta H_f \text{ (n,n-dimethyl alkanes)} = 4 * GC_{CH3} + N_{CH2} * GC_{CH2} + 1 * GC_C \quad (5b)$$

These are the logical extensions of the equations that were applied in the work of [1]. The averaged absolute difference between model and experimental values was found to be 4.77 kJ/mol with about 50% of the individual values beyond chemical accuracy, with the largest deviation being 12 kJ/mol for 3,3-dimethylhexane. All data have been collected in Table S2 columns 4–6. However, based on the discussion in Section 3.1, we add corrective terms accounting for nearest neigbour interactions and changes in the main-chain conformation. A substitution at the 2 position was accompanied by a (geminal) neighbor effect (the neighboring methyl groups) of −2.1 kJ/mol. When we translate this to the dimethylalkanes, we expect an additional contribution to our model of −2.1 kJ/mol for 2,4-dimethylhexane and 2,6-dimethyloctane, and the results shown in Table S2 indeed confirm the better agreement with experimental data. Species including 2,4-dimethylpentane, 2,5-dimethylhexane, 2,6-dimethylheptane, and 2,7-dimethyloctane have two near terminal methyl groups well separated, and we, therefore, count these as two 2-positions and consequently a correction of −4.2 kJ/mol, with the results showing the clear improvement compared to the corresponding experimental data.

For the 2,3-methyl-substituted species, we, as before, adopt a neighbor effect for the 2-methyl with the terminal methyl of −2.1 kJ/mol, and when we then adopt +3.8 kJ/mol for the interaction between the neighboring 2- and the 3-methyl groups, we see suitable agreement between model and experiment, see 2,3-dimethylbutane, 2,3-dimethylbutane, and 2,3-dimethylbutane in Table S2. For 3,4-dimethylhexane, we only have the +3.8 kJ/mol neighbor effect for two neighboring methyl groups in the chain, whereas for 2,4-dimethylhexane and 2,6-dimethyloctane, we only have the −2.1 kJ/mol contribution due to the geminal effect between the 2-methyl and the terminal methyl group.

Finally, we deal with the n,n-methylsubstituted species. For the 2,2-dimethylalkanes, we have an effect of −4.2 kJ/mol due to the interactions of the two methyls with the terminal methyl groups. Furthermore, after careful analysis of the data, we need to introduce a further +11 kJ/mol contribution accounting for the overall effect due to two methyl groups attached to the same (main-chain) carbon. These contributions together lead to an +6.8 kJ/mol substitution contribution (accounting for both methyl-methyl vanderWaals interactions and the resulting conformational changes). Consequently, for 3,3-dimethylpentane and 3,3-dimethylhexane, the neighbor interaction to be added is only the new correction +11 kJ/mol.

The dimethylalkanes are still relatively unstrained in the sense as discussed by Rüchardt and Beckhaus [16,17]. This explains why the results shown in Table S2 are an indication that by adding moderate corrections, one can obtain excellent agreement between model and experimental results. It is interesting to note that the "correction" we introduced for 2,2-dimethylpentane is 6.8 kJ/mol, whereas Rüchardt and Beckhaus (case 9d, Table 2 in reference [17]) have calculated the strain energy is 6.7 kJ/mol, an agreement suggesting that our corrections are the strain enthalpies assigned by Rüchardt and Beckhaus.

In summary, from the data presented in Table S2, "model − experiment + neighbour interaction correction", we see that we finally reach an averaged absolute difference of 1.08 kJ/mol and all individual differences within chemical accuracy. For a 2-methyl substitution, we have added −2.1 kJ/mol for two neighboring methyl groups along the alkyl chain +3.8 kJ/mol, and n,n-disubstituted methyls an additional contribution of +11 kJ/mol. As we will see in the next section, these corrections will be found successfully applicable

in many of the higher substituted alkanes. We introduced three additional parameters to obtain overall chemical accuracy for all individual species. When we briefly make some comparison to the MG approach and the results from the ICAS23 software, the latter reveals an almost 6 kJ/mol energy difference with Rossini's experimental value for 2,3-dimethylpentane, thus beyond chemical energy, even though two second-order groups were involved. For the complete set of dimethylalkanes in Table S2, the ICAS23 results lead to an averaged absolute difference of 1.72 kJ/mol and three species showing an error beyond chemical accuracy, even though various second-order group contribution parameters were involved. Surprisingly, in particular, when considering the discussion presented by Rüchardt and Beckhaus, for 3,3-dimethylpentane and 3,3-dimethylhexane, only first-order GC parameters were involved (CH_3, CH_2, C). This illustrates that not taking into account known chemistry or physical knowledge leads to incorrect parametrization and consequently lower predictability for unknowns, i.e., species not used for the parametrization of the GC parameters.

3.1.3. Tri-, Tetra-, and Pentamethyl Alkanes

Trimethylalkane experimental data are shown in Table S3. The actual values for the neighbor interactions (corrections) were adopted from the previous section. For example, for 2,2,3-trimethylpentane, we have twice the correction −2.1 kJ/mol for the geminal 2-methyl − terminal methyl group interactions, twice the +3.8 kJ/mol in-chain methyl-methyl interactions, and finally +11 kJ/mol for the interactions between the 2,2 methyl groups. We observe, viz. Table S3, that many of the model values, when including the named corrections, agree within chemical accuracy with the experimental values, and the averaged absolute deviation is 3.86 kJ/mol. Still, for several other species, the mismatch is up to 8 kJ/mol. Even though this is, as overall behavior, slightly better than previously published results [6,8] where maximum deviations of 10 kJ/mol and more were present, we were aiming for a model exhibiting chemical accuracy throughout which we do not yet achieve. The experimental error is estimated (see Rossini c.s.) between 1 and 1.5 kJ/mol, which is too small to account for the larger deviations for 2,2,4-trimethylpentane, 2,2,4-trimethylhexane, and 2,4,4-trimethylhexane. When we inspect the 3-dimensional structures of these three exceptions, all deviating by about 8 kJ/mol from our model, with the model exhibiting the more negative value, the reason for this deviation becomes clear. In the Scheme 2, from the Structures (5) and (6) we see that methyl substituents at the neighboring main-chain carbon are on different sides of the main-chain, viz. the left-hand 2,2,3-trimethylhexane structure, whereas for the 2,2,4-trimethylhexane (right-hand structure), there is a more close interaction between the methyl substituents at different hexane (main-chain) carbons. This is even more clear when one looks at the corresponding space-filling models.

Structure (5)   Structure (6)
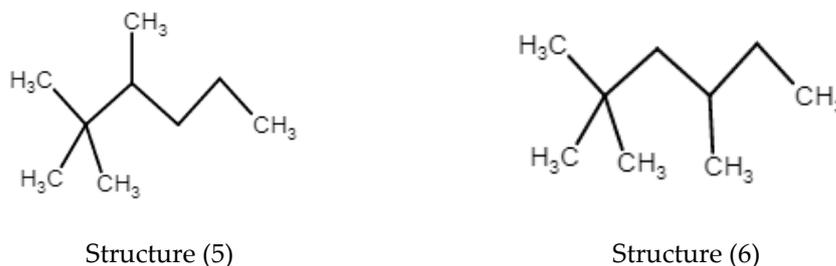
Scheme 2. Chemical Structures to illustrate different methyl-methyl interactions as discussed in the text. Structure (5): 2,2,3-trimethylhexane structure, Structure (6): 2,2,4-trimethylhexane structure.

In addition, we have performed quantum calculations involving optimization (minimizing the total energy) of these various structures. Here we noticed that the main chain is more seriously deformed (valence and torsional angles) compared to the free unsubstituted hexane in the case of the right-hand structure. In addition, the van der Waals overlap between the methyl substituents in these structures implies a destabilizing energy contri-

bution. Consequently, we introduce a further correction parameter, +8.5 kJ/mol, a value slightly larger than the deviations revealed for the trimethylalkanes but more appropriate when we look at the tetramethylalkanes below. This correction parameter can be named Me-Me-1,5-interaction. Importantly, we note that this correction is only involved when one of the main-chain carbon atoms is doubly methyl-substituted because, in the case of single substitution, the methyl groups can avoid van der Waals overlap.

When we now turn to the tetra-methyl-substituted alkanes, data collected in Table S4, we also see that some values are within chemical accuracy after adding the corrections discussed earlier, whereas other values are not, viz. column entitled "model-exp + correction". In particular, the value for 2,2,4,4-tetramethylpentane is off by 22 kJ/mol. Interestingly, Rüchardt and Beckhaus (case 9e, Table 2 in reference [17]) have calculated the strain energy for this species to be 31 kJ/mol, which is well within the deviation we observe (35 kJ/mol, column model-exp). We repeat that the strain energy is not only methyl-methyl interaction energies (the corrections we applied, final columns in Table S4) but also includes changes in the geometric parameters of a molecule. Now considering the Me-Me-1,5-interaction for the appropriate case, i.e., 2,2,3,4-tetramethylpentane and 2,2,4,4-tetramethylpentane (twice), we arrive at the final model values as shown in the last column of Table S4. The averaged absolute deviation reads 3.27 kJ/mol, and only 2,2,3,3-tetramethylpentane is the clear deviation from chemical accuracy.

The higher the degree of substitution, the fewer solid experimental data are available. Table S5 shows the results for only two pentamethyl alkanes, and the final model values are beyond chemical accuracy. This indicates that conformational changes as induced by a higher degree of substitution cannot be sufficiently, in view of the chemical accuracy we are requiring, mimicked by a few parameters with fixed (non-continuous) values.

3.1.4. Methyl-Ethyl and (iso)Propyl-Alkanes

To conclude the alkyl-substituted alkanes, when an ethyl group is involved (experimental data from Prosen and Rossini [18] and Labbauf et al. [19]), we clearly observe that for a mono-ethyl-substituted alkane, viz. the first six entries in Table S6, the standard model for mono-substituted alkanes (Equation (1)) applies reasonably well (no nearest neighbor corrections involved) with all values within chemical accuracy from the experimental values (column 4 in Table S6). For all multiple substituted alkanes in Table S6, the deviations are larger; however, after the nearest neighbor corrections we introduced earlier, most improve significantly, see the column entitled "model –exp(erimental value) + correction". The corrections introduced thus far are related to methyl groups and methyl-methyl interactions only. When we now add an additional term related to an ethyl group substitution with a magnitude +6 kJ/mol, we arrive at the values in the last two columns of Table S6, which are now all within chemical accuracy, whereas the overall averaged absolute deviation is 1.62 kJ/mol. Next, when we apply the same correction to propyl and isopropyl substituents, we see that we also obtain results within chemical accuracy. Based on these observations, based on the available experimental data, we assume that this will likely apply to more of the larger alkyl substituents. Further experimental data should further clarify this.

When we look at the last three entries in Table S6, we see that for the tetrasubstituted ethylalkanes, the results are similar to the tetramethylalkanes in the previous section. Most are within or close to chemical accuracy, 3,3,4,4-tetraethylhexane is very much beyond, and this species was excluded by Verevkin et al. [9] because of non-additive behavior. The GC model of Verevkin et al. suggested −314.8 kJ/mol. Our model, according to a straightforward application of a formula similar to Equation (1) with CH, CH_2, and CH_3 GC parameters only predicted a heat of formation of −379.9 kJ/mol, which differs from the experimental value by 115.6 kJ/mol with the understanding that in reality, the experimental structure is less stable than what the model suggests. When we adopt the corrections that we introduced in the previous section, the difference reduces to 54 kJ/mol, which is still much larger than chemical accuracy. In addition, the MG [4]/ICAS [5] approach yields a

value, -336.9 kJ/mol, much of the experimental value, and with the peculiarity that one third-order term is involved, which is named $(CH_3)_3C\text{-}(CH_2)_m\text{-}CH_m(CH_3)_n$ m in 0, 1, n in 2, 3 whereas no t-butyl unit $(CH_3)_3C$ can be recognized in the structure.

To obtain more insight into the issue, we have considered the two structures 3,3,4,4-tetraethylhexane (7) and 2,2,5,5-tetraethylhexane (8), Scheme 3. Applying the standard GC model with CH_3, CH_2, and CH parameters only, the heat of formation of both structures are identical. By applying quantum calculations (B3LYP//6-311+G**), we found a difference of 112 kJ/mol, which can be compared to the 115.6 kJ/mol quoted above. This great agreement is inevitably partly fortuitous. Still, it is the explanation that interactions between the four branches, with partly interpenetrating van der Waals radii, are very significant and cause a breakdown of the simple additive GC model, and we need to accept that the heat of formation of structures such as 3,3,4,4-tetraethylhexane cannot be described within any reasonable accuracy by the GC method.

Structure (7)   Structure (8)
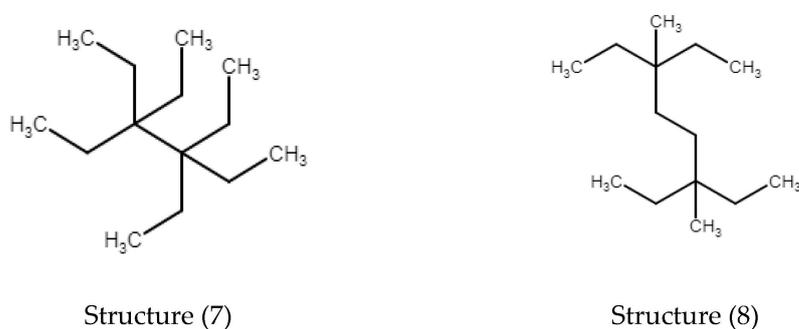
Scheme 3. Chemical Structures for 3,3,4,4-tetraethylhexane (7) and 2,2,5,5-tetraethylhexane (8) illustrating differences in methyl-methyl interactions as discussed in the text.

3.1.5. Summary Alkylalkanes

The heats of formation of the mono- di- and tri-substituted alkylalkanes can be appropriately described within chemical accuracy applying the proposed GC parametrization. For the tetra- and penta-substituted alkylalkanes, part of the results for the species considered is within chemical accuracy. Some are less (pentamethylpentanes), whereas 3,3,4,4-tetraethylhexane is truly far beyond. In order to account for the steric and conformational effects when substituents are present, we have introduced five additional nearest neighbor parameters to account for substitution effects in the alkylalkanes. All methyl-related corrections are illustrated in Scheme 4 below. For a 2-methyl substitution (geminal effect), we add -2.1 kJ/mol, for two neighboring methyl groups along the alkyl chain $+3.8$ kJ/mol, for n,n-disubstituted methyls an additional contribution of $+11$ kJ/mol, in case an ethyl or (iso-)propyl group is present an additional correction of $+6$ kJ/mol applies, and finally a 1,5-interaction term (as in 2,4-disubstituted alkylalkanes) of $+8.5$ kJ/mol. For the monomethylalkanes, this led to an improvement over our previous findings [1] with an averaged absolute deviation (AAD) of 1.08 kJ/mol, whereas the previous value was 1.91 kJ/mol. For the dimethylalkanes, we also obtained an AAD of 1.08 kJ/mol. For the tri- and tetrasubstituted methylalkanes, we obtained an ADD of 1.18 and 3.27 kJ/mol, respectively. For the ethylalkanes and a few (iso)propylalkanes, we observed the same trends as for the methylalkanes: for the 1-, 2-, and 3-substituted species, we observed suitable agreement within chemical accuracy, whereas for the tetrasubstituted species, deviations are larger and generally beyond chemical accuracy.

When we compare the new results to the up till now as claimed best results, i.e., the MG approach [4] as implemented in ICAS23 [5,6], for the mono- up till tri-substituted ethyl alkanes, the latter yields an AAD of 5.4 kJ/mol with 12 of the 30 entries having a predicted heat of formation beyond chemical accuracy, the largest deviation being 13.6 kJ/mol, and this twice, whereas our new model yielded an ADD of 1.62 kJ/mol and all individual values within chemical accuracy (see Table S6).
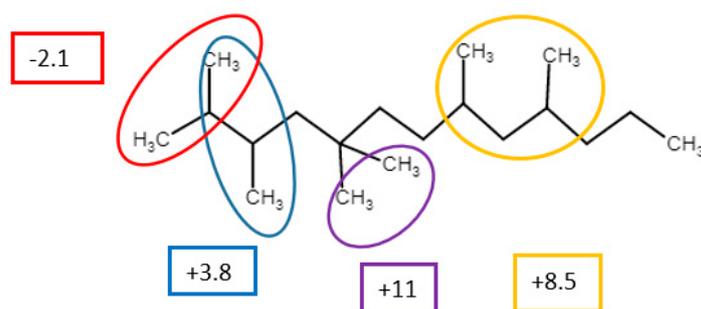
Scheme 4. Illustration of the corrections accounting for methyl-methyl interactions we have introduced (for details see text).

Whereas we defined interaction parameters that we related to neighbor effects and enabled us to describe the heat of formation for almost all species with chemical accuracy, it needs to be emphasized, very importantly, that each substitution pattern has its own characteristics, in particular its specific effect on valence and torsional angles. These cannot be generally described by an interaction parameter with default and fixed numerical value. Of course, this is all partly related to the fact that we want to achieve chemical accuracy in predictions. The agreement for, e.g., the pentamethylpentanes and the trimethyl-ethyl-alkanes can most likely be improved by accounting for the species individual changes that were induced due to substitution. This, however, would lead to very individual parameters for an individual species, which leads to a very large number of parameters and destroys the idea of a method that potentially can predict the heat of formation for large classes of molecules. However, then it must be accepted that the heats of formation for some of the highly substituted alkanes can not be appropriately described using a linear additive method such as the group contribution method.

In this context, it is also interesting and relevant to refer to what Smith [20] stated some two decades ago. He came to the same conclusion for the higher substituted alkanes as Rüchardt and Beckhaus [16,17]. In his paper, Smith argued that the thermochemistry of alkanes is dominated by the energies of the HOMOs (higher occupied molecular orbitals), more specifically, C-C π antibonding MOs (molecular orbitals). It was concluded that "the large discrepancies for, e.g., 2,2,3-trimethylbutane or 2,2,3,3-tetramethylbutane, show that there is no simple linear relationship". At the end of his paper, Smith wrote what we think is still ever so true even 20 years after Smith's publication: "Even in the age of supercomputers capable of tackling enormous basis sets, back-of-the-envelope calculations are not to be despised; they have the not inconsiderable merit that it is possible to follow the working, and hence to rationalise the results. Analyses such as those presented here may help to build bridges between purely empirical methods such as Benson's group additivity scheme and the most accurate quantum mechanical computations".

3.2. Alkyl-Alcohols

In the work of [1], only linear primary and secondary alcohols were considered, with most experimental data taken from the NIST database, which come from different original sources. In the present study, we use a more extensive collection of data on alcohols originating from a single group well known for its experimental capabilities and that also entails methyl branched alcohols [21], leading to about twice the number of species involved compared to the set considered in the work of [1] and, more importantly, with greater variety (methyl branched). We revisited the GC parameters, now adopting distinct parameters for a primary and for a secondary OH group.

The results are collected in Table S7. The AAD for the non-branched species is 1.40 kJ/mol with a single deviation beyond chemical accuracy: 6.57 kJ/mol for 4-octanol. Verevkin c.s. have given an error of 2.5 kJ/mol for the enthalpy of formation for this species, clearly higher than for all other primary and secondary alcohols, which potentially brings the value within chemical accuracy. It could also be [private communication

Prof. S.Verevkin], however, that a spurious amount of water present in the sample causes a deviation, as a correction for the presence of water would lead to a more negative value, which is in agreement with our more negative model value (−375.4 kJ/mol) compared to the experimental value (−368.8 kJ/mol). For the (m)ethyl branched species, the AAD equals 1.37 kJ/mol, and all individual values are within chemical accuracy.

Similar to the highly branched alkanes, the higher substituted species (last three entries in Table S7) sometimes reveal high deviations indicating these cannot be appropriately dealt with simple linear GC models, even though we already added the corrections we introduced for the branched alkanes (see above). Whereas our current model reveals differences of 2, 1, and 33 kJ/mol for these three species, the ICAS23 model [5] yields deviations of 12, 34, and 97 kJ/mol, whereas Verevkin c.s. [21] have reported 0.4, 19, and 48 kJ/mol. So our new model only totally fails for 2,2,4,4-tetramethyl-3-iPr-3-pentanol.

3.3. Alkylethers

Whereas previous methods generally provide suitable results for this class of compounds, one still observes deviations and also unexpected behavior. One of these is the Marrero-Gani method [4,5] as implemented in ICAS23 [5]. Examples include 6 kJ/mol for di-s-butylether, 11 kJ/mol for t-butyl s-butylether, whereas interestingly di- t-butylether is predicted very well (although higher branched), whereas both ethyl and butyl t-amylether are somewhat beyond chemical accuracy (4.5 and 5.7 kJ/mol, respectively). When we look more closely into the GC parameters, the three aforementioned s- and t-butylethers were all described by first-order [5] GC parameters only, whereas the t-amyl ethers involved a third-order parameter, which is somewhat counterintuitive. Verevkin [22] had very excellent results for almost all species using a GC-type model, though, e.g., tBu-O-sBu ether was off by 16.5 kJ/mol and tBu-O-tBu was off by 22.5 kJ/mol. Consequently,0 there is a reason to reevaluate this class, also because we want to develop a GC method that applies to a broad range of classes of molecules such as the MG-ICAS approach [4,5].

In reference [1], we considered a rather limited data set originating from (mainly) NIST and CAPEC database values for the gas phase heat of formation. For the present work, we adopted a more extensive and consistent list of experimental data was retrieved from Verevkin [22]. From the data in Table 5 in Ref [22] (pages 1083 and 1084), inspection of the series tBu-O-n-alkyl and the series t-amyl-O-n-alkyl show that the energy difference between the methyl species (e.g., tBu-O-Me) and the related ethyl or higher species is significantly higher (around 30 kJ/mol) than between the ethyl and higher species which more closely reflect the typical increment related to a CH_2 group (around −21 kJ/mol). Consequently, this should be reflected in the GC parameters if we want to achieve chemical accuracy. In addition, we have to account for steric effects in congested species.

In Table 2, experimental values have been listed along without model values. This table consists of two parts comprising the Me-OC-X series and the X-COC-Y series, respectively, with X and Y being various alkyl substituents. Because of the noted differences between Me or higher alkyl substitution when using the same GC parameters for COC and the formerly determined parameter for CH_2, a dedicated GC parameter was adopted for dimethylether Me-O-Me, which then obviously equals the experimental value 184.1 kJ/mol. With increasing steric factors and induced changes in the structure of the basic structure, in particular, but not exclusively the COC valence angle, we adopted different GC parameters for the COC structural element depending on the substitution pattern. This is what can be seen from the one but last column in Table 2. Note that molecules that were assigned a specific GC parameter for the COC entity all have COC angles that are very close or even identical; see the last column in Table 2. In accordance with qualitative expectations, the larger the substituents, the less negative the COC group contribution or, in other words, the less stable the species. For the Me-COC-X, these vary from 184.1 kJ/mol for Me-O-Me to −156 kJ/mol for the tertiary carbon substituents (tBu and t-amyl). The same trend is observed in the second part of Table 2, involving the X-CXOC-Y species. Overall we obtain very suitable results, with all individual model values within chemical accuracy from the

experimental values, an ADD of 0.83 kJ/mol for the methylalkyl ethers and 1.94 kJ/mol for the other ethers. These results look logical based on the formerly discussed considerations on the impact of increasing steric crowdedness, including the effect on the conformational parameters of the backbone, e.g., the COC valence angle (last column in Table 2) but also torsional angles. The various GC parameter values were, thus, determined on the basis of careful analysis of the experimental data along with the chemical structures.

Table 2. Experimental and model values for various ethers. All values in kJ/mol. Experimental data were taken from Verevkin [22]. The model values dHf were calculated using the GC formula for the methylalkylethers ΔH_f (methylalkylethers) = GC_{CH3} + N_{CH2} * GC_{CH2} + $GC_{(H3)COC(-)}$ (see Reference [1]). However, to obtain chemical accuracy, the last term is to be considered substitution-dependent. The effect of the substitution on the conformational details of the backbone of the ethers was found (see text) to be suitably accounted for by adding a correction parameter related to the C-O-C valence angle, see the last two columns. The red values are averaged absolute differences between the final model (for discussion, see text) and the experimental values.

Methyl-Alkyl-Ethers	Verevkin 2002	Model dHf	Model-Exp	ABS (Model-Exp)	Ether Group Constitution	GC Value Ether Group	COC Valence Angle
dimethylether	−184.1	−184.1	0.00	0.00	Me-O-Me	−184.1	112.7
methyl ethyl ether	−216.4	−217.36	−0.96	0.96	Me-O-C-R	−175	113.1
methyl propyl ether	−238.4	−237.99	0.41	0.41	Me-O-C-R		113.1
methyl n-butyl ether	−258.3	−258.62	−0.32	0.32	Me-O-C-R		113.1
methyl decyl ether	−381.1	−382.4	−1.30	1.30	Me-O-C-R		113.1
methyl isopropyl ether	−252	−252.72	−0.72	0.72	Me-O-CRR′	−168	115.1
methyl t-butylether	−283.4	−282.08	1.32	1.32	Me-O-CRR′R″	−156	118.4
methyl t-amylether	−301.1	−302.71	−1.61	1.61	Me-O-CRR′R″		118.7
averaged absolute difference				0.83			
Di-Alkyl Ethers	**Verevkin 2002**	**Model dHf**	**Model-Exp**	**ABS (Model-Exp)**	**Ether Group Constitution**	**GC Value Ether Group**	**COC Valence Angle**
diethylether	−252.1	−252.72	−0.62	0.62	R-COC-R′	−168	113.5
ethyl propyl ether	−272.4	−273.35	−0.95	0.95	R-COC-R′		113.5
ethyl butyl ether		−293.98			R-COC-R′		113.5
di-n-propylether	−293.1	−293.98	−0.88	0.88	R-COC-R′		113.5
di-n-butylether	−332.9	−335.24	−2.34	2.34	R-COC-R′		113.7
di-n-pentylether	−380.4	−376.5	3.90	3.90	R-COC-R′		113.4
ethyl t-amylether	−333.5	−336.07	−2.57	2.57	R-COC-R′R″R‴	−146	119.4
butyl t-amylether	−375.7	−377.33	−1.63	1.63	R-COC-R′R″R‴		119.1
ethyl t-butylether	−316.8	−315.44	1.36	1.36	R-COC-R′R″R‴		118.8
propyl t-butylether	−339.3	−336.07	3.23	3.23	R-COC-R′R″R‴		118.7
n-butyl t-butylether	−360.1	−356.7	3.40	3.40	R-COC-R′R″R‴		118.6
amyl t-butylether	−380.6	−377.33	3.27	3.27	R-COC-R′R″R‴		118.6
di-i-propylether	−319.4	−318.44	0.96	0.96	RR′-COC-R″R‴	−149	116

Table 2. *Cont.*

Di-Alkyl Ethers	Verevkin 2002	Model dHf	Model-Exp	ABS (Model-Exp)	Ether Group Constitution	GC Value Ether Group	COC Valence Angle
di-sec-butylether	−361.3	−359.7	1.60	1.60	RR′-COC-R″R‴		116.5
t-butyl s-butylether	−379	−381.43	−2.43	2.43	RR′-COC-R″R‴R⁗	−149	119.8
t-butyl i-propylether	−360.1	−360.8	−0.70	0.70	RR′-COC-R″R‴R⁗		119.9
t-butyl i-butylether	−367.9	−364.8	3.10	3.10	RR′-COC-R″R‴R⁗		119
di- t-butylether	−361.2	−361.16	0.04	0.04	tBU-COC-tBu	−107	128
averaged absolute difference				1.94			

Still, we prefer to understand a little better and establish physico-chemical arguments. The magnitude of the substitution effects can be estimated from quantum calculations to evaluate energy differences on different but similar structures, focusing on the potential steric and conformational effects. We can perform these calculations on structures for which we have no experimental data but that have the same number and type of groups: CH_3, CH_2, CH, C, ether moiety. This enables us to obtain a fair impression of certain effects, in particular those with steric congestion. Such relative energies are known to be more reliable than calculations of absolute energy quantities. From quantum calculations (B3LYP//6-31+G* method), the COC angle for Structure (9) in Scheme 5 was calculated as 113.5°, or di-n-propylether we found 113.5°, for di-isopropylether 116.0°, and for sec-butyl isobutyl ether 115.8°, so all relatively close. For Structure (10) in Scheme 5, di-t-amylether, we obtained a value of 128.9° and thus significantly larger (di-t-buytlether has a very similar value). The computed energy difference between Structures (9) and (10) was found to be close to 29 kJ/mol (B3LYP//6-31+G* method), with Structure (9) being the more stable structure. When we take Structure (9) and constrain the COC angle to 128.9°, the COC angle associated with Structure (9), the energy difference is 20 kJ/mol. This suggests that the steric effects due to the four interacting Me-groups have significant effects on the backbone conformation and, in particular, by inducing a larger COC angle associated with 20 kJ/mol energy difference. Such effects are the ones we discussed before and were referred to by Rüchardt and Beckhaus [16].

Structure (9)   Structure (10)
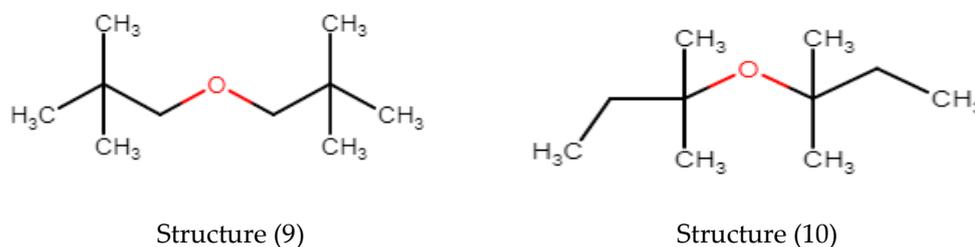
Scheme 5. Two structures revealing the difference in COC valence angle, as computed (see text), depending on the location of the methyl substituents. Structure (9) COC angle = 113.5°; Structure (10) COC angle = 128.9°.

3.4. Dienes

We have a limited number of reliable data on dienes, but a sufficient number to determine the relevant GC parameters and to make some interesting observations. First of all, we need to recall the different types of dienes, which are molecules with two double bonds.

- Double bonds are *isolated* if they are separated by two or more single bonds so that they cannot interact with each other;

- Double bonds are *conjugated* if they are separated by just one single bond. Because of the interaction between the double bonds, systems containing conjugated double bonds are more stable than similar systems with isolated double bonds;
- Successive double bonds with no intervening single bonds are called *cumulated* double bonds. Systems containing cumulated double bonds are less stable than similar systems with isolated double bonds.

In particular, in the case of conjugation, we need to consider neighbor effects when we take individual double bonds as groups. Very suitable results, i.e., all model values within chemical accuracy, can be obtained when we define the following individual groups and corrections:

terminal C=C- +62.5 kJ/mol
trans R-C=C-R' +73.5 kJ/mol
cis R-C=C-R' +78 kJ/mol
C=C=C +205 kJ/mol

and a conjugation correction of −17 kJ/mol for all *conjugated* species (up till and including 1,3-butadiene). Whereas these parameters generally provide suitable model values (within chemical accuracy from the experiment), initially, there were two exceptions, namely propadiene and 2,3-pentadiene. For propadiene, the reason is that we have defined C=C=C- as a group [1], but propadiene does not have a further bond; it is simply C=C=C. When we compare the group value −62.5 kJ/mol for C=C- with ethylene 52.4 kJ/mol, the C=C group value for the non-bonded species is 10 kJ/mol lower. When we apply such a correction to propadiene, in fact, we will adopt −11 kJ/mol as this is more consistent with what comes next (2,3-pentadiene), and it provides a better agreement between model and experimental value for propadiene, we obtain the suitable model value shown in Table 3. Similarly, for 2,3-pentadiene, the suitable value in Table 3 was obtained after an additional correction of +11 kJ/mol, which is the difference between the double bond group value of 62.5 kJ/mol previously established for 1-alkenes C=C- and 73.5 kJ/mol for disubstituted double bonds R-C=C-R'.

Table 3. Experimental [23–26] and model data for a set of dienes. The first three species comprise isolated double bonds, the next six contain conjugated double bonds, and finally, the last four reveal cumulative double bonds. The model values dHf were evaluated using the previously (reference [1]) established GC parameters for CH$_3$ and CH$_2$ groups and specific GC parameters for the three different type double bond containing species (isolated, conjugated, cumulative) as well as distinct parameters for the trans and cis forms in R-C=C-R'. The red value on the last row is the averaged absolute difference between model and experiment, 1.58 kJ/mol, whereas also all individual values are within chemical accuracy (1 kcal/mol or 4.2 kJ/mol).

Dienes	Rossini 1949	Prosen, Maron, and Rossini 1951	Fraser and Prosen 1955	Fang and Rogers 1992	Model dHf	Model-Exp	ABS (Model-Exp)
1,4-pentadiene	105.5		106.4		104.37	−2.03	2.03
1,4-hexadiene				77	73.01	−3.99	3.99
1,5-hexadiene				85	83.74	−1.26	1.26
cis-1,3-hexadiene	59				60.51	1.51	1.51
trans-1,3-hexadiene	54				56.01	2.01	2.01
cis 1,3-pentadiene, (Z)-			82.8		81.14	−1.66	1.66
trans 1,3-pentadiene			75.8		76.64	0.84	0.84
trans-2,4-hexadiene				44	45.28	1.28	1.28
1,3-butadiene	112	108.9			108	−0.90	0.9

Table 3. *Cont.*

Dienes	Rossini 1949	Prosen, Maron, and Rossini 1951	Fraser and Prosen 1955	Fang and Rogers 1992	Model dHf	Model-Exp	ABS (Model-Exp)
Propadiene	192.3				194	1.70	1.7
1,2-butadiene	165.6	162.3			162.64	0.34	0.34
1,2-pentadiene	145.7		140.85		142.01	1.16	1.16
2,3-pentadiene	138.6		133.1		131.28	−1.82	1.82
averaged absolute difference							1.58

4. Final Discussion and Conclusions

Our goal was to establish a group contribution approach revealing "chemical accuracy", i.e., maximum 1 kcal/mol (4.2 kJ/mol) difference between experimental and model values. Moreover, we wanted no or exceptionally few outliers, and if we understand why these are outliers, a goal we have achieved. Finally, overfitting is definitely to be avoided as this will affect the reliability of predictions.

Combining the results from the present and the previous paper [1], we accounted for the n-alkanes, mono-, di-, tri- tetra-, and penta-substituted methylalkanes, ethyl alkanes, various classes of alkenes, alkynes, primary and secondary alcohols, n-alkylamines, n-aldehydes, methyl- and di-alkylethers, 2-alkanones, mono- and di-carboxylic acids, dienes, mono- and di-nitriles, alkyl-substituted benzenes, alkyl-substituted naphthalenes, and alkyl-substituted cycloalkanes. Not only averaged absolute deviations but also individual results were within chemical accuracy, except for some more heavily alkyl-substituted molecules. These suitable results, when compared to other studies, were the result of taking into account the following components. We primarily used experimental data sets from a few reliable sources only. This proved to be one essential ingredient to the quality of the results, with more consistent CH_2-increments through the various classes of molecules. Secondly, we optimized class by class, so we first established parameters for CH_3 and CH_2 groups, etcetera. By doing so and determining the GC parameters by hand, specific trends and features became apparent, e.g., methyl versus other alkyl ethers. This way, we obtained uniquely defined and optimal GC parameters. The only exceptions were some more heavily alkyl-substituted species, for which we also discussed the origin. The heavier the substitutions, the more serious the influence on the conformation of the entire molecule; in particular, valence angles (CCC, COC) and torsional angles are affected. The effect on the heat of formation depends on the substitution pattern, and therewith a continuous parameter and not to be accounted for by a few additional GC parameters. We have to accept that this implies a breakdown of the group contribution concept. As long as we apply the GC model to molecules for which GC parameters could be established successfully, we still end up with an approach that is highly useful for practical applications requiring chemical accuracy.

model (model—exp). The red value 1.18 kJ/mol in the last row is the averaged absolute difference between the final model (for discussion see text) and the experimental values, Table S4: Experimental and model values for tetramethylalkanes. All values in kJ/mol. Rossini c.s. data [19] except for 2,2,5,5-tetramethylhexane (CAPEC data base) were used when comparing with the present model (model—exp). The red value 3.27 kJ/mol in the last row is the averaged absolute difference between the final model (for discussion see text) and the experimental values, Table S5: Experimental and model values for pentamethylalkanes. All values in kJ/mol. When available, Experimental data from Rossini c.s. [19] were used when comparing with the present model (model—exp). Corrections as described in the text, Table S6: Experimental and model values for ethylalkanes. All values in kJ/mol. Experimental data were taken from Rossini c.s., i.e., the first two entries from Prosen and Rossini [18] and all others from Labbauf et al. [19]. Upper table: mono up to tri-substituted alkanes. The red value 1.62 kJ/mol in the last row is the averaged absolute difference between the final model (for discussion see text) and the experimental values. Lower table: tetra-substituted alkanes, Table S7: Experimental and model values for various alcohols. All values in kJ/mol. Experimental data were taken from Verevkin [21]. The red values are averaged absolute differences between the final model (for discussion see text) and the experimental values.

References

1. Meier, R.J. Group Contribution Revisited: The Enthalpy of Formation of Organic Compounds with "Chemical Accuracy". *ChemEngineering* **2021**, *5*, 24. [CrossRef]
2. Van Krevelen, D.W.; Chermin, H.A.G. Estimation of the free enthalpy (Gibbs free energy) of formation of organic compounds from group contributions. *Chem. Eng. Sci.* **1951**, *1*, 66–80. [CrossRef]
3. Pijpers, A.P.; Meier, R.J. Core level photoelectron spectroscopy for polymer and catalyst characterisation. *Chem. Soc. Rev.* **1999**, *28*, 233–238. [CrossRef]
4. Marrero, J.; Gani, R. Group-contribution based estimation of pure component properties. *Fluid Phase Equilibria* **2001**, *183–184*, 183–208. [CrossRef]
5. Proped (Property Prediction) Module Version 4.7 in ICAS23. Available online: https://www.kt.dtu.dk/english/research/kt-consortium/software (accessed on 10 November 2021).
6. 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 Equilibria* **2013**, *348*, 23–32. [CrossRef]
7. Kadda, A.; Mustapha, B.A.; Yahiaoui, A.; Toubal Khaled, T.; Hadji, D. Enthalpy of Formation Modeling Using Third Order Group Contribution Technics and Calculation by DFT Method. *Int. J. Thermodyn.* **2020**, *23*, 34–41. [CrossRef]
8. Das, G.; dos Ramos, M.C.; McCabe, C. Accurately modeling benzene and alkylbenzenes using a group contribution based SAFT approach. *Fluid Phase Equilibria* **2014**, *362*, 242–251. [CrossRef]
9. 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*, 033102. [CrossRef]
10. Mathieu, D. Atom Pair Contribution Method: Fast and General Procedure to Predict Molecular Formation Enthalpies. *J. Chem. Inf. Model.* **2018**, *58*, 12–26. [CrossRef]
11. Available online: https://webbook.nist.gov/ (accessed on 10 November 2021).
12. Nielsen, T.; Abildskov, J.; Harper, P.; Papaeconomou, I.; Gani, R. The CAPEC Data Base. *J. Chem. Eng. Data* **2001**, *46*, 1041–1044. [CrossRef]
13. *Spartan '10*; Wavefunction Inc.: Irvine, CA, USA. Available online: www.wavefun.com (accessed on 10 November 2021).

14. Prosen, E.J.; Johnson, W.H.; Rossini, F.D. Heats of combustion and formation at 25 °C of the alkylbenzenes through C10H14. and of the higher normal monoalkylbenzenes. *J. Res. Natl. Bur. Stand.* **1946**, *36*, 455–461. [CrossRef]

15. Hukkerikar, A.S. Development of Pure Component Property Models for Chemical Product-Process Design and Analysis. Ph.D. Thesis, Danish Technical University (DTU), Kgs. Lyngby, Denmark, 2013. Available online: https://backend.orbit.dtu.dk/ws/portalfiles/portal/59650003/Amol+S.+Hukkerikar_PEC13-42.pdf (accessed on 10 November 2021).

16. Rüchardt, C.; Beckhaus, H.-D. Consequences of Strain for the Structure of Aliphatic Molecules. *Angew. Chem. Int. Ed. Engl.* **1985**, *24*, 529–538. [CrossRef]

17. Rüchardt, C.; Beckhaus, H.-D. Steric and Electronic Substituent Effects on the Carbon-Carbon Bond. *Top. Curr. Chem.* **1986**, *130*, 1–22.

18. Prosen, E.J.; Rossini, F.D. Heats of combustion and formation of the paraffin hydrocarbons at 25 °C. *J. Res. Natl. Bur. Stand.* **1945**, *34*, 263–269. [CrossRef]

19. Labbauf, A.; Greenshields, J.B.; Rossini, F.D. Heats of formation, combustion, and vaporization of the 35 nonanes and 75 decanes. *J. Chem. Eng. Data* **1961**, *6*, 261–263. [CrossRef]

20. Smith, D.W. Carbon.carbon π-antibonding effects on the thermochemistry of alkanes, elucidated by angular overlap and MO calculations. *Phys. Chem. Chem. Phys.* **2001**, *3*, 3562–3568. [CrossRef]

21. Roganov, G.N.; Pisarev, P.N.; Emel'yanenko, V.N.; Verevkin, S.P. Measurement and Prediction of Thermochemical Properties. Improved Benson-Type Increments for the Estimation of Enthalpies of Vaporization and Standard Enthalpies of Formation of Aliphatic Alcohols. *J. Chem. Eng. Data* **2005**, *50*, 1114–1124. [CrossRef]

22. Verevkin, S.P. Improved Benson Increments for the Estimation of Standard Enthalpies of Formation and Enthalpies of Vaporization of Alkyl Ethers, Acetals, Ketals, and Ortho Esters. *J. Chem. Eng. Data* **2002**, *47*, 1071–1097. [CrossRef]

23. Kilpatrick, J.E.; Beckett, C.W.; Prosen, E.; Pitzer, K.S.; Rossini, F.D. Heats, Equilibrium Constants, and Free Energies of Formation of the C3 to C5 Diolefins, Styrene, and the Methylstyrenes. *J. Res. Natl. Bur. Stand.* **1949**, *42*, 225–240. [CrossRef]

24. Prosen, E.J.; Maron, F.W.; Rossini, F.D. Heats of combustion, formation, and insomerization of ten C_4 hydrocarbons. *J. Res. Natl. Bur. Stand.* **1951**, *46*, 106–112. [CrossRef]

25. Fraser, F.M.; Prosen, E.J. Heats of Combustion and Isomerization of Six Pentadienes and Spiropentane. *J. Res. Natl. Bur. Stand.* **1955**, *54*, 143–148. [CrossRef]

26. Fang, W.; Rogers, D.W. Enthalpy of hydrogenation of the hexadienes and cis- and trans-1,3,5-hexatriene. *J. Org. Chem.* **1992**, *57*, 2294–2297. [CrossRef]