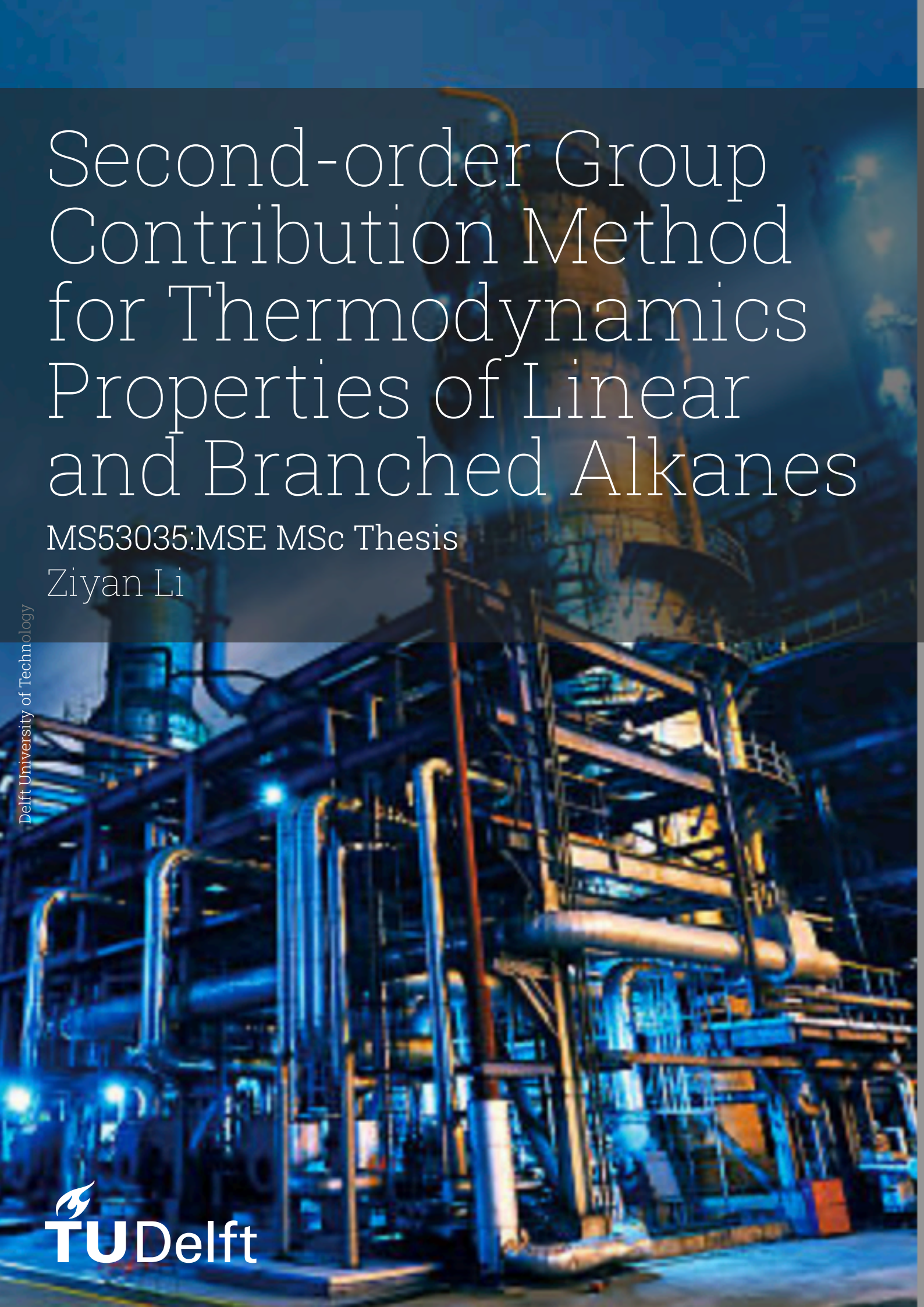
Second-order Group Contribution Method for Thermodynamics Properties of Linear and Branched Alkanes

MS53035:MSE MSc Thesis

Ziyan Li

Delft University of Technology



TU Delft

Review of Group Contribution Methods for Prediction of Thermodynamic Properties of Long-chain Hydrocarbons

by

Ziyan Li

Abstract:
This review investigates various group contribution methods (GCMs) used for the prediction of thermodynamic properties of long-chain hydrocarbons. The discussion includes first order and second order GCMs, hybrid techniques, and emerging machine learning approaches. The performance, advantages, and limitations of GCMs are critically analyzed based on recent literature.

Instructor: Prof.dr.ir. T.J.H. Vlugt
Dr. P. Dey
Faculty: Faculty of Mechanical Engineering

TU Delft

1. Introduction

With the increasing demand for transition to renewable energy and products, the prediction of thermodynamic properties of organic chemicals is critical for the design, simulation, and optimization of chemical processes as well as the design of new products with improved environmental properties [1]. Accurate and reliable property predictions are crucial for applications ranging from reaction and separation processes to energy storage and sustainable chemical production [2, 3]. For example, sustainable aviation fuel (SAF), as an alternative for reducing greenhouse gas emissions from the aviation sector, requires various production pathways, such as hydroisomerization and hydrocracking of long-chain alkanes, for tailoring the properties to meet stringent performance and environmental requirements [4]. To enable accurate modeling and process design for SAF production, it is essential to understand the thermodynamic properties of both linear and branched hydrocarbons involved in these catalytic transformations [5]. The experimental determination of thermodynamic properties of branched hydrocarbons can be costly, time-intensive, and impractical for many compounds, particularly for large branched molecules [6].

Group contribution methods (GCMs) have become a widely-used approach for estimating thermodynamic properties based on the total summation of contributions of molecular substructures [7]. In these methods, the properties of a compound are estimated as a summation of the contributions of simple aspects of the structural groups, which are named as first order groups. In this way, properties can be gained fast by just examining the molecular structure. GCMs provide the important advantage of quick estimates without requiring substantial computational resources [8]. These methods have evolved significantly and incorporated first order and more complex groups, namely higher order groups, to enhance accuracy and account for molecular complexities such as branching and isomerism [6]. GCMs are versatile and capable of predicting a wide range of thermodynamic and transport properties. Commonly estimated properties are the critical temperature (T_c), pressure (P_c), volume (V_c), boiling (T_b) and melting points (T_m), Gibbs energies of formation ($G^0 - H^0(0\,\mathrm{K})$), enthalpies of formation ($H^0 - H^0(0\,\mathrm{K})$), absolute entropies (S^0), Gibbs free energies of formation (ΔG_f), enthalpies of formation (ΔH_f), and constant pressure heat capacities (c_p^0). The temperature over which GCMs can be applied normally covers a range of 298 to 1500K but GCMs made specifically for temperature outside this range are scarce as GCMs are developed based on experimental data [9].

Despite the success in short-chain organic compounds and mixtures, classical methods, in which a compound is estimated as a summation of the contributions of first order groups that can occur in the molecular structure, often face limitations in predicting properties for highly branched compounds, long-chain molecules, or systems with incomplete group libraries [10]. Recent advancements [11–13] have focused on improving these methods by refining group contributions, integrating statistical models, and enhancing computational power. A key innovation has been the development of approaches that use higher order group interactions, which refers to interactions beyond isolated functional group like the influence of neighboring atoms or branches on a central group, to capture molecular intricacies more effectively and accurately [5, 8]. These methods have demonstrated enhanced accuracy for complex systems, often exceeding the precision of earlier models, while reducing the reliance on extensive parameterization.

In this paper, to support the urgent need for more advanced GCMs, the research

status of GCM was reviewed and some of the most representative GCMs will be examined for a better reference for the application at different scenario. The review is organized into several key sections, beginning with the fundamental principles of GCMs, followed by a detailed evaluation of classical and modern methods, including Lydersen, Joback, CG94, and Sharma models. The review also introduces applications in pure components, mixtures, and process design, before concluding with recent advancements and future research directions.

2. Fundamentals of group contribution methods

2.1. Principles and approaches

The fundamental principle of a GCM is that the physical properties of a molecule are determined by its molecular structure and the interactions and chemical bonds between its atoms. In an ideal scenario, if a complete understanding of the interactions acting on each atom is known, all the physical and chemical properties of a molecule can be accurately determined. However, approximations are necessary because of the complexity of the intramolecular interactions and intermolecular interactions. Intramolecular interactions, such as steric hindrance, resonance, and ring strain, play a key role in determining properties related to internal energy and flexibility of the molecular structure like T_c, S^0, while intermolecular interactions significantly affect properties like T_b and enthalpy of vaporization, where the strength of interactions between molecules determine phase behavior.

Since the energy of atoms is mainly affected over very short distances, it is generally assumed that an atom is primarily influenced by its immediate neighboring atoms. This implies that if an atom has the same local environment in different molecules, it should behave similarly, regardless of the overall molecular structure. For instance, a carbon atom bonded to three hydrogen atoms and one nitrogen atom exhibit the same contribution to the properties of a molecule whether it is part of N-methylpyrrolidine or trimethylamine. This approximation forms the basis of GCMs. It is worth mentioning that GCMs provide an empirical approximation to the full quantum mechanical (QM) solution by capturing average structural contributions rather than explicitly solving the Schrödinger equation for each molecule. Therefore, GCMs are typically more effective for estimating intramolecular properties like ΔG_f and ΔH_f, but may be less accurate for properties that strongly depend on intermolecular interactions, such as T_b and T_m [14].

In GCMs, specific atomic arrangements, namely "groups" that can consist of one or more central carbon atoms, are assigned numerical values representing their contributions to a given molecular property. By adding up the contributions of all functional groups within a molecule, the total sum correlates with or directly determines the overall physical properties of the molecule. This approach simplifies the property prediction and is widely used in computational chemistry and molecular design [15]. To determine the numerical values of group contributions, linear or non-linear regression analysis is commonly applied in the development of GCM. A set of experimental property data for known molecules is used to fit a linear (or non-linear) model, where the contribution of each functional group is treated as a regression coefficient. The general form of the equation used is:

$$P = \sum_i G_i N_i + \epsilon \tag{1}$$

where P is the target physical property or a function of the target property, G_i represents the contribution of each functional group, N_i is the number of times that group appears in the molecule, and ϵ is a fitting parameter that also determines P. By applying statistical methods, such as least-squares regression, the optimal values of G_i are obtained. One of the main difficulties in this approach is that an extremely large number of groups would be required to describe all possible combinations of atoms and nearest neighbors. Even if these groups could be identified, the lack of sufficient experimental data would prevent statistically significant estimates of their contribution values. To address this, further approximations are made. One key assumption is that the contributions of many groups are not highly dependent on the exact identity of all nearest neighbors as organic compounds often include different types of atoms [15]. For instance, in the case of a carbon atom bonded to three hydrogen atoms and one nitrogen atom, its contribution to molecular properties could be approximated as a carbon bonded to three hydrogen atoms and any non-hydrogen atom, rather than specifically a nitrogen atom. This simplification reduces the number and complexity of group definitions and increases the statistical reliability of contribution values [16].

2.2. First order group contributions

In GCMs, first order groups represent basic functional groups in a molecule that contribute to the estimation of physical properties. These groups typically consist of single atoms or small clusters of atoms that define the primary structure of a molecule, without considering more complex interactions. First order groups are directly assigned contribution values, which are summed up to estimate pure component properties like T_b, T_c, and $H^0 - H^0(0\,\mathrm{K})$. These groups are independent of the molecular environment and contribute the same value to the properties of a molecule , regardless of the rest of the molecular structure [17]. Examples of first order groups can be a methyl group ($-CH_3$), a methylene group ($-CH_2-$), a hydroxyl group ($-OH$) and a carboxyl group ($-COOH$). Each of these groups has a predefined numerical value G_i (see Eq. 1) that contributes to the calculation of various thermodynamic and physical properties. The simplicity of first order groups lies in the use of pre-determined values, eliminating the need for complex calculations. Property estimation is achieved by directly summing the contributions of all first order groups, without relying on additional QM calculations or extensive experimental data [18]. While first order groups are highly useful for simple molecules, the limitation in capturing molecular interactions, such as hydrogen bonding, isomers or steric effects, results in the failure of being accurate for polar compounds and complex molecular structures. To improve the accuracy and applicability of GCM, higher order group corrections that capture more complex interactions are often necessary.

2.3. Second order group and higher order contributions

The limitations of early GCMs arise because first order groups treat each functional group without accounting for the influence of neighboring atoms or groups, which

can lead to a decreased predictive accuracy in systems where intergroup interactions, steric hindrance, conjugation, or electronic delocalization significantly affect thermophysical properties. A simple example is the comparison between 2-methylhexane and 2-ethylpentane with the same molecular formula (C_6H_{14}). Although these two compounds have different structures, first order GCMs do not adequately distinguish because only the number of first order groups, such as CH_3 or CH_2, is considered and the configuration of connectivity of these groups is ignored. Second order groups solve this issue by considering not just the group itself, but also the nature of its neighboring atoms, which allows the model to better account for branching and other structural effects, thereby improving the prediction of isomer-specific properties.

Second order group contributions are introduced to address these limitations and extend the first order approximation by capturing the interactions between adjacent first order groups. Importantly, second order groups are not merely refinements and are based on underlying molecular theories, which consider how the local chemical environment modifies group behavior. For example, in the group $CH_2(CH_3)(CH)$, the central CH_2 is influenced by both neighboring substituents CH_3 and CH, capturing branching effects more accurately than first order representations. Such groups can be identified using QM calculations [18]. One of the earliest methods to consider corrections beyond simple first order additivity was proposed by Benson [9], which introduced systematic rules for estimating thermodynamic properties, including the heat of formation, using group contributions while explicitly incorporating symmetry and steric effects as corrections to the standard group additivity framework. This can be viewed as an initial conceptual foundation for higher order contributions, especially in the context of heat of formation, by recognizing the importance of molecular topology and electronic structure in modifying group behavior.

Multiple definitions of second order groups of hydrocarbons exist in the literature, depending on the structural phenomena being captured—such as steric bulk (e.g., the tert-butyl group $–C(CH_3)_3$), symmetry, or resonance [19]. These groups are sometimes parameterized through ab initio calculations or detailed structural enumeration, offering corrections to first order GCMs and helping reduce systematic bias [20]. In addition, statistical methods like sensitivity analysis can be used to evaluate the impact of each second order group on the overall prediction performance by analyzing how small changes in the presence or frequency of specific groups affect the error of the model [21]. In this way, which groups contribute most significantly to predictive accuracy can be identified.

Higher order group based on second order groups contributions go even further by considering long range intramolecular interactions, molecular topology, three-body effects, and even connectivity indices. These are particularly valuable for modeling properties in large, branched hydrocarbons or macromolecules, where simple additive models break down [22]. Higher order terms based on molecular theories help capture the non-linearity and cooperative effects that first and second order GCMs may miss, making them useful in fields such as drug design, polymer science, and advanced material development [23]. Nevertheless, it is worth noting that property estimation methods beyond second order groups (third order, fourth order, etc.) introduce a large number of model parameters which can satisfactorily regenerate the experimental data used to be introduced, but have questionable extrapolating behaviors. Despite the potential to improve accuracy, second and higher order GCMs come at the cost of increased complexity and are more difficult to parameterize. As the number of interactions and structural variations grows, so too does the number of re-

quired parameters and the size of the training dataset needed for reliable calibration. To overcome this, recent advances advocate for hybrid models that combine GCM frameworks with data-driven techniques—including machine learning and Artificial Intelligence (AI) [24]. These approaches allow for flexible parameter tuning, automated feature discovery, and scalable correction models that preserve interpretability while improving performance [11, 25]. At the same time, it should be emphasized that high interpolation accuracy does not guarantee reliable extrapolation performance.

3. Applications of group contribution methods for pure component properties of hydrocarbons

The number of possible molecular structures increases exponentially with carbon number (Table 1), making it practically impossible to synthesize and experimentally measure all compounds [25]. For instance, there are over 75,000 isomers of branched alkanes with ten or more carbon atoms. This complexity is particularly relevant in the context of hydroisomerization processes, where accurate thermodynamic property estimation is essential for designing optimal reaction conditions [5]. GCM has developed rapidly over the past half decade and many mature models have emerged as a convenient way to provide estimations of thermodynamics proprieties for pure organic molecules lack of experiment data [7]. More information is provided to three group contribution methods [8, 15, 26] which, based on the literature [27] have been evolutionary in the development of group contributions and have been extensively used. A GCMs developed specifically for alkanes [5] is also examined, which exhibits improved predictive performance for branched isomers and offers valuable insights for advancing the development of structurally sensitive GCM frameworks.

3.1. Lydersen Method

The Lydersen [26] method is one of the earliest GCMs used to predict properties for organic compounds where critical parameters T_c, V_c and P_c were investigated and equations of groups contributions of these properties were given as:

$$V_c\ /[\mathrm{cm}^3/\mathrm{mol}] = 40 + \sum G_i N_i \tag{2}$$

$$P_c\ /[\mathrm{bar}] = \frac{M_w}{(0.34 + \sum G_i N_i)^2} \tag{3}$$

$$T_c\ /[\mathrm{K}] = \frac{T_b}{0.567 + \sum G_i N_i - (\sum G_i N_i)^2} \tag{4}$$

where G_i is the numberical values of group contributions (see Eq. 1) which are shown in Table 2 N_i is the occurrence of the groups and M_w is the molar mass of molecule in [g/mol]. It is worthy mentioning that equation for T_c was gained based on Guldberg Rule [28] that T_b expressed in Kelvin is approximately two-thirds of T_c. The Lydersen Method is simple and widely used for estimating critical properties when experimental data is unavailable. Comparisons with experimental data show that the calculated values deviated by less than 1% for T_c and 3.8% for P_c [15]. However, the method lacks high accuracy when it came to highly branched or heteroatomic compounds where molecular interactions and steric effects shall be considered [15]. One

major disadvantage of this method is that the experimental value of T_b is required to estimate T_c, which may not always be available. Nevertheless, Lydersen method has been considered as the prototype for and ancestor of many new models like Joback [15], Constantinou and Gani [8] and others [29, 30].

3.2. Joback method

The Lydersen method is refined in the Joback Method [15] to improve accuracy and applicability through the incorporation of additional molecular features that had previously not been considered. While the Lydersen method was built on first order groups that assumed the group contribution of each functional as independent contribution to the whole molecule, Joback introduced topological corrections for cyclic systems, conjugated bonds, and hydrogen bonding effects, which is a significant improvement over the original model of Lydersen where simple additivity was assumed without considering more detailed molecule features. In addition to improving the estimation for the critical parameters, Joback method expands to 12 thermodynamics properties, namely T_c, P_c, V_c, T_b, T_m, ideal gas heat capacity C_p, ΔH_f, ΔG_f, heat of vaporization (ΔH_vap), entropy of vaporization (ΔS_vap), heat of fusion (ΔH_fus) and some transport properties like the liquid dynamic viscosity (η_L). Joback investigated a broader and more detailed set of first order groups that covered more heteroatoms, such as nitrogen and halogens, and more functionalized carbon groups, like aromatic carbon. Groups specific to cyclic structures, conjugated systems, and double/triple bonds were also considered. The Joback method provides special attention to polar groups like $-$OH, $-\mathrm{NH_2}$ and $-$COOH, which strongly influence intermolecular interactions like hydrogen bonding. These groups have more complex contributions to properties affected by intermolecular forces, particularly to η_L, heat of vaporization ΔH_vap and T_b.

Although the Joback method is still based on first order group contributions, the design of the group values implicitly reflects common neighboring effects without formally including second order groups, based on the assumption that groups are defined narrowly enough that the surrounding environment is partially accounted for and adjusted group definitions to capture the influence of ring strain in small cyclic molecules, resonance effects in conjugated systems such as aromatics, and branching penalties by distinguishing between groups in branched and linear environments. These considerations allow the method to be more structurally aware and sensitive that reflected complex molecular geometries and interactions without explicitly introducing higher order interaction terms called "pseudo-second order" effect within a first order framework, which provided a new path for definition of high-order groups for future models.

Some of the thermodynamics properties equations derived by Joback are shown below:

$$T_m/[\mathrm{K}] = 122.5 + \sum G_i N_i \tag{5}$$

$$T_b/[\mathrm{K}] = 198 + \sum G_i N_i \tag{6}$$

$$T_c/[\mathrm{K}] = T_b \left[0.584 + 0.965 \sum G_i N_i - \left(\sum G_i N_i\right)^2\right]^{-1} \tag{7}$$

Unlike Lydersen method where T_b was used empirical data, the Joback method directly provided an equation for estimating T_b when experimental values are not available. The critical volume and pressure were given:

$$V_c\ /[\mathrm{cm}^3/\mathrm{mol}] = 17.5 + \sum G_i N_i \tag{8}$$

$$P_c\ /[\mathrm{bar}] = \left[0.113 + 0.0032 N_a - \sum G_i N_i\right]^{-2} \tag{9}$$

where N_i is the number of groups of type i in the molecular structure. This equation improved on Lydersen method by incorporating the impact of molecular structure, ensuring that branched and cyclic molecules were handled with better accuracy.

As mentioned before, Joback also investigated some new properties like ΔH_{vap} and ΔH_{fus}. Notably, the method was extended to include certain transport properties, such as liquid viscosity η_L at standard temperature and pressure, broadening the applicability of GCMs beyond just thermodynamic properties:

$$\Delta H_{\mathrm{vap}}/[\mathrm{kJ/mol}] = 15.30 + \sum G_i N_i \tag{10}$$

$$\Delta H_{\mathrm{fus}}/[\mathrm{kJ/mol}] = -0.88 + \sum G_i N_i \tag{11}$$

$$\eta_L/[\mathrm{Pa \cdot s}] = M_w \exp\left[\frac{(\sum \eta_a N_i - 597.82)}{T} + \sum \eta_b N_i - 11.202\right] \tag{12}$$

where M_w is the molecular weight in [g/mol]. The values G_i for Joback method are shown in Table 3. The Joback method extended group contribution methods by enabling the estimation of a broader set of thermodynamic properties. Compared with the Lydersen method, the Joback method introduced improved group definitions and took the structural effects like branching, and ring strain into account to enhance the accuracy for a wider variety of organic compounds. Nevertheless, while the Joback Method improved performance for moderately complex molecules, newer developments in GCMs have revealed its limitations in extrapolating to large and complex molecules, as well as in handling strongly associating groups [15]. These shortcomings are likely due to the relatively small and limited database used by Joback and Reid to derive the group parameters [31]. It also has unrealible extrapolating behaviour for properties such as T_c and P_c [8]. Despite these limitations, Joback Method remains a widely used group contribution method due to its wide applicability and simplicity, and one of the most influential of the original framework of Lydersen.

3.3. Constantinou and Gani method (CG94)

The Constantinou-Gani Method [8] (CG94) was proposed in 1994. Based on the foundational work of earlier models such as Lydersen and Joback methods, which relied on first order group contributions and assumed additive behavior of independent functional groups, the CG94 was developed to overcome the limitations of earlier GCMs. While Lydersen and Joback methods provided decent and reasonable accuracy for small, simple molecules, these methods struggle with molecules containing multiple functional groups, isomeric variations, and extended conjugated systems where localized interactions and structural dependencies could no longer be neglected [30]. Constantinou and Gani introduced a more systematic and theoretically grounded group contribution framework capable of accounting for the effects of molecular topology, neighboring group interactions, and the effect of specific bonding arrangements on thermodynamic properties. The method was designed to estimate a broad range of critical and thermodynamic properties, including T_c, P_c, V_c, ΔH_f and ΔG_f, directly from molecular structure with improved accuracy and wider applicability. By the incorporation of a hierarchical group definition system and the more detailed consideration of molecular complexity, CG94 was seen as an important step forward in the evolution of group contribution methodologies and established a versatile and reliable tool for property prediction for a diverse set of organic compounds to date cited more than 1000 times. The most notable refinement made in CG94 is its two-level group contribution framework. In CG94, first order groups were defined representing basic functional fragments of molecules like those used in Lydersen and Joback methods, which captured the primary contributions of individual atoms and small functional units. Constantinou and Gani introduced second order groups to account for structural dependencies, local interactions, and topological effects based on the principle that molecular properties are not solely dictated by isolated groups but are also significantly influenced by the surrounding molecular environment. These second order groups served as correction factors that modify the contributions of first order groups based on specific arrangements of neighboring groups, the presence of conjugation, and the molecular context in which a functional group exists. The proposed equation for CG94 equals:

$$f(X) = \sum_i N_i C_i + W \sum_j M_j D_j \tag{13}$$

where where $f(X)$ represents the function of estimated value of the target property X, N_i and M_j are the occurrence of first order groups and second order groups, and C_i and D_j represent the group contributions. W is assigned as follow:

$$W = \begin{cases} 1, & \text{both first order and second order groups are used(full model)} \\ 0, & \text{only first order groups are used (basic model for a first order approximation)} \end{cases} \tag{14}$$

The determination of the adjustable parameters and second order groups N_i and M_j is achieved via a two-step and regression procedure designed to ensure accuracy and independence between the two levels of group contributions:

(1) Regression analysis is performed exclusively on the first order groups, with the

factor W set to zero, effectively deactivating the influence of second order groups. This allows for the precise estimation of C_i based solely on the primary functional groups present in the molecule without interference from structural corrections.

(2) The previously determined first order contributions C_i held constant, and the second order approximation is activated and fitted by setting $W = 1$. With the first order contributions fixed, regression is then used to determine the values of the second order group contributions D_j. This sequential approach ensures that the second order groups serve as optimization or corrections to the initial first order approximation, capturing structural and topological effects such as neighboring group interactions and conjugation.

By dividing the regression into two distinct phases, the method maintains the independence of first order contributions while enhancing overall accuracy through the systematic incorporation of second order corrections. The selection of second order groups in the fitting method shall be considered carefully for a better accuracy of the whole model beheavior. As mentioned before, the definition of first order group in CG94 was similar to the previous methods, which involved small functional units such as $-CH_3$ and $-CH_2-$. To be consistent with the group contributions for mixtures, Constantinou and Gani used the UNIFAC [32] first order groups. Rather than representing a unique atom or bond type, second order groups were characterized as a structural relation or interaction between adjacent groups or atoms. These include systems of conjugated double bonds, cyclic structures, and specific branching patterns. The incorporation of second order groups enables the model to detect and correct for structural motifs that influence physical properties beyond the sum of isolated functional groups. One of the most important structural effects accounted for in second order groups definition is conjugation, which refers to the delocalization of π-electrons across alternating single and multiple bonds and then creates a system where electron density is spread over several atoms [33]. This delocalization stabilizes molecules and change the physical and thermodynamic properties [34]. In CG94, Constantinou and Gani incorporated conjugation by specific second order groups that recognize conjugated patterns within the molecular structure based on ABC approach (Atoms, Bonds, and Conjugates) [35], where conjugation was treated as a core structural principle, knowing that many molecules exist as hybrids of multiple conjugate forms. In this way, the classes of conjugate forms having the strongest conjugation effects can be identified by analyzing the contributions of the associated operators. The group identification focuses on the operators which correspond to the important conjugate forms, that is, the operators with significantly higher contribution than others. Using these principles, Constantinou and Gani have defined 30 second order groups. For parameter estimation and model calibration, Constantinou and Gani used an extensive and diverse dataset composed of 370 organic compounds, varying from hydrocarbons, alcohols, ketones, acids, esters, ethers, and amines, as well as more structurally complex compounds such as aromatics, conjugated dienes, and cyclic systems. Eight pure component properties including T_b, T_m and critical parameters are investigated:

$$\exp\left[\frac{T_c/[\mathrm{K}]}{t_{c0}}\right] = \sum_i N_i C_i + W \sum_j M_j D_j \tag{15}$$

$$(P_c/[\mathrm{bar}] - p_{c1})^{-0.5} - p_{c2} = \sum_i N_i C_i + W \sum_j M_j D_j \tag{16}$$

$$V_c/[\mathrm{cm}^3/\mathrm{mol}] - v_{c0} = \sum_i N_i C_i + W \sum_j M_j D_j \tag{17}$$

with selected parameters shown in Table 4 and 5. By incorporating second order corrections to account for structural complexities such as conjugation and branching, the model achieved a high level of consistency for a diverse range of properties. For critical properties, the method produced an average absolute percentage deviation of 0.89% for T_c, 2.89% for P_c and 1.79% for V_c, which shows a better performance than Joback method.

In summary, CG94 represented a significant advancement in the field of GCMs by introducing a systematic and comprehensive framework of second order groups, which in a way addressed the limitations of earlier models. By the integration of both first order and second order groups, this method successfully captured not only the fundamental contributions of individual functional groups but also the critical influence of structural features such as conjugation, branching, and ring strain. A clear set of principles for defining second order groups made sure that the method can accurately represent localized molecular interactions essential for reliable property estimation. With its hierarchical design and consistent group structure applied across multiple thermodynamic and physical properties, CG94 achieved a superior accuracy and capability at that time. Its performance over a broad dataset and its ability to maintain internal consistency across diverse properties established itself as one of the most influential frameworks in GCMs [36]. Despite its significant contributions and improved accuracy over earlier methods, the CG94 model has several major limitations. First, its development was based on the experimental data available at the time, which imposed a constraint on the diversity of compounds used for model calibration [37]. For instance, the experimental data set used for the regression included a limited number of complex aromatic, cyclic, and highly branched molecules, which may be less accurate to such structures, despite its strong extrapolating performance. While the model introduced second order groups to better account for local interactions, its physical foundation is somewhat limited—only the second order terms were derived based on structural theory and the first order contributions remained in the form of a conventional additive group term (as per the group contribution principle) [38]. Due to the sparsity of reliable data, several groups were not assigned any contribution values during regression, leading to gaps in model generality. The method also struggles to fully describe long-range interactions and global molecular effects such as conformational flexibility or electronic delocalization, which are critical for large or highly interactive systems. As a result, CG94 may fail to capture the complexity of modern molecular design challenges without further refinement or integration with hybrid, data-driven approaches especially when distinguishing between branched alkane isomers [5]. Therefore, CG94 requires additional reinforcement with molecular theories to be in position to capture the reliably the estimation of very complex organic structures or isomer behaviours extensively. Nevertheless, CG94 still remained one of the

most reliable and transferable approaches available and has laid the foundation for the development of even more sophisticated GCMs. It introduces the second order approximation and demonstrates how molecular theories can be embedded into group contributions through second order groups without sacrificing the simplicity and applicability of group contributions [39]. It has also been implemented to one of the most extensive list of pure component properties, properties of polymers, temperature dependent properties and properties of mixtures [40, 41]. Finally, it was proved to have robust extrapolating behaviour [42]. Building on the principles of CG94, Marrerro and Gani [18] added a correction to the original model by introducing a third order term. Using an extended database, the updated approach improved applicability in interpolation behaviour. Nevertheless, it may not be able to distinguish among the complex isomer structures, to extensively encounter molecular symmetry and steric effects. Furthermore, the third order term introduced additional complexity and a very big number of adjustable parameters in the regression disproportional to the number of experimental data were used [43]. Finally, Stefanis and Panagiotou [44] successfully adopted the CG94 method to the estimation of the Hansen solubility parameters .

3.4. Sharma Method

Recent GCMs have aimed to systematically include higher order structural effects to improve predictive accuracy particularly for large and highly branched molecules, which is essential to industrial processes such as hydroisomerization, hydrocracking, and fuel production where precise knowledge of properties like ΔG_{f} and ΔH_{f} is essential for predicting reaction equilibria, optimizing catalyst performance, and designing energy-efficient processes [45]. In isomerization applications, where small structural differences critically affect thermodynamic stability and equilibrium, conventional GCMs fall short due to the inability to distinguish long-chain isomers as the reliability of first order group counts ignores the positional and topological context of each group. As a result, molecules with identical group compositions but different connectivity, such as linear and branched isomers, are often treated similarly, leading to a less accuracy. [46]. To address these limitations, Sharma et al. [5] introduced an advanced group contribution framework that fully relies on a comprehensive set of second order groups into a linear regression scheme for the accurate prediction of ΔG_{f} and ΔH_{f} of hydrocarbons. Unlike earlier models that often relied on predefined functional or simple group contributions and limited structural corrections, this method came up with an exhaustive enumeration of only second order groups to account for the detailed local environments within a molecule, which can provide greater structural sensitivity and enhanced predictive performance. The main feature of the Sharma method is its comprehensive and systematic use of second order groups to accurately represent the local structural environment for long-chain and highly branched alkanes. Unlike previous group contribution methods that apply a limited or heuristic selection of second order corrections, this model exhaustively enumerates all the possible atom combinations surrounding a central atom and forms second order groups present within a molecule. It captures the effects of adjacent group interactions, branching patterns, and local connectivity, which are particularly important in iso-alkanes where subtle variations in branching can significantly affect the thermochemical properties. This complete representation of second order structural features allowed the model to distinguish between molecules with identical first order group compositions but different topologies, addressing a major limitation of conventional

GCMs when applied to large isomeric systems. The definition of second order groups were built by identifying all the possible atom combinations surrounding a central atom. These pairs account for direct connections between neighboring carbon atoms with the respective substitution patterns to reflect the branching and connectivity. The procedure required enumerating every bonded pair of carbons, recording the nature of surrounding substituents of each pair to capture the subtle differences between linear, branched, and highly branched structures. Through this exhaustive identification of bonded pairs across the entire molecule, the model generates a complete set of second order descriptors, ensuring that no significant local structural variation is overlooked. This rigorous method allows the model to systematically capture the influence of localized branching, steric effects, and chain topology, all of which are critical for accurately estimating the thermochemical properties of long-chain and highly branched alkanes.

The model was calibrated using a dataset of 970 long-chain alkane isomers. These were selected from iso-alkanes ranging from C_7 to C_{20} to ensure diversity in chain length and branching. This calibration was enabled by the thermodynamic tables compiled by Scott [47]. These tables cover a broad range of linear and branched alkanes based on high-level quantum chemical calculations and provide consistent and reliable thermochemical data. Examples of second order group defined in the Sharma method can be seen in the treatment of local environments surrounding a central CH_2 unit, as shown in Figure 1. In both cases (a) and (b), the central united atom is CH_2, and if only first order group contributions would be used, these two environments would be treated identically, which may lead to inaccurate property predictions because the influence of neighboring groups would be neglected. By defining second order groups that explicitly incorporate the identities of adjacent groups, the model can distinguish between these environments. This differentiation distinguishes the variations in branching and local connectivity, which can significantly affect thermochemical properties. By considering not only the central group but also its immediate neighbors, the second order group contribution framework can improve structural sensitivity and enhance predictive accuracy, particularly for isomeric systems where subtle differences in local structure have pronounced thermodynamic effects. Following this idea, 69 kinds of second order groups are defined and can be found in Table 6.

Another important feature of the Sharma method is its implementation of temperature-dependent group contributions through polynomial regression. The model uses a quadratic polynomial function to each second order group, enabling the accurate prediction of properties for a wide temperature range from 0 to 1000 K based on the fact that the influence of structural features on thermodynamic properties varies with temperature. This approach allowed the model to move beyond the fixed, temperature-independent parameters of traditional GCMs, and provided a more flexible and physically consistent framework that accounts for the thermal behavior of complex hydrocarbons. The resulting model is therefore highly tailored to the specific challenges of long-chain alkanes. The model demonstrated excellent predictive accuracy for key thermochemical properties, achieving chemical accuracy within 1 kcal/mol for properties such as ΔG_f and ΔH_f. Despite these advances, the Sharma method remains limited to alkanes and its applicability to other classes of compounds is uncertain. This is because extending the second order group definitions to molecules containing heteroatoms (such as oxygen, nitrogen, or sulfur) would require an explosion in the number of group types, which significantly increases model complexity and reducing practicality. While the single usage of second order groups effectively

captured local interactions, the method does not explicitly account for long-range intramolecular effects or global conformational changes, which can affect the properties of very large or highly flexible molecules. In summary, Sharma et al. have provided a new thought for GCMs by only defining a set of second order groups. Through its systematic second order group framework, temperature-dependent corrections, and robust regression strategy, the method has successfully presented a significant progress in predicting thermochemical properties for complex, branched hydrocarbons.

3.5. Applications for the Prediction of Thermodynamics Properties of Pure Compounds

Presently, a large number of GCMs have been widely used for prediction and estimations of thermodynamic properties of pure compounds in both academics and industry. One example is the work of Nannoolal et al. [31], who developed a refined group contribution approach for estimating T_c, V_c and P_c. This method was particularly applied to create a comprehensive and consistent database of critical properties for a wide range of organic compounds for process design and simulation in the petrochemical industry where accurate critical property data are required for phase equilibrium modeling, distillation column design, and safety assessments of hydrocarbons. Similarly, Marrero and Gani [18] presented a group contribution framework to estimate multiple thermophysical properties for the application of solvent screening for separation processes, where hundreds of potential solvents must be evaluated for the volatility, thermal stability, and compatibility with target solutes. In such design scenarios, GCMs enable rapid pre-screening of candidates based on predicted properties before experimental validation, which can save significant time and resources. Another example is from Rarey et al. [48], who focused on predicting the thermal conductivity of organic liquids through group contributions. This model has been particularly relevant in thermal management of heat transfer fluids, where the thermal conductivity of potential components was investigated for designing efficient heat exchangers, cooling systems, and other thermal processes in the chemical and energy sectors. More recently, Csemány et al. [12] evaluated various material property estimation methods, including several GCMs, for alkanes with a focus on modeling droplet evaporation processes in combustion applications. The study showed the importance of accurate prediction of critical properties, vapor pressure, and transport properties (such as thermal conductivity and viscosity) where experimental data are scarce, demonstrating the continuing relevance of GCMs in the energy sector. Groniewsky and Hégely [49] proposed an extension of CG94 through an automated group conversion procedure to broaden the applicability of GCMs to predict vapor pressure and improve acentric factor estimation, which is especially significant for applications in process engineering where accurate vapor-liquid equilibrium (VLE) data are essential, but traditional GCM frameworks face limitations due to incomplete group databases. Overall, these methods show a wide range of ways group contribution methods that are used across different industries, including petrochemical process design, separation technologies, thermal systems, and sustainable chemical development. In all these fields, GCMs are valuable because of the quick, reliable, and broadly applicable property predictions, helping researchers and engineers make informed decisions when experimental data are limited or unavailable.

4. Applications to mixture property estimation

In industrial practice, most operations, like distillation, extraction, absorption, and reaction, deal with complex mixtures rather than individual pure components [50]. As a result, understanding how pure component properties combine to define the behavior of mixtures is essential for accurate process simulation and equipment design. In these situations, GCMs are commonly combined with excess Gibbs energy models to predict mixture properties that are interamolecular like activity coefficients and phase equilibria [51]. These models use group contribution concepts to estimate the non-ideal interactions between different functional groups in a mixture [52]. Such predictions are especially important when dealing with multicomponent systems where experimental data are unavailable or incomplete, as is often the case for new solvents, specialty chemicals, or bio-based feedstocks. A typical example of mixture property estimation using GCMs is in solvent recovery and recycling, where knowledge of VLE behavior helps optimize distillation sequences and minimize energy consumption [53]. Another important application is in extractive distillation, where the interaction between solvent and solute molecules governs the separation efficiency [54]. Fuller et al. [55] developed one of the earliest and most widely adopted GCMs for estimating binary gas-phase diffusion coefficients, which introduced the concept of diffusion volumes assigned to molecular fragments and calculated the diffusion coefficient via structural information from each component. The method was calibrated against a large dataset of 340 binary systems and remained favored in engineering applications due to its balance of simplicity, broad applicability, and reasonable accuracy. Similarly, in the design of liquid-liquid extraction processes, GCM-based mixture models can help identify suitable solvent systems by predicting phase separation and component distribution without the need for exhaustive experimental testing [56]. A significant advancement in predicting mixture properties, especially for systems involving complex and associating molecules, has been the integration of GCMs with molecular-based equations of state. One widely recognized framework is the Statistical Associating Fluid Theory (SAFT) [57], which models fluids based on rigorous statistical thermodynamics by considering molecular size, shape, polarity, and hydrogen bonding interactions explicitly [58]. Variants such as PC-SAFT (Perturbed Chain Statistical Associating Fluid Theory) extend the SAFT framework by modeling molecules as chains of spherical segments with explicit association sites, which enables the accurate description of long-chain, associating, and polar compounds [40]. Papaioannou et al. [59] propose a novel Group Contribution PC-SAFT methodology to predict the thermodynamic behavior of complex heteronuclear molecules and multicomponent mixtures where group contributions are defined to fit PC-SAFT parameters systematically, enabling the estimation of key thermodynamic properties such as phase equilibria and densities even in the absence of extensive experimental data. The approach proved particularly effective for mixtures involving polar, associating, and long-chain components, thereby broadening the applicability of predictive molecular thermodynamics to challenging chemical systems. GCMs can also be powerful tools for the estimation of activity coefficients, excess enthalpies, and phase equilibria in binary and multicomponent systems [41]. In mixtures, GCMs like UNIFAC [60] (UNIQUAC [61] Functional-group Activity Coefficients) and its variants (e.g., Effective UNIFAC) can capture both combinatorial effects (due to molecular size and shape) and residual interactions (due to energy differences among functional groups), enabling accurate modeling of vapor–liquid, liquid–liquid, and solid–liquid equilibria [61]. For instance, these methods can be used to predict azeotropic behavior, miscibility gaps, and solubility of solids in solvents—all crucial

in process design. When enhanced with more precise group definitions or hybridized with data-driven models, GCMs offer a versatile platform for handling the increasing complexity of real-world mixture behavior [14]. Isamu Nagata and Jitsuo Koyabu [62] presente a modified GCM—termed the Effective UNIFAC model—designed to improve the estimation of activity coefficients for predicting phase equilibria. Detailed group volume and area parameters, group-interaction matrices, and validation examples such as binary alcohol-hydrocarbon systems and ternary systems like ethanol–water–ethyl acetate are included and predictions for solubility and excess enthalpy are provided. This method is rooted in an extension of the effective UNIQUAC equation and is tailored to handle vapor–liquid, liquid–liquid, and solid–liquid equilibria, especially in complex binary and multicomponent mixtures. The model relies on group-interaction parameters derived from experimental data and achieves improved prediction accuracy, particularly for systems involving alcohols, hydrocarbons, and polar components.

In summary, GCMs have proven to be a tool for the estimation of mixture properties, for accurate predictions of phase equilibria, activity coefficients, and other thermodynamic behaviors in multicomponent systems. By extending the principles of pure component property estimation to mixtures, GCMs are capable for the modeling of complex interactions between different functional groups without requiring extensive experimental data. By the integration of GCMs with established excess Gibbs energy models, such as UNIFAC and its variants, these methods provide reliable and transferable predictions that support process design, optimization, and scale-up across a wide range of chemical industries.

5. Application to process and product design, optimization and reaction kinetics

Besides the estimation of pure component and mixture properties, GCMs can also be used in the broader context of process and product design and optimization [63]. In modern chemical engineering, process development increasingly relies on predictive models that can evaluate not only the feasibility of individual compounds but also the performance of entire process systems. GCMs provide a critical foundation in this regard by enabling the rapid estimation of key thermodynamic and physical properties that feed directly into process simulation, flowsheet design, energy integration, and economic evaluation. By coupling GCMs with process modeling environments, engineers can efficiently explore a wide range of design scenarios, even when working with hypothetical molecules or novel compounds for which no experimental data exist [64]. This capability is particularly important in the early stages of process synthesis, where screening thousands of possible chemical pathways requires fast and reliable property predictions [19]. GCMs are also deeply integrated into Computer-Aided Molecular Design (CAMD) frameworks to support the simultaneous optimization of both molecular structure and process performance [65]. In such workflows, GCMs allow the property estimation of candidate molecules to be directly linked to process objectives such as minimizing energy consumption, maximizing product yield, or reducing environmental impact [66]. Harper et al. [67] developed an extended CAMD methodology by integrating traditional GCMs with molecular modeling tools and including a structured approach that progressively refines candidate molecules across four levels, from group vector assembly to atomic-level structure generation and 3D molecular modeling using Chem3D. Two industrial application examples were presented in Ref. [67]: the first involved the design of alternative solvents

to replace toluene for phenol removal from wastewater, where candidate solvents were generated, screened, and optimized based on multiple property constraints including environmental impact and separation performance. The second example focused on designing solvents for extractive distillation to separate a close-boiling organic acid mixture, demonstrating the ability of the CAMD method to handle complex molecular architectures and phase behavior modeling. This work shows how combining GCMs with detailed molecular modeling can improve property prediction accuracy, especially when traditional GCMs are insufficient, and the potential for CAMD in industrial solvent design, process optimization, and material innovation. A key advantage of using GCMs in process optimization is the ability to support multi-scale decision-making, bridging molecular design, mixture behavior, and full-process performance. For instance, in the development of alternative refrigerants, green solvents, or bio-based feedstocks, group contribution-based models are used not only to predict the properties of new substances but also to evaluate the performance within the overall process, accounting for separation efficiency, energy demands, and lifecycle emissions [68]. One example is the work of Gmehling [69], where GCMs were applied to optimize complex distillation processes, including the design of distillation columns and the selection of suitable solvents for azeotropic separations. These applications enabled efficient process simulation and flowsheet development, particularly when experimental data are limited or when working with multicomponent systems that exhibit strong non idealities. This study also emphasized the potential for further improvement of GCMs to handle increasingly complex mixtures and systems encountered in industrial practice. Traditionally GCMs have been predominantly used for the prediction of thermodynamic and physical properties. However, the application has been successfully extended into the field of chemical kinetics. In this context, GCMs are used to estimate kinetic parameters such as reaction rate constants (k) and activation energies (E_{act}) by decomposing complex molecules into functional groups whose individual contributions can be systematically summed [70]. This approach enables the rapid prediction of reaction behavior for large sets of molecules, significantly aiding in the design and environmental assessment of chemical processes where experimental kinetic data may be sparse or unavailable [15]. Minakata et al. [70] applied a group contribution method to predict aqueous phase hydroxyl radical reaction rate constants for various organic compounds, supporting environmental assessments of pollutant degradation. Saeys et al. [71] propose an "ab initio group contribution method" for estimating activation energies in hydrogen abstraction reactions, which combined quantum mechanical knowledge with group contributions to improve reaction rate predictions. In these studies, GCMs have proven effective in providing kinetic parameters across diverse reaction systems, from biological networks to environmental processes and catalytic mechanisms. This fast, transferable, and systematic way of estimating reaction kinetics based on molecular structure is becoming a handy tool for both fundamental research and industrial process development.

6. Future Directions and Research Opportunities

Recent advances in GCM have significantly improved the accuracy and generalizability of property prediction, yet several challenges and opportunities remain. Second order group contributions appear to be a particularly promising direction, which pro-

vided an effective balance between structural complexity and model accuracy, capturing critical local interactions such as branching, conjugation, and hydrogen bonding [8, 39]. Current studies suggest that while second order groups considerably improve model fidelity, pursuing higher order group contributions offers limited additional benefit and often leads to diminishing returns with increased model complexity [72, 73]. Therefore, the focus remains on refining second order contributions rather than extending to higher orders [48]. A key priority is to reinforce the group contribution framework with molecular-level theories while maintaining the second order approximation. Quantum chemical calculations and statistical mechanics may help ground the definition of second order groups in physically meaningful terms. This integration can support the consistent treatment of effects like molecular symmetry, which is especially important for accurately modeling isomeric species and complex structural motifs [74]. Such refinement would also help overcome one of the major current limitations: the inability of many GCMs to distinguish between structurally similar but functionally different molecules. Despite the rapid adoption of machine learning and AI in this field, human expertise remains essential, particularly in thermodynamics, process design, and model interpretation. Chemical insight is still required to define meaningful molecular fragments, detect thermodynamic inconsistencies, and guide extrapolation in underexplored chemical domains [25]. Future efforts should also focus on strengthening extrapolation capabilities. While GCMs perform well in interpolation, errors and failures occur when extended beyond original training domains. To address this, better utilization of existing databases—for both interpolation and extrapolation—is needed [7]. Group identification tools and robust similarity metrics can play a key role in mapping new molecules to known group environments. Simultaneously, the availability and coverage of property data must be improved, which in turn supports more reliable process simulation and inverse property estimation frameworks. With these strengthened foundations, GCMs will be better positioned to contribute to sustainable process and product development. Reliable property prediction enables more informed decisions in solvent selection, material screening, and lifecycle design, ultimately accelerating innovation in green chemistry and circular manufacturing.

One of challenges in GCMs is the ability to extrapolate to large hydrocarbons reliably beyond the range of data used for parameterization. While GCMs have proven effective for predicting the properties of well characterized molecules within established chemical families, the accuracy decreases when applied to compounds with novel structural motifs, extreme operating conditions, or larger molecular sizes. This limitation arises because traditional GCMs are often calibrated using datasets that primarily cover small to medium-sized molecules, with limited diversity in functional groups and topologies [32]. Future work must focus on the development of robust extrapolation strategies that can extend the applicability of GCMs to broader chemical spaces. One promising direction involves the use of hierarchical models, where first order and second order contributions are complemented by additional correction terms that capture higher order interactions and long range structural effects [75]. Hybrid approaches that integrate group contribution frameworks with physics-based models, such as equations of state or molecular simulations, may provide better extrapolative capabilities by grounding empirical contributions in more fundamental thermodynamic principles [76, 77]. As traditional GCMs largely rely on empirical correlations derived from experimental data, most models often struggle with accuracy when applied to compounds with unusual electronic structures or rare functional groups. Recent research [7, 13] has explored the integration of physical chemistry

principles and quantum mechanical calculations into GCMs. This approach aims to enhance the fundamental understanding of molecular interactions, thus improving both the predictive accuracy and the transferability of GCMs to systems beyond the range of available data. By combining quantum mechanical, such as electron density distributions, orbital interactions, and non-covalent forces, phenomena like hydrogen bonding, resonance stabilization, and polarizability features that are difficult to capture using only first order additive group contributions can be better described. Hybrid models, which combine quantum chemical calculations with group contribution schemes, offer a pathway to generate group parameters grounded in first-principles calculations rather than relying exclusively on experimental fitting. For example, Gani [7] discussed how integrating QM-derived parameters into property estimation models helps to address cases involving highly polar, reactive, or strained molecules that challenge conventional GCMs, the development of semiempirical quantum mechanical methods, as shown by Christensen et al. [13], computationally efficient ways to calculate molecular interactions and thermodynamic properties can be incorporated into GCMs to enhance the accuracy for systems involving non-covalent interactions. In any case, the reinforcement of molecular theories to group contributions should not be done at the expense of simplicity and wide applicability. Therefore, embedding those theories into the definition of first and/or second order groups will maintain the ability of group contributions to support any process and product design framework in a simple and effective manner.

Another limitation of current GCMs is that the reliance on the availability of high quality experimental data for parameter development and validation. Most existing GCMs are built on databases dominated by small, well-characterized organic molecules containing common functional groups [42]. As a result, these models often fail when applied to complex compounds, such as large biomolecules, polymers, ionic liquids, pharmaceuticals, and heavily branched hydrocarbons, for which experimental property data are lacking. Without robust experimental data covering a diverse range of chemical structures, the extrapolation of GCMs to new molecular spaces becomes unreliable, compromising the accuracy and generalizability of property predictions. A key future direction is the systematic expansion of experimental data libraries, focusing specifically on underrepresented compound classes and functional groups. This includes generating accurate thermophysical and thermochemical property data (such as critical properties, phase behavior, heat capacities, viscosities, and enthalpies of formation) for structurally complex molecules. High-throughput experimental techniques, along with advanced calorimetry, spectroscopy, and chromatographic methods, are now making it more feasible to measure such properties across larger chemical spaces [78]. Additionally, collaborative efforts to develop standardized and open-access databases are essential for ensuring broad usability of new data for GCM development [79]. Recent initiatives such as the NIST ThermoData Engine [80] and Dortmund Data Bank [52] have played critical roles in expanding accessible datasets, but further efforts are required to include more heteroatomic species, ionic species, biodegradable compounds, and environmentally relevant pollutants. These enriched datasets will not only strengthen the accuracy of existing GCMs but also enable the creation of next-generation models capable of handling the increasing molecular diversity encountered in areas like biotechnology, green chemistry, and pharmaceutical design [81, 82]. Without sufficient experimental data, even the most sophisticated GCMs are constrained in predictive power. Therefore, continuous expansion and curation of high-quality, diverse experimental data libraries remain essential to the future

advancement of GCMs.

The future of GCMs is closely related to advanced computational techniques like ML, AI and big data analytics. Traditional GCMs rely on linear and additive models, which is effective for many systems yet can fail to capture the non-linear, multi-scale interactions present in complex molecular systems. By combining GCM frameworks with modern computational strategies, it is now possible to overcome these limitations, improve predictive performance, and automate the development of new group parameters [83]. ML algorithms, such as neural networks, decision trees, and support vector machines, can analyze large molecular datasets to identify hidden patterns and complex relationships between structural features and properties. These data-driven models complement GCMs by refining group contributions, correcting systematic errors, and providing corrections where traditional group-additive assumptions break down. One example can be found in the work of Hwang [11] where the Group Contribution Graph Convolution Neural Network (GC-gcn), a hybrid model that combines the conventional GCM framework with graph convolution networks (GCNs) was introduced to estimate pure component thermodynamic properties. Instead of representing molecules with detailed atomic graphs, the GC-gcn model uses functional groups as nodes, drastically reducing the number of adjustable parameters and making it feasible to train accurate models even with limited thermophysical data. This showed an opportunity to combine machine learning and modern computational science with GCMs. Similarly, Mann et al. [84] propose a new perspective to address the limited flexibility in capturing complex molecular phenomena of traditional GCMs by introducing the integration of deep learning models to dynamically define and adjust group contributions, allowing more nuanced structural features to be encoded. In addition, the framework emphasized the use of explainable AI (XAI) to ensure that the resulting models retain physical interpretability, a crucial factor for reliable property prediction. This hybrid approach aims to improve both the predictive accuracy and generalizability of GCMs, especially for novel and diverse chemical spaces.

7. Conclusions

GCMs have been developed as a practical alternative tool to experimental methods that are often costly, time-consuming, or even impossible for complex and novel molecules for the prediction of thermodynamic properties of organic compounds. In this review, we have examined the development and evolution of GCMs, starting from early models like the Lydersen and Joback methods, which focused on first order groups, to more advanced approaches, such as CG94 and the Sharma method, which used second order group corrections to capture intricate molecular interactions, branching effects, and conjugation patterns. These innovations significantly improved the accuracy and applicability of GCMs, particularly for large, highly branched, or isomeric molecules. In Table 7, an overview of widely used and recently developed GCMs is provided as a reference for the application to different properties. Despite these advances in GCMs, challenges still remain. Current GCMs face limitations when extrapolating to new chemical spaces, handling extreme conditions, especially when temperature is lower than 298K, or dealing with complex compounds lacking sufficient experimental data, especially for isomers. Consequently, future research should focus on enhancing extrapolation models, integrating quantum mechanical insights, expanding experimental datasets for underrepresented molecules, and combining advanced computational techniques, including machine learning, to build more robust, transfer-

exhaustive experimentation impractical. GCMs provide a computationally efficient alternative by predicting properties from molecular structure. The accuracy of GCMs is inherently limited by the quality of the experimental or QM data used for parameter fitting. As expectations for predictive models grow, the gap between available experimental accuracy and GCM output must be critically examined. Traditional GCMs often show reduced reliability when extrapolated to complex molecules or multicomponent systems while recent efforts integrated ML with GCM frameworks enables models to learn from data patterns and correct systematic errors [89]. Ultimately, aligning the accuracy of GCM predictions with the fidelity of underlying data is critical for advancing in reliable process modeling and sustainable chemical design. In evaluating future directions for GCM development, it is essential to recognize the inherent trade-off between simplicity and accuracy. Simpler models that rely on fewer parameters and offer greater robustness and easier extrapolation to new chemical spaces albeit sometimes at the expense of fine-grained predictive accuracy. In sharp contrast, more advanced ML-enhanced approaches can achieve higher predictive accuracy but may suffer from reduced extrapolation capabilities and risk overfitting to specific datasets [87]. Therefore, the desired level of accuracy must be carefully considered based on the application context, as different engineering problems demand different degrees of precision in thermophysical property estimation. Balancing model simplicity, computational cost, interpretability, and the required predictive fidelity will remain a central theme in advancing future GCM frameworks. In summary, GCMs continue to evolve as powerful, adaptable tools that bridge the gap between molecular structure and macroscopic properties. With rapid developments, GCMs are expected to play an increasingly important role in advancing green chemistry, process optimization, and the design of next-generation materials and fuels for more efficient, sustainable, and innovative chemical processes.

Table 1. The number of structural isomers for linear and branched alkanes as a function of alkanes. The exponential growth in isomer number with increasing chain length shows the combinatorial complexity of molecular structures [87].

Molecular Formula	Chain Length	Total No. of Structural Isomers
C_3H_8	3	1
C_4H_{10}	4	2
C_5H_{12}	5	3
C_6H_{14}	6	5
C_7H_{16}	7	9
C_8H_{18}	8	18
C_9H_{20}	9	35
$C_{10}H_{22}$	10	75
$C_{11}H_{24}$	11	159
$C_{12}H_{26}$	12	355
$C_{13}H_{28}$	13	802
$C_{14}H_{30}$	14	1858
$C_{15}H_{32}$	15	4347
$C_{16}H_{34}$	16	10359
$C_{17}H_{36}$	17	24894
$C_{18}H_{38}$	18	60523
$C_{19}H_{40}$	19	148284
$C_{20}H_{42}$	20	366319
$C_{21}H_{44}$	21	910726
$C_{22}H_{46}$	22	2278658

Table 2. Group contribution values used in the Lydersen Method for the estimation of T_c, P_c, and V_c as used in Eqs. (2)–(4).

Group	T_c /[K]	P_c /[bar]	V_c /[cm^3/mol]
–CH_3, –CH_2–	0.020	0.227	55.0
>CH	0.012	0.210	51.0
–C<	–	0.210	41.0
=CH_2	0.018	0.198	45.0
=C<, =C=	–	0.198	36.0
–CH_2– (ring)	0.013	0.184	44.5
>CH– (ring)	0.012	0.192	46.0
>C< (ring)	-0.007	0.154	31.0
–OH	0.082	0.060	18.0
–O–	0.021	0.160	20.0
–COOH	0.085	0.400	80.0
–NH_2	0.031	0.095	28.0
–CN	0.060	0.360	80.0
–NO_2	0.055	0.420	78.0
–S–	0.015	0.270	55.0
=S	0.003	0.240	47.0

Table 3.: Joback method group contributions values (G_i) for various thermodynamic properties including T_b, T_m, T_c, P_c, V_c, ΔH_f, ΔG_f,ΔH_{vap}, ΔH_{fus}, and liquid viscosity coefficients (η_a, η_b) (see Eq. 9). Each G_i represents the additive contribution of the corresponding structural group to the estimated property [15].

Group	T_b/[K]	T_m/[K]	T_c/[K]	P_c/[bar]	V_c/[cm^3/mol]	ΔH_f/[kJ/mol]	ΔG_f/[kJ/mol]	ΔH_{vap}/[kJ/mol]	ΔH_{fus}/[kJ/mol]	η_a	η_b
–CH_3	23.58	-5.10	0.0141	-0.0012	65	-76.45	-43.96	2.373	0.908	548.29	-1.719
–CH_2–	22.88	11.27	0.0189	0.0000	56	-20.64	8.42	2.226	2.590	94.16	-0.199
-CH–	21.74	12.64	0.0164	0.0020	41	29.89	58.36	1.691	0.749	-322.15	1.187
-C-	18.25	46.43	0.0067	0.0043	27	82.23	116.02	0.636	-1.460	-573.56	2.307
=CH_2	18.18	-4.32	0.0113	-0.0028	56	-9.63	3.77	1.724	-0.473	495.01	-1.539
=CH–	17.34	13.72	0.0065	-0.0010	41	12.95	34.72	1.000	0.000	n.a.	n.a.
≡CH	9.20	-11.18	0.0027	-0.0008	46	79.30	77.71	1.155	2.322	n.a.	n.a.
=C=	26.15	17.78	0.0026	0.0028	36	142.14	136.70	2.661	4.720	n.a.	n.a.
=C-	24.14	11.14	0.0117	0.0011	38	83.99	92.36	2.138	3.063	n.a.	n.a.
–NH_2	73.23	66.89	0.0243	0.0109	38	-22.02	14.07	n.a.	n.a.	n.a.	n.a.
-NH (non-ring)	50.17	52.66	0.0295	0.0077	35	53.47	89.39	n.a.	n.a.	n.a.	n.a.
–OH	63.56	20.09	0.0031	0.0084	63	-17.33	-22.99	n.a.	n.a.	n.a.	n.a.
–S– (non-ring)	68.78	34.40	0.0119	0.0049	54	41.87	33.12	n.a.	n.a.	n.a.	n.a.
–S– (ring)	52.10	79.93	0.0019	0.0051	38	39.10	27.76	n.a.	n.a.	n.a.	n.a.
–NO_2	152.54	127.24	0.0437	0.0064	91	-66.57	-16.83	n.a.	n.a.	n.a.	n.a.
-F	-0.03	-15.78	0.0111	-0.0057	27	-251.92	-247.19	-0.670	1.398	n.a.	n.a.
-Cl	38.13	13.55	0.0105	-0.0049	58	-71.55	-64.31	4.532	2.512	625.45	-1.814
-Br	66.86	43.43	0.0133	0.0057	71	-29.48	-38.06	6.582	3.603	738.91	-2.038
-I	93.84	41.69	0.0068	-0.0034	97	21.06	5.74	9.52	2.724	809.55	-2.224

Table 4.: First order group contribution values C_i for the estimation of T_c, P_c, and V_c in CG94 [8]. These values represent the contributions (G_i) of each group to the respective thermodynamic property and are used in linear summation models to predict critical properties.

Group	T_c/[K]	P_c/[bar]	V_c/[cm^3/mol]
CH_3	1.6781	0.019904	0.07504
CH_2	3.4920	0.010558	0.05576
CH	4.0330	0.001315	0.03153
C	4.8823	-0.010404	-0.00034
CH_2=CH	5.0146	0.025014	0.11648
CH=CH	7.3691	0.017865	0.09541
CH_2=C	6.5081	0.022319	0.09183
CH=C	8.9582	0.012590	0.07327
C=C	11.3764	0.002044	0.07618
CH_2=C=CH	9.9318	0.031270	0.14831
CHO	10.1986	0.014091	0.08635
CH_3O	6.4737	0.020440	0.08746
ACH	3.7337	0.007542	0.04215
AC	14.6409	0.002136	0.03985
$ACCH_3$	8.2130	0.019360	0.10364
$ACCH_2$	10.3239	0.012200	0.10099
ACCH	10.4664	0.002769	0.07120
OH	9.7292	0.005148	0.03897
ACOH	25.9145	-0.007444	0.03162
CH_3CO	13.2896	0.025073	0.13396
CH_2CO	14.6273	0.017841	0.11195
CHO	10.1986	0.014091	0.08635
CH_3COO	12.5965	0.029020	0.15890
CH_2COO	3.8116	0.021836	0.13649
HCOO	11.6057	0.013797	0.10565
CH_3O	6.4737	0.020440	0.08746
CH_2O	6.0723	0.015135	0.07286
CH–O	5.0663	0.009857	0.05865
FCH_2O	9.5059	0.009011	0.06858
CH_2NH_2	12.1726	0.012558	0.13128
$CHNH_2$	10.2075	0.010694	0.07527
CH_3NH	9.8544	0.012589	0.12152
CH_2NH	10.4677	0.010390	0.09956
CHNH	7.2121	-0.000462	0.09165
CH_3N	7.6924	0.015874	0.12598
CH_2N	5.5172	0.004917	0.06705
$ACNH_2$	28.7570	0.001120	0.06358
C_5H_4N	29.1528	0.029565	0.24831
C_5H_3N	27.9464	0.025653	0.17027
CH_2CN	20.3781	0.036133	0.15831
COOH	23.7593	0.011507	0.10188

Table 5.: Second order group contribution values D_j for the estimation of T_c, P_c, and V_c in CG94 [8]. These corrections are added to account for structural effects such as branching, ring strain, and conjugation that are not captured by first order groups.

Group	T_c/[K]	P_c/[bar]	V_c/[cm^3/mol]
$(CH_3)_2CH$	-0.5334	0.000488	0.00400
$(CH_3)_2C$	-0.5143	0.001410	0.00572
$CH(CH_3)CH(CH_3)$	1.0699	-0.001849	-0.00398
$CH(CH_3)C(CH_3)_2$	1.9886	-0.005198	-0.01081
$C(CH_3)_2C(CH_3)_2$	5.8254	-0.013230	-0.02300
3 membered ring*	-2.3305	0.003714	0.00401
4 membered ring*	-1.2978	0.001171	-0.00851
5 membered ring*	-0.6785	0.000424	-0.00866
6 membered ring*	0.8479	0.002257	0.01636
7 membered ring*	3.6714	-0.009799	-0.02700
$CH_n{=}CH_m\text{-}CH_p{=}CH_k$	0.4402	0.004186	-0.00781
$CH_3\text{-}CH_m{=}CH_n$	0.0167	-0.000183	-0.00098
$CH_2\text{-}CH_m{=}CH_n$	-0.5231	0.003538	0.00281
$CH\text{-}CH_m{=}CH_n$ or $C\text{-}CH_m{=}CH_n$	-0.3850	0.005675	0.00826
Alicyclic side chain $C_{cyclic}C_m$ (m> 1)	2.1160	-0.002546	-0.01755
CH_3CH_3	2.0427	0.005175	0.00227
CHCHO or CCHO	-1.5826	0.003659	-0.00664
CH_3COCH_2	0.2996	0.001474	-0.00510
CH_3COCH or CH_3COC	0.5018	-0.002303	-0.00122
$C_{cyclic}(=O)$	2.9571	0.003818	-0.01966
ACCHO	1.1696	-0.002481	0.00664
CHCOOH or CCOOH	-1.7493	0.004920	0.00559
ACCOOH	6.1279	0.000344	-0.00415
CH_3COOCH or CH_3COOC	-1.3406	0.000659	-0.00293
$COCH_2COO$ or COCHCOO or COCCOO	2.5413	0.001067	-0.00591
CO–O–CO	-2.7617	-0.004877	-0.00144
ACCOO	-3.4235	-0.000541	0.02605
CHOH	-2.8035	-0.004393	-0.00777
COH	-3.5442	0.000178	0.01511

* Stress-strain ring corrections treated as second order terms.

Table 6.: Second order group definitions used in the Sharma method for branched alkanes [5].

Category	**Groups**
No neighboring groups (CH_4)	CH_4
1 neighboring group (CH_3 as the center atom)	$CH_3(C)$, $CH_3(CH)$, $CH_3(CH_2)$, $CH_3(CH_3)$
2 neighboring groups (CH_2 as the center atom)	$CH_2(C)(C)$, $CH_2(C)(CH)$, $CH_2(C)(CH_2)$, $CH_2(C)(CH_3)$, $CH_2(CH)(CH)$, $CH_2(CH)(CH_2)$, $CH_2(CH)(CH_3)$, $CH_2(CH_2)(CH_2)$, $CH_2(CH_2)(CH_3)$, $CH_2(CH_3)(CH_3)$
3 neighboring groups (CH as the center atom)	$CH(C)(C)(C)$, $CH(C)(C)(CH)$, $CH(C)(C)(CH_2)$, $CH(C)(C)(CH_3)$, $CH(C)(CH)(CH)$, $CH(C)(CH)(CH_2)$, $CH(C)(CH)(CH_3)$, $CH(C)(CH_2)(CH_2)$, $CH(C)(CH_2)(CH_3)$, $CH(C)(CH_3)(CH_3)$, $CH(CH)(CH)(CH)$, $CH(CH)(CH)(CH_2)$, $CH(CH)(CH)(CH_3)$, $CH(CH)(CH_2)(CH_2)$, $CH(CH)(CH_2)(CH_3)$, $CH(CH)(CH_3)(CH_3)$, $CH(CH_2)(CH_2)(CH_2)$, $CH(CH_2)(CH_2)(CH_3)$, $CH(CH_2)(CH_3)(CH_3)$, $CH(CH_3)(CH_3)(CH_3)$
4 neighboring groups (C as the center atom)	$C(C)(C)(C)(C)$, $C(C)(C)(C)(CH)$, $C(C)(C)(C)(CH_2)$, $C(C)(C)(C)(CH_3)$, $C(C)(C)(CH)(CH)$, $C(C)(C)(CH)(CH_2)$, $C(C)(C)(CH)(CH_3)$, $C(C)(C)(CH_2)(CH_2)$, $C(C)(C)(CH_2)(CH_3)$, $C(C)(C)(CH_3)(CH_3)$, $C(C)(CH)(CH)(CH)$, $C(C)(CH)(CH)(CH_2)$, $C(C)(CH)(CH)(CH_3)$, $C(C)(CH)(CH_2)(CH_2)$, $C(C)(CH)(CH_2)(CH_3)$, $C(C)(CH)(CH_3)(CH_3)$, $C(C)(CH_2)(CH_2)(CH_2)$, $C(C)(CH_2)(CH_2)(CH_3)$, $C(C)(CH_2)(CH_3)(CH_3)$, $C(C)(CH_3)(CH_3)(CH_3)$, $C(CH)(CH)(CH)(CH)$, $C(CH)(CH)(CH)(CH_2)$, $C(CH)(CH)(CH)(CH_3)$, $C(CH)(CH)(CH_2)(CH_2)$, $C(CH)(CH)(CH_2)(CH_3)$, $C(CH)(CH)(CH_3)(CH_3)$, $C(CH)(CH_2)(CH_2)(CH_2)$, $C(CH)(CH_2)(CH_2)(CH_3)$, $C(CH)(CH_2)(CH_3)(CH_3)$, $C(CH)(CH_3)(CH_3)(CH_3)$, $C(CH_2)(CH_2)(CH_2)(CH_2)$, $C(CH_2)(CH_2)(CH_2)(CH_3)$, $C(CH_2)(CH_2)(CH_3)(CH_3)$, $C(CH_2)(CH_3)(CH_3)(CH_3)$, $C(CH_3)(CH_3)(CH_3)(CH_3)$

Table 7. A summary of GCMs applied to the prediction of thermophysical and thermochemical properties, including rate constants (k) and activation energies (E_{act}) for hydroxyl radical reactions and hydrogen abstraction reactions, along with representative references

Property	**Method/Author**	**Reference**
T_b, T_m	Joback	15
	CG94	8
	Marrero-Gani	18
	Hwang	11
	Nannoolal, et al.	75
	He et al.	88
	Xue et al.	89
	Hou et al.	90
	Simamora et al.	91
P_c, V_c, T_c	Lydersen	26
	Joback	15
	CG94	8
	Marrero-Gani	18
	Wilson-Jasperson	92
	Nannoolal et al.	31
	Lan et al.	93
	He et al.	88
	Mann et al.	2
	Xue et al.	89
ΔH_f, ΔG_f	Benson	9
	CG94	8
	Holderbaum-Gmehling	32
	Mann et al.	84
	Sharma method	5
	Hwang	11
	Liu et al.	94
	Li et al.	95
C_p	Gardas	96
	Pietro et al.	97
	Han et al.	98
	Haghbakhsh et al.	99
	Ahmadi et al.	100
	Albert et al.	101
	CG94	8
	Villazón-León et al.	102
k, E_{act}	Minakata et al.	70
	Saeys et al.	71
	Liu et al.	103

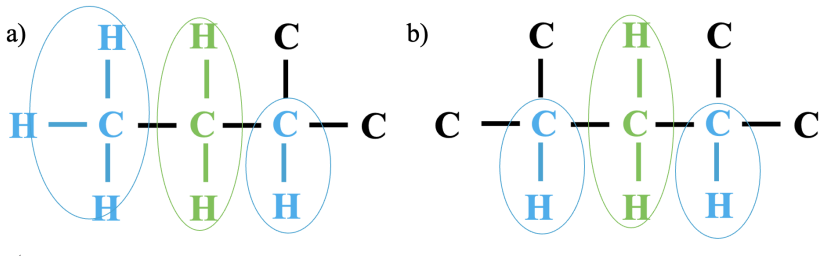


Figure 1. Typical examples of second order groups (a) $CH_2(CH_3)$-(CH) and (b) $CH_2(CH)(CH)$ with CH_2 as the central united atom. Both with CH_2 as the central united atom. Each second order group includes the central atom (green) and its two adjacent bonded fragments (blue) [5].

References

[1] M.S. Rigutto, R. van Veen, and L. Huve, *Zeolites in hydrocarbon processing*, in *Studies in Surface Science and Catalysis*, Vol. 168, Elsevier, 2007, pp. 855–XXVI.

[2] V. Mann, K. Brito, R. Gani, and V. Venkatasubramanian, *Hybrid, interpretable machine learning for thermodynamic property estimation using grammar2vec for molecular representation*, Fluid Phase Equilibria 561 (2022), p. 113531.

[3] Y. Yang, X. He, G. Wu, Y. Yuan, and Z. Liu, *Recent progress in dendrimer-based materials for advanced lithium batteries*, ACS Energy Letters 5 (2020), pp. 2125–2143.

[4] L. Tao, J.J. Jacobson, L. Zhang, M.A. Jackson, D.B. Hodge, C. Kinchin, and M. Wang, *Techno-economic analysis and life-cycle assessment of cellulosic isobutanol and comparison with cellulosic ethanol and n-butanol*, Biofuels, Bioproducts and Biorefining 11 (2017), pp. 965–980.

[5] S. Sharma, J.J. Sleijfer, J. Op de Beek, S. van der Zeeuw, D. Zorzos, S. Lasala, M.S. Rigutto, E. Zuidema, U. Agarwal, R. Baur, S. Calero, D. Dubbeldam, and T.J.H. Vlugt, *Prediction of thermochemical properties of long-chain alkanes using linear regression: Application to hydroisomerization*, Journal of Physical Chemistry B 128 (2024), pp. 9619–9629.

[6] J. Gmehling, *Group contribution methods for the estimation of activity coefficients*, Fluid Phase Equilibria 30 (1986), pp. 119–134.

[7] R. Gani, *Group contribution-based property estimation methods: advances and perspectives*, Current Opinion in Chemical Engineering 23 (2019), pp. 184–196.

[8] L. Constantinou and R. Gani, *New group contribution method for estimating properties of pure compounds*, AIChE Journal 40 (1994), pp. 1697–1707.

[9] S.W. Benson and J.H. Buss, *Additivity rules for the estimation of molecular properties. thermodynamic properties*, The Journal of Chemical Physics 29 (1958), pp. 546–572.

[10] R. Gani, B. Nielsen, and A. Fredenslund, *A group contribution approach to computer-aided molecular design*, AIChE Journal 37 (1991), pp. 1318–1332.

[11] S.Y. Hwang and J.W. Kang, *Group contribution-based graph convolution network: Pure property estimation model*, Springer Nature (2022).

[12] D. Csemány, L. Lisztes, G. Jancsó, and B. Hégely, *Comparative assessment of material property estimation methods for droplet evaporation modeling of n-alkanes, 1-alcohols and methyl esters*, Fluid Phase Equilibria 535 (2021), p. 112998.

[13] A.S. Christensen, T. Kubar, Q. Cui, and M. Elstner, *Semiempirical quantum mechanical methods for noncovalent interactions for chemical and biochemical applications*, Chemical Reviews 116 (2016), pp. 5301–5337.

[14] A. Klamt and F. Eckert, *COSMO-RS: an alternative to simulation for calculating thermodynamic properties of liquid mixtures*, Annual Review of Chemical and Biomolecular Engineering 1 (2002), pp. 101–124.

[15] K.G. Joback and R.C. Reid, *Estimation of pure-component properties from group-contributions*, Chemical Engineering Communications 57 (1987), pp. 233–243.

[16] E.A. Brignole, S. Bottini, and R. Gani, *A strategy for design and selection of solvents for separation processes*, Fluid Phase Equilibria 29 (1986), p. 125.

[17] S. Macchietto, O. Odele, and O. Omatsome, *Design of optimal solvents for liquid-liquid extraction and gas absorption processes*, Trans. IChemE 68, Part A (1990), p. 429.

[18] J. Marrero and R. Gani, *Group-contribution based estimation of pure component properties*, Industrial & Engineering Chemistry Research 40 (2001), pp. 5256–5267.

[19] D. Dalmazzone, A. Salmon, and S. Guella, *A second order group contribution method for the prediction of critical temperatures and enthalpies of vaporization of organic compounds*, Fluid Phase Equilibria 242 (2006), pp. 29–42.

[20] S.I. Sandler and J. Wu, *Use of ab initio quantum mechanics calculations in group contribution methods. 1. theory and the basis for group contributions*, Industrial & Engineering Chemistry Research 41 (2002), pp. 1281–1290.

[21] I.M. Sobol, *Global sensitivity indices for nonlinear mathematical models and their Monte*

Carlo estimates, Mathematics and Computers in Simulation 55 (2001), pp. 271–280.
[22] A. Rahmanian, A. Mohammadi, and D. Richon, *A new group contribution method based on topological indices for estimation of standard enthalpy of formation*, Fluid Phase Equilibria 414 (2016), pp. 80–88.
[23] W. Chanachichalermwong, A. Charoensaeng, and U. Suriyapraphadilok, *Krafft point prediction of anionic surfactants using group contribution method: First-order and higher-order groups*, Journal of Surfactants and Detergents 22 (2019), p. 907–919.
[24] A.R.N. Aouichaoui, F. Fan, S.S. Mansouri, J. Abildskov, and G. Sin, *Combining group-contribution concept and graph neural networks toward interpretable molecular property models*, Journal of Chemical Information and Modeling 63 (2023), pp. 725–744.
[25] Y. Chung, J. Pfaendtner, and E.J. Maginn, *Group contribution and machine learning approaches to predict abraham solute parameters*, ACS Physical Chemistry Au 2 (2022), pp. 570–582.
[26] A.L. Lydersen, *Estimation of critical properties of organic compounds by the method of group contributions*, Engineering Experiment Station Report 3, College of Engineering, University of Wisconsin, Madison, Wisconsin, 1955.
[27] B.E. Poling, J.M. Prausnitz, and J.P. O'Connell, *The Properties of Gases and Liquids*, 5th ed., McGraw-Hill, New York, 2001.
[28] P. Waage and C.M. Guldberg, *Studies concerning affinity*, Forhandlinger: Videnskabs–Selskabet I Christinia (1864), p. 35.
[29] K.M. Klincewicz and R.C. Reid, *Estimation of critical properties with group contribution methods*, AIChE Journal 30 (1984), pp. 137–142.
[30] D. Ambrose, *Correlation and estimation of vapour-liquid critical properties. i. critical temperatures of organic compounds*, NPL Reports Chemistry 92, National Physical Laboratory, 1978.
[31] Y. Nannoolal, J. Rarey, and D. Ramjugernath, *Estimation of pure component properties. part 2. estimation of critical property data by group contribution*, Fluid Phase Equilibria 252 (2007), pp. 1–27.
[32] T. Holderbaum and J. Gmehling, *PSRK: A group contribution equation of state based on unifac*, Fluid Phase Equilibria 70 (1991), pp. 251–265.
[33] A.R. Katritzky, M. Kuanar, S. Slavov, C.D. Hall, M. Karelson, I. Kahn, and D.A. Dobchev, *Quantitative correlation of physical and chemical properties with chemical structure: Utility for prediction*, Chemical Reviews 110 (2010), pp. 5714–5789.
[34] J. Thiele, *Zur kenntnis der ungesättigten verbindungen*, Justus Liebig's Annalen der Chemie 306 (1899), pp. 87–142.
[35] M.L. Mavrovouniotis, *Estimation of properties from conjugate forms of molecular structures: The abc approach*, Industrial & Engineering Chemistry Research 29 (1990), pp. 1943–1953.
[36] E. Stefanis, L. Constantinou, and C. Panayiotou, *Accurate group-contribution method for predicting pure component properties of biochemical and safety interest*, Industrial & Engineering Chemistry Research 43 (2004), pp. 6253–6262.
[37] E. Stefanis, L. Constantinou, and C. Panayiotou, *Estimation of temperature dependent properties through group contributions*, International Journal of Thermophysics 26 (2005), pp. 1370–1388.
[38] E. Stefanis and C. Panayiotou, *Prediction of hansen solubility parameters with a new group-contribution method*, International Journal of Thermophysics 29 (2008), pp. 568–585.
[39] L. Constantinou, R. Gani, and J.P. O'Connell, *Estimation of the acentric factor and liquid molar volume at 298K through a new group contribution method*, Fluid Phase Equilibria 103 (1995), pp. 11–22.
[40] A. Tihic, N. von Solms, M.L. Michelsen, G.M. Kontogeorgis, and L. Constantinou, *Analysis and applications of a group contribution sPC-SAFT equation of state*, Fluid Phase Equilibria 281 (2009), pp. 60–69.
[41] A. Tihic, N. von Solms, M. Michelsen, G. Kontogeorgis, and L. Constantinou, *A pre-*

dictive group-contribution simplified pc-saft equation of state: Application to polymer systems, Industrial & Engineering Chemistry Research 47 (2008), pp. 5092–5101.

[42] J. Abildskov, L. Constantinou, and R. Gani, *Towards the development of a second-order approximation in activity coefficient models based on group contributions*, Fluid Phase Equilibria 118 (1996), pp. 1–12.

[43] Y. Nannoolal, *Development and critical evaluation of group contribution methods for the estimation of critical properties, liquid vapour pressure and liquid viscosity of organic compounds*, Ph.d. thesis, University of KwaZulu-Natal, Durban, South Africa, 2006.

[44] A. Vishnyakov and J.I. Siepmann, *Group contribution method for henry's law constants of polar and nonpolar solutes in polymeric solvents*, International Journal of Thermophysics 30 (2009), pp. 1220–1230.

[45] A. Chakraborty and W.H. Green, *Group contribution and machine learning approaches to predict abraham solute parameters, solvation free energy, and solvation enthalpy* (2023).

[46] H. Calis, W. Lüke, I. Drescher, and A. Schütze, *Synthetic diesel fuels*, in *Energy Sources for Transportation: Handbook of Fuels*, Wiley, 2021, pp. 161–200.

[47] D.W. Scott, *Correlation of the chemical thermodynamic properties of alkane hydrocarbons*, Journal of Chemical Physics 60 (1974), pp. 3144–3165.

[48] Y. Nannoolal, J. Rarey, and D. Ramjugernath, *Krafft point prediction of anionic surfactants using group contributions*, Journal of Chemical & Engineering Data 64 (2019), pp. 4898–4909.

[49] A. Groniewsky and B. Hégely, *Extension of Constantinou and Gani's group contribution method for vapor pressure and acentric factor estimation through automated group conversion*, Fluid Phase Equilibria 577 (2024), p. 113990.

[50] D.W. Green and R.H. Perry, *Perry's Chemical Engineers' Handbook*, 8th ed., McGraw-Hill Education, New York, 2007.

[51] A. Fredenslund, J. Gmehling, and P. Rasmussen, *Vapor-Liquid Equilibria Using UNIFAC: A Group Contribution Method*, Elsevier Scientific Publishing Company, Amsterdam, Oxford, New York, 1977.

[52] J. Gmehling, B. Kolbe, M. Kleiber, and J. Rarey, *Chemical Thermodynamics for Process Simulation*, John Wiley & Sons, 2012 01.

[53] R. Gani and E.A. Brignole, *Molecular design of solvents for liquid extraction based on unifac*, Fluid Phase Equilibria 13 (1983), pp. 331–340.

[54] J.D. Seader and E.J. Henley, *Separation Process Principles*, 2nd ed., John Wiley & Sons, Hoboken, NJ, 2006.

[55] E.N. Fuller, P.D. Schettler, and J.C. Giddings, *A new method for prediction of binary gas-phase diffusion coefficients*, Industrial & Engineering Chemistry 58 (1966), pp. 18–27.

[56] R. Rajesh, R. Morales-Rodríguez, and R. Gani, *Group-contribution methods for the prediction of thermodynamic properties: Recent developments, limitations, and future improvements*, Industrial & Engineering Chemistry Research 61 (2022), pp. 17469–17493.

[57] W. Chapman, K. Gubbins, G. Jackson, and M. Radosz, *SAFT: Equation-of-state solution model for associating fluids*, Fluid Phase Equilibria 52 (1989), pp. 31–38.

[58] J. Gross and G. Sadowski, *Perturbed-chain SAFT: An Equation of State Based on a Perturbation Theory for Chain Molecules*, Industrial & Engineering Chemistry Research 40 (2001), pp. 1244–1260.

[59] V. Papaioannou, T. Lafitte, C. Avendaño, C.S. Adjiman, G. Jackson, E.A. Müller, and A. Galindo, *Group contribution methodology based on the statistical associating fluid theory for heteronuclear molecules formed from mie segments*, Journal of Chemical Physics 140 (2014), p. 054107.

[60] A. Fredenslund, R.L. Jones, and J.M. Prausnitz, *Group-contribution estimation of activity coefficients in nonideal liquid mixtures*, AIChE Journal 21 (1975), p. 1086.

[61] D.S. Abrams and J.M. Prausnitz, *Statistical thermodynamics of liquid mixtures: A new expression for the excess Gibbs energy of partly or completely miscible systems*, AIChE

Journal 21 (1975), pp. 116–128.
[62] I. Nagata and J. Koyabu, *Phase equilibria by effective unifac group-contribution method*, Thermochimica Acta 48 (1981), pp. 187–211.
[63] L. Constantinou and V. Vassiliades, *Computer-aided molecular design using a linear group contribution method for the prediction of pure component properties: Applications to solvent selection*, in *Chemical Product Design: Towards a Perspective through Case Studies*, Elsevier, 2007, p. 34.
[64] A. Dimian, C. Bildea, and A. Kiss, *Integrated Design and Simulation of Chemical Processes*, Elsevier, 2014 09.
[65] R. Gani and E.A. Brignole, *Computer-aided molecular design: Theory and practice*, Computers & Chemical Engineering 26 (2002), pp. 883–888.
[66] L.E.K. Achenie, R. Gani, and V. Venkatasubramanian, *Computer Aided Molecular Design: Theory and Practice*, 1st ed., Elsevier Science B.V., Amsterdam, The Netherlands, 2003.
[67] P.M. Harper, R. Gani, P. Kolar, and T. Ishikawa, *Computer-aided molecular design with combined molecular modeling and group contribution*, Fluid Phase Equilibria 158-160 (1999), pp. 337–347.
[68] F.E. Pereira, E. Keskes, A. Galindo, G. Jackson, and C.S. Adjiman, *Integrated framework for molecular-based process design using a SAFT–VR equation of state*, Computers & Chemical Engineering 35 (2011), pp. 474–491.
[69] J. Gmehling, *Present status and potential of group contribution methods for process development*, Journal of Chemical Thermodynamics 41 (2009), pp. 731–747.
[70] D. Minakata, P. Keli Westerhoff, and J. Crittenden, *Development of a group contribution method to predict aqueous phase hydroxyl radical (HO•) reaction rate constants*, Environmental Science & Technology 43 (2009), pp. 6220–6227.
[71] M. Saeys, M.F. Reyniers, V.V. Speybroeck, M. Waroquier, and G.B. Marin, *Ab initio group contribution method for activation energies of hydrogen abstraction reactions*, ChemPhysChem 7 (2006), pp. 188–199.
[72] R. Gani, *Chemical product design: challenges and opportunities*, Computers & Chemical Engineering 28 (2004), pp. 2441–2457.
[73] J. Gmehling, D. Constantinescu, and B. Schmid, *Group contribution methods for phase equilibrium calculations*, Annual review of chemical and biomolecular engineering 6 (2015), pp. 267–92.
[74] H.S. Wu and S.I. Sandler, *Use of ab initio quantum mechanics calculations in group contribution methods. 1. theory and the basis for group identifications*, Industrial & Engineering Chemistry Research 30 (1991), pp. 881–889.
[75] Y. Nannoolal, J. Rarey, D. Ramjugernath, and W. Cordes, *Estimation of pure component properties. part 1. estimation of the normal boiling point of non-electrolyte organic compounds via group contributions and group interactions*, Fluid Phase Equilibria 226 (2004), pp. 45–63.
[76] A. Fredenslund, J. Gmehling, M.L. Michelsen, P. Rasmussen, and J.M. Prausnitz, *Computerized design of multicomponent distillation columns using the unifac group contribution method for calculation of activity coefficients*, Industrial & Engineering Chemistry Process Design and Development 16 (1977), pp. 450–462.
[77] T. Zhou, R. Gani, and K. Sundmacher, *Hybrid data-driven and mechanistic modeling approaches for multiscale material and process design*, Engineering 7 (2021), pp. 582–604.
[78] Y. Liu, Z. Hu, Z. Suo, L. Hu, L. Feng, G. Xiuqing, Y. Liu, and J. Zhang, *High-throughput experiments facilitate materials innovation: A review*, Science China Technological Sciences 62 (2019).
[79] T. Suzuki, *A nonlinear group contribution method for predicting the free energies of inclusion complexation of organic molecules with α- and β-cyclodextrins*, Journal of Chemical Information and Computer Sciences 41 (2001), pp. 1266–1273.
[80] M. Frenkel, R. Chirico, V. Diky, X. Yan, Q. Dong, and C. Muzny, *Thermodata engine (tde): Software implementation of the dynamic data evaluation concept*, Journal of

Chemical Information and Modeling 45 (2005), pp. 816–38.
[81] M. Frenkel, R.D. Chirico, V. Diky, Q. Dong, S. Frenkel, P.A. Franchois, and D.L. Embry, *ThermoData Engine (TDE): Software Implementation of the Dynamic Data Evaluation Concept. 6. Dynamic Web-Based Data Dissemination through the NIST Web Thermo Tables*, Journal of Chemical Information and Modeling 51 (2011), pp. 1506–1512.
[82] V. Diky, *TRC/NIST: Public data domain*, Presented at the 8th International Research and Practical Conference "Chemical Technology: Science, Econonmy And Production" (2024).
[83] J. Yi, Q. Lei, W.M. Gifford, J. Liu, J. Yan, and B. Zhou, *Fast Unsupervised Location Category Inference from Highly Inaccurate Mobility Data*, in *Proceedings of the 2019 SIAM International Conference on Data Mining (SDM)*. SIAM, 2019, pp. 324–332.
[84] V. Mann, R. Gani, and V. Venkatasubramanian, *Group contribution-based property modeling for chemical product design: A perspective in the AI era*, Fluid Phase Equilibria 568 (2023), p. 113734.
[85] NIST, *Selected values of properties of chemical compounds*, Tech. Rep. 590, National Institute of Standards and Technology, 1984.
[86] Y. Chung, F.H. Vermeire, H. Wu, P.J. Walker, M.H. Abraham, and W.H. Green, *Group contribution and machine learning approaches to predict Abraham solute parameters, solvation free energy, and solvation enthalpy*, Journal of Chemical Information and Modeling 62 (2022), pp. 633–644.
[87] S.R. Rieder, M.P. Oliveira, S. Riniker, and P.H. Hünenberger, *Development of an open-source software for isomer enumeration*, Journal of Cheminformatics 15 (2023), p. 10.
[88] Y. He, Y. Feng, L. Qiu, and D. Tang, *Data-driven approach augmented by attention mechanism in critical and boiling thermophysical properties prediction of fluorine/chlorine-based refrigerants*, Energy 306 (2024), p. 132490.
[89] J. Xue, X. Feng, Q. Jia, Q. Wang, and F. Yan, *A universal spatial group contribution method by 3d-structures for predicting the thermodynamic properties*, AIChE Journal (2025).
[90] X. Hou, L. Yu, C. He, and K. Wu, *Group and group-interaction contribution method for estimating the melting temperatures of deep eutectic solvents*, AIChE Journal 68 (2021).
[91] P. Simamora and S.H. Yalkowsky, *Group contribution methods for predicting the melting points and boiling points of aromatic compounds*, Industrial & Engineering Chemistry Research 33 (1994), pp. 1405–1409.
[92] G.M. Wilson and L.V. Jasperson, *Critical Constants Tc, Pc. Estimation Based on Zero, First, Second-Order Methods*, in *AIChE Annual Meeting*, New Orleans, LA. 1996, p. 21.
[93] T. Lan, Y. Wang, R. Ali, H. Liu, X. Liu, and M. He, *Prediction and measurement of critical properties of gasoline surrogate fuels and biofuels*, Fuel Processing Technology 251 (2024), p. 108918.
[94] Y. Liu, Y. Wang, Z. Wang, X. Wang, and Y. Duan, *Hybrid group contribution and machine learning model for prediction of normal boiling points of organic compounds*, Journal of Chemical Engineering of Chinese Universities 36 (2022), pp. 985–993.
[95] Q. Li, J. Ren, Y. Liu, and Y. Zhou, *Prediction of critical properties and boiling point of fluorine/chlorine-containing refrigerants*, International Journal of Refrigeration 143 (2022), p. 5548.
[96] R. Gardas, *A group contribution method for heat capacity estimation of ionic liquids*, Industrial & Engineering Chemistry Research 47 (2008), pp. 5751–5757.
[97] T.D. Pietro, L. Cesari, and F. Mutelet, *Group contribution models for densities and heat capacities of deep eutectic solvents*, Fluid Phase Equilibria 572 (2023), p. 113854.
[98] J. Han, M. Li, N. Tian, C. Liu, Y. Zhang, Z. Ji, and X. Sun, *Prediction of heat capacity of ionic liquids: A simple group contribution method*, Fluid Phase Equilibria 589 (2023), p. 113729.
[99] R. Haghbakhsh, S. Raeissi, and A. Duarte, *Group contribution and atomic contribution models for the prediction of various physical properties of deep eutectic solvents*, Scientific Reports 11 (2021).

[100] A. Ahmadi, R. Haghbakhsh, S. Raeissi, and V. Hemmati, *A simple group contribution correlation for the prediction of ionic liquid heat capacities at different temperatures*, Fluid Phase Equilibria 401 (2015), pp. 64–74.

[101] J. Albert and K. Müller, *A group contribution method for the thermal properties of ionic liquids*, Journal of the Korean Industrial and Engineering Chemistry 53 (2014), pp. 17522–17526.

[102] V. Villazón-León, A. Bonilla-Petriciolet, J. Tapia, and G. Luna-Barcenas, *Calculation of liquid Cp of pure compounds using an improved thermodynamic model based on group contributions and artificial neural networks*, Fluid Phase Equilibria 576 (2023), p. 113938.

[103] Z. Liu, L. Shang, K. Huang, Z. Yue, A. Han, D. Wang, and H. Zhang, *Combining group contribution method and semisupervised learning to build machine learning models for predicting hydroxyl radical rate constants of water contaminants*, Environmental Science & Technology 59 (2024).

Second-order Group Contribution Method for Thermodynamics Properties of Linear and Branched Alkanes

by

Ziyan Li

Abstract:
Accurate prediction of thermodynamic properties of hydrocarbons is essential for chemical process modeling. Conventional group contribution methods are often used to predict these properties. However, these methods often require extensive parameter sets to handle structural complexities. A refined group contribution method for predicting thermodynamic properties of hydrocarbon isomers with reduced complexity and improved accuracy is presented and discussed. By combining the structural framework of Constantinou and Gani (CG94) with a sensitivity-based selection of second-order groups, a reduced yet highly effective set of twelve second-order groups is identified. This reduced set retains the predictive power comparable to more complex models while significantly reducing the number of parameters. Linear regression is applied to model standard enthalpies and Gibbs free energies of formation for a wide temperature range. To test broader applicability, the model is further extended to properties that require nonlinear regression, including critical temperatures, critical pressures, acentric factors, and liquid densities. For all cases, the proposed model achieves high predictive accuracy, demonstrating its robustness and generalizability. This methodology balances interpretability, efficiency, and performance, making it suitable for both research and industrial thermodynamic modeling.

Instructor: Prof.dr.ir. T.J.H. Vlugt
Dr. P. Dey
Faculty: Faculty of Mechanical Engineering

TU Delft

1. Introduction

The accurate prediction of thermodynamic properties of hydrocarbons is a fundamental requirement for the design, simulation, and optimization of chemical processes, as well as innovative products with improved environmental and safety properties, particularly in the context of the global transition toward sustainable fuels and chemicals [1]. Iso-alkanes with high degrees of branching are preferred constituents in sustainable aviation fuels (SAF), lubricants, and phase change materials due to their desirable thermophysical properties such as high energy density, low freezing point, and cold flow properties [2]. Consequently, catalytic processes such as hydroisomerization, which convert linear alkanes into branched isomers inside shape-selective zeolites, are of growing industrial relevance [3]. Experimental determination of thermodynamic properties like the standard Gibbs free energy ($\Delta G_{\mathrm{f}}^{0}$), standard enthalpy of formation ($\Delta H_{\mathrm{f}}^{0}$) and entropy ($\Delta S^{0}$) for the myriad of possible branched alkanes, particularly those with more than ten carbon atoms, is often infeasible due to the large number of isomers and practical limitations of laboratory measurements [4, 5]. To address this, group contribution methods (GCMs) have emerged as a widely-used and efficient approach to estimate thermodynamic properties based on molecular structure [6]. These methods predict properties by summing contributions from predefined structural fragments, termed "groups," which are generally classified as first-order groups that are basic functional units or higher order groups that capture local structural environments and neighboring atom effects [7].

Classical GCMs such as those of Lydersen [8], and Joback and Reid [9] have provided reasonably accurate predictions for small and moderately branched molecules. The Constantinou and Gani (CG94) method [10] improved many of those deficiencies by introducing a two-level structure: first-order groups capture basic functional fragments, while another set of groups, i.e. second-order groups, account for local structural effects like branching and conjugation. This methodology managed to improve accuracy and applicability of group contributions and partially capture the isomer effect. In this method, through chemical intuition, the typical first-order groups for alkanes are used, and second-order groups are defined by specifying a central atom or group and its first neighboring atoms or groups, thereby encoding the local chemical environment more explicitly. This allows for more accurate differentiation between isomers and improves predictions for molecules with complex or branched structures. Unfortunately, the accuracy of predictions still decreases for highly branched long-chain alkanes [11]. This limitation often arises primarily from the reliance on first-order groups and limited inclusion of second-order corrections [12]. Recent research [13–15] has increasingly focused on refining group definitions, expanding group libraries to integrate more structural effects, and applying new computational advances to improve the prediction of thermodynamic properties of complex isomers.

To overcome the shortcomings of existing GCMs, Sharma et al. [13] proposed a novel linear regression-based second-order group contribution method for alkanes that explicitly captures the interactions between neighboring atoms. By training the model on a dataset of C_1–C_{10} alkane isomers and systematically incorporating all possible second-order groups, an accuracy beyond 1 kcal/mol was achieved in predicting $\Delta H_{\mathrm{f}}^{0}$ and $\Delta G_{\mathrm{f}}^{0}$ for alkanes longer than C_{10}. While highly accurate, the Sharma et al. method has certain limitations that lies in the complexity introduced by the use of 69 distinct second-order groups to represent local atomic environments. Although this comprehensive enumeration improves prediction for long and branched alkanes, it

significantly increases the dimensionality of the model, which can make the regression process more complex, and thus reduce interpretability.

This paper presents a novel idea that combines the basic principles of Constantinou and Gani (CG94) [10] and the Sharma et al. methods [13] to maintain the accuracy while reducing the complexity. By identifying and selecting the most relevant second-order groups defined in the Sharma method, and adopting the second-order approximation strategy of CG94 in a data-driven framework, we aim to balance model accuracy and complexity. The proposed method holds potential to for predicting properties of more structurally complex hydrocarbons that contain additional functional groups beyond those found in alkanes. This paper is organized as follows: first, the theoretical background of linear regression, the CG94 framework, and the Sharma et al. method are presented, followed by details of the methodology for selecting key second-order groups through sensitivity analysis. We analyze the predictive accuracy for both $\Delta H_{\mathrm{f}}^{0}$ and $\Delta G_{\mathrm{f}}^{0}$ for a wide temperature range from 0–1500K, and assess how temperature affects outcomes and model robustness. This study concludes with a summary of key findings and a discussion on the implications for a scalable and interpretable GCM. We specifically focus on $\Delta H_{\mathrm{f}}^{0}$ and $\Delta G_{\mathrm{f}}^{0}$ due to the fundamental role in determining chemical equilibrium and thermodynamic feasibility. Other important properties such as critical constants (T_{c}, P_{c}), molar volume at standard condition (V_{m}), and acentric factor (ω) are also included in this study. In the Supporting Information, we provide detailed list of all training data. In SI1.xlsx, the sheet titled DHf0 and DGf0 include training data of $\Delta H_{\mathrm{f}}^{0}$ and $\Delta G_{\mathrm{f}}^{0}$ from the Scott tables [16]. The critical temperature, critical pressure, and acentric factor sheets present experimental values of T_c, P_c, and ω from Ref.[17] used for training our model. The liquid density (298K) sheet provides density training data at 298K from Ref.[18], and the molar volume (298K) sheet contains the corresponding molar volume values derived from the liquid density data. SI1.xlsx also includes the predictions of these properties using CG94 first-order group contribution method, CG94 second-order groups contributions method, Sharma et al. method and our method. In SI2.xlsx, the first- and second-order group contributions using different methods for these properties will be provided. SI3.py provides the script for capturing the first- and second-order groups in CG94 from SMILE strings and Fig. 1 provides an example for using SI3.py.

2. Theory

2.1. Linear Regression

Linear Regression (LR) is commonly used to predict the thermochemical properties, such as $\Delta H_{\mathrm{f}}^{0}$ and $\Delta G_{\mathrm{f}}^{0}$, of alkanes, using the occurrences of first-order or second-order groups as independent variables

$$y = K + \sum_{i=1} C_i N_i \tag{1}$$

where y is the target property, N_i is the occurrence of a first or a second-order group i in the molecule, and C_i is the group contribution of the group i. K serves as fitting residual. To know which variants, or "groups", are relatively more important, a sensitivity analysis is used [19]. In a LR model of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon, \tag{2}$$

the coefficients β_j indicate the marginal change in the response y per unit change in the predictor x_j keeping all other variables constant. When predictors are measured on different scales or units, as is common in GCMs for thermochemical properties, direct comparison of β_j values can be misleading. To assess sensitivity, all variables are transform into standardized form

$$z_j = \frac{x_j - \bar{x}_j}{s_j}, \quad \text{and} \quad z_y = \frac{y - \bar{y}}{s_y}, \tag{3}$$

where $\bar{x}_j$ and s_j are the mean and standard deviation of predictor x_j, respectively, and similarly for the response y. The standardized regression model becomes

$$z_y = \beta_1^* z_1 + \beta_2^* z_2 + \cdots + \beta_p^* z_p + \varepsilon, \tag{4}$$

where β_j^* is the standardized coefficient of predictor x_j computed from β_j via

$$\beta_j^* = \beta_j \cdot \frac{s_j}{s_y}. \tag{5}$$

The standardized coefficient β_j^* quantifies the number of standard deviations the response will change given a one standard deviation increase in x_j, keeping other variables constant. Therefore, the absolute value $|\beta_j^*|$ gives a direct and interpretable measure of the sensitivity of the output to that predictor [20, 21].

2.2. Constantinou and Gani method (CG94)

The Constantinou and Gani [10] (CG94) method, introduced in 1994, features both first-order groups and second-order groups. first-order groups represent basic functional units like -CH_3 or -CH_2-, while second-order groups serve as correction factors that capture structural dependencies, such as branching, conjugation, and neighboring group interactions [7]. The definition of the second-order groups was based on the conjugation principle as presented in the open literature. When applied to alkanes in a united-atom representation, only four first-order groups (CH_3, CH_2, CH and C) and five second-order groups (shown in Fig. 2 (a)) are considered. An innovative element of the CG94 is its two-step property estimation by using the model below:

$$f(X) = \sum_i N_i C_i + W \sum_j M_j D_j + K \tag{6}$$

where where $f(X)$ represents the function (linear or non-linear) of estimated value of the target property X, N_i and M_j are the occurrence of first-order groups and second-order groups, and C_i and D_j represent the group contributions. Initially, the model fits the contributions of first-order groups by ignoring second-order effects ($W = 0$). Once these base values of C_i and K are established, second-order group contributions are introduced and optimized in a separate regression step (W=1), while keeping C_i and

K constant. This ensures that the second-order effects D_j are treated as corrections to the first order approximation. Note that C_i, D_j and K are temperature-dependent parameters, allowing the model to capture the thermodynamic variability of the target property. This approach maintains the independence of first-order groups and allows second-order groups to capture subtle topological and interaction-based corrections without excessive adjustable parameters [7]. Despite its advancements over the earlier GCMs, CG94 still requires some improvements in specific areas. For example, the conjugation principle in CG94 does not always long-range interactions and overall molecular effects like conformational flexibility or electronic delocalization, which are important for modeling large or highly interactive molecular systems [22]. Therefore, CG94 can be further supported by molecular-level theories in order to improve the accuracy of the estimation of properties of highly complex organic structures and accurately capture isomer-specific behavior [23].The CG94 provided the foundation for several other efforts in group contributions aiming to improve GCMs by refining group definitions, expanding group libraries, and incorporating more structural effects. For example, Marrero and Gani [24] added a third order correction to the Constantinou and Gani second order approximation model. However, this introduces a significant number of additional adjustable parameters and implementation complexity. Similarly, Constantinou et al. [25] and later researchers [15, 26–28] explored approaches that integrate ring corrections, stereochemistry, and group interactions beyond nearest neighbors of pure compounds and mixtures. These developments may be perceived as as an intermediate stage, bridging classical dual-level models with modern machine-learning frameworks. A comprehensive critical review of GCMs can be found in Ref.[7].

2.3. Sharma et al. method

The Sharma et al. method [13] represents a recent advancement in CGMs specifically designed to improve the prediction of thermodynamic properties for long-chain and highly branched alkanes developed considering hydroisomerization as an application. Unlike earlier models that rely primarily on first-order groups, the method uses a comprehensive and systematic enumeration of second-order groups as the sole molecular descriptors. This method exhaustively enumerates all the possible atom combinations surrounding a central atom and forms second-order groups present within a molecule. In this way, 69 second-order groups are defined for branched alkanes. This definition of second-order groups captures the influence of neighboring group interactions, branching patterns, and local connectivity, which are factors especially crucial in iso-alkanes where small differences in branching can lead to significant changes in thermochemical properties [7]. Unlike CG94 where both first and second-order groups are used in a two-step regression, the Sharma et al. method exclusively considers second-order groups in LR using the data set provided by Scott [16]. Each of the 69 defined second-order groups is treated as an independent variable and its contribution is directly estimated through the regression coefficients. The extensive use of all 69 distinct second-order groups introduces a notable level of complexity. While this richness and exhaust improves the predictive accuracy, it also makes the model harder to interpret, more data-intensive, and less generalizable. Although the Sharma et al. method marks a significant leap in structural sensitivity, its high dimensionality raises challenges for practical implementation and may limit scalability. Therefore, a sensitivity analysis is used to determine which groups have more impact on predicting the thermodynamics

properties.

3. Results and discussion

Fig. 3 shows the sensitivity analysis for ΔH_f^0 at 500 K in the Sharma et al. method, where each second-order group is characterized by its $|\beta_j^*|$ and its occurrence for all molecules provided by the Scott tables [16]. A higher value of $|\beta_j|$ indicates a large sensitivity, meaning the corresponding group has a stronger influence on the predicted thermodynamic property. The circles within the blue ellipses show both high sensitivities, i.e. strong influence on predicted enthalpy and high frequency of occurrence, which indicates that these groups are not structurally but statistically significant, making them the most important contributors in the model. In sharp contrast, many of the groups concentrated near the origin have either negligible $|\beta_j^*|$ values, low occurrence, or both. These groups contribute little to the overall variance in ΔH_f^0 and may be considered less relevant in terms of predictive power. It is also worth mentioning that some groups have very low occurrence, which may be attributed to the limitation of the training dataset, which includes only C_1–C_{10} isomers and thus lacks highly branched structures only found in heavier alkanes. The combination of high $|\beta_j^*|$ and high occurrence therefore serves as a useful criterion for identifying the most influential structural motifs in the regression model. This trend was consistently observed for all temperatures from 0K to 1500K, for both ΔH_f^0 and ΔG_f^0, indicating the robustness of group importance for thermal variations.

Based on the observation, the 12 second groups falling in the blue circles, which are characterized by both high sensitivity and high frequency occurrence, are proposed to be elected as a new representative set and are shown in Fig. 2 (b). This subset captures the majority of group features while significantly decreasing model complexity by reducing the number of second-order groups needed to fit. This reduced group set (as denoted by: our new model) is then used to develop a new linear regression model, which is systematically compared to the Sharma et al. method, which includes all 69 second-order groups, and the CG94 method, which incorporates both first and second-order groups. All five second-order groups defined in CG94 (Fig. 2 (a)) can be fully represented using combinations of the more detailed second-order groups selected in our new method (Fig. 2(b)). For example, the CG94 group corresponding to $CH(CH_3)_2$ can be assembled from two $CH_2(CH_3)$ and one $C(CH_3)$ groups. Similarly, the CG94 group $CH(CH_3)CH(CH_3)$ corresponds to two $CH_2(CH_3)$ units connected via a central carbon. This demonstrates that the CG94 groups are a subset or simplified combinations of the second-order groups selected through our sensitivity-based approach. Therefore, our new set preserves the representational capacity of CG94 while offering a finer structural resolution.

While our sensitivity analysis is conducted specifically for ΔG_f^0 and ΔH_f^0, this focus is rooted in the original design of the Sharma et al. method, which was developed and calibrated mainly for these two thermodynamic properties. Since the 69 second-order groups in the Sharma et al. method were trained and validated using ΔG_f^0 and ΔH_f^0 data, the selection of a reduced group set should start from the same context. Interestingly, the selected subset of second-order groups emerging from our analysis shows a high degree of chemical intuitiveness. Many of these groups represent prototypical local environments that reflect key branching and substitution patterns, such as $CH_2(C)(CH_3)$ or $CH_2(CH)(CH)$, which are expected to influence a

wide range of thermodynamic and physical properties. This structural logic suggests that the most influential groups for ΔG_f^0 and ΔH_f^0 may also play important roles in other properties like critical parameters and the acentric factor (ω). Therefore, while our method is derived from sensitivity analysis on a limited property domain, its generalizability is empirically plausible and chemically justifiable. In the later sections of this work, we test whether this same set retains strong predictive performance for multiple temperature-dependent properties, providing a first assessment of its broader applicability. The fist- and second-order group contributions C_i adn D_i of these 12 groups used in our method for ΔH_f^0 and ΔG_f^0 can be found in Tables 1, 2, 3 and 4.

The predicted ΔH_f^0 at 500 K for various C_6–C_7 iso-alkanes using first-order group contributions only and CG94 with the reference values from the Scott tables are shown in Fig. 4 and compared. For molecules with simple or slightly branched structures, such as C_6, 2-m-C_5, 3-m-C_5, the first-order model performs reasonably well, showing small deviations from the reference values. As branching increases, the accuracy of the first-order model reduces significantly, while CG94, by incorporating second-order structural correction, improves predictions for some isomers. CG94 still fails to fully capture fine-grained structural effects. In particular, when second-order groups are sparse or structurally ambiguous, the contribution may be undervalued. This still shows the importance of using second-order groups as a correction. Capturing the local structural environment and surrounding atom effects is essential for accurately predicting thermochemical properties of complex branched isomers [29].

For a complete comparison, Figs. 5 and 6, show the predicted ΔG_f^0 at 500K and 800K for six branched isomers. At 500K, all three models show reasonable results to the training data, although clear distinctions begin to emerge. The Sharma et al. method consistently aligns closely with the Scott tables, indicating its robustness in capturing subtle structural effects using more fitted parameters. Our new method, while slightly more variable, maintains comparable accuracy for most compounds. For instance, prediction for 3,4,4-m-C_7 is nearly identical to that from the Sharma et al. method. One can not observe a huge distinction or advantage over CG94 as both models use the same two-level LR procedure, and the main difference lies in the new method incorporating a few additional fitted parameters, slightly enhancing the accuracy. However, at 800K, the differences between models become much more evident. The Sharma et al. method continues to yield excellent agreement with Scott tables. Notably, our new method yields predictions that closely match those of the Sharma et al. method and remain in good agreement with the Scott tables for nearly all species, including highly branched structures. This suggests that at elevated temperatures, where entropic contributions become more dominant and more uniformly distributed, the reduced set of 12 high-sensitivity second-order groups remains sufficient to model structural effects effectively. In sharp contrast, CG94 deviates become more pronounced at 800K, with overestimations reaching up to several kJ/mol for some compounds. This further supports the idea that its second-order correction are needed to accurately model the entropy-sensitive behavior of branched isomers, particularly when temperature amplifies structural contributions. While differences between models are modest at 500K, our method shows predictive performance close to that of the Sharma et al. method at 800K, despite using significantly fewer parameters. This finding demonstrates that a carefully selected subset of second-order groups can retain both sensitivity and accuracy for a broad temperature range.

To further assess the temperature dependence of model performance, Fig. 7 presents the R^2 values for ΔG_f^0 and ΔH_f^0 predictions, respectively, for a range of

temperatures from 400K to 1500K. For both $\Delta G_{\mathrm{f}}^{0}$ and $\Delta H_{\mathrm{f}}^{0}$, the Sharma et al. method consistently maintains the highest R^2 values, exceeding 0.99 at nearly all temperatures, which re-affirms its robustness and accuracy. Our new method exhibits a performance curve that closely follows that of Sharma et al. method, achieving R^2 values above 0.995 over a wide temperature range. Interestingly, its accuracy improves steadily with temperature up to around 1000K, before a slight decline appears. In contrast, CG94 starts with significantly lower R^2 values, below 0.96 at 400K and then shows a gradual rise, reaching a plateau around 0.996 at mid-range temperatures, before dropping sharply at 1500K. For $\Delta H_{\mathrm{f}}^{0}$, a similar trend is observed. The Sharma et al. method again delivers near-perfect R^2, while the our method remains stable around 0.990 with minor fluctuations. CG94 shows improved accuracy with increasing temperature but consistently underperforms relative to the other models. Notably, the gap between our method and Sharma is slightly more pronounced for $\Delta H_{\mathrm{f}}^{0}$ than for $\Delta G_{\mathrm{f}}^{0}$, possibly reflecting that enthalpy is more sensitive to specific group contributions. It is also worth noting that all three models exhibit a decline in R^2 values for $\Delta G_{\mathrm{f}}^{0}$ at 1500K. While still maintaining relatively high accuracy, this simultaneous drop for all models suggests that prediction becomes inherently more challenging at extreme temperatures. One possible explanation may be the increasing dominance of entropic contributions at high temperatures [30]. R^2 values for all temperatures can be found in Tables 5 and 6.

The R^2 analysis for a wide temperature range confirms that Sharma et al. method shows the most accurate and consistent performance, which is expected as it used all 69 type of second-order group as fitted parameters. In sharp contrast, our new method, despite using only four first-order and twelve "important" second-order groups, manages to achieve R^2 values above 0.995 at most temperatures examined. This indicates that the reduced model is not only significantly simpler but also highly efficient in capturing the essential structural effects. Compared to CG94, which shows notably lower R^2 values, particularly at lower and higher temperatures, our new method shows a clear advantage in balancing model complexity with predictive reliability.

To better compare the model performance in practical applications, Fig. 8 presents $\Delta H_{\mathrm{f}}^{0}$ at 298 K for linear alkanes of C_4 to C_{20} using these three different GCMs. The predictions for C_4 to C_{10} represent the fitted results, while the experimental data for C_{11} to C_{20} are extrapolated values, intended to evaluate generalizability of each model beyond the training range. When extrapolated to longer alkanes, significant differences in performance emerge. The Sharma et al. method exhibits the best consistency with the experimental data, followed by our method, while the CG94 shows the largest deviations, particularly for higher carbon numbers. This comparison shows the improved extrapolation capability of our method over CG94.

Having demonstrated the strong performance of the reduced second-order group set in predicting $\Delta G_{\mathrm{f}}^{0}$ and $\Delta H_{\mathrm{f}}^{0}$ for a wide temperature range, we next explored whether this our method also retains predictive power for other key thermodynamic properties. Specifically, we applied the same group framework to estimate critical temperatures (T_{c}), critical pressures (P_{c}), acentric factors (ω), and liquid densities at standard conditions (ρ_{l}), to assess the broader applicability and structural relevance of these selected groups. Unlike before, these properties require functional forms that can accommodate diminishing returns or saturation as molecular size increases, which shows non-linear beheviors [31]. This justifies fitting curves such as power-laws or logarithmic relation rather than relying on a simple additive linear model. Such an approach aligns with prior works [23, 32–34] in the field, where GCMs using nonlinear

regression have successfully improved accuracy for critical properties of hydrocarbons. The fitting equations for T_c, P_c and ω are as follows:

$$e^{\frac{T_c}{T_0}} = \sum_i N_i C_i + W \sum_j M_j D_j + K \tag{7}$$

$$P_c = P_0 + \left(\sum_i N_i C_i + W \sum_j M_j D_j + K \right)^a \tag{8}$$

$$\omega = \alpha \left[\ln \left(\sum_i N_i C_i + W \sum_j M_j D_j + K \right) \right]^\beta \tag{9}$$

These equations are fitted through non-linear regression using the `curve_fit` function from the `scipy.optimize` module in Python. All training data used in this non-linear regression were obtained from Yaws' Handbook [17]. Similarly, the regression was conducted in a two-step procedure consistent with the philosophy illustrated in Eq. 6. First, only first-order group parameters C_i were fitted with $W = 0$. Once the contribution values for first-order groups C_i were established, second-order group effects D_i were introduced and optimized in a separate regression step by setting $W = 1$. It is important mentioning that all the parameters, including T_0, P_0, a, α, β and K, in Eqs. 7, 8 and 9 were fitted together with C_i (when $K = 0$) and these fitted parameters can be found in Table 9. The group counts N_i and M_j are determined from SMILES string using Python. The first- and second-order groups contributions C_i and D_j for T_c, P_c and ω can be found in Tables 7 and 8. All the fitted parameters can be found in the file SI2.xlsx in the Supporting Information.

Fig. 9 shows the parity plots of the predictive performance of our proposed model for T_c, P_c and ω. For all three properties, the predicted values exhibit a strong linear correlation with experimental data, as evidenced by the close alignment of the data points along the ideal $y = x$ reference line. The distribution is particularly concentrated for ω, with minimal dispersion and virtually no systematic deviation. The prediction of T_c and P_c also demonstrates excellent accuracy, although a few deviations appear in more complex or highly branched compounds. This high degree of agreement reflects the ability and robustness of our model to incorporate non-linear structural effects through tailored functional forms. Quantitative performance metrics including R^2, the mean absolute error (MAE), and the average relative deviation (ARD) for each property are summarized in Table 10, which shows a decent accuracy for these three properties.

To analyze the liquid densties (ρ_l) at 298K and 1 bar pressure of alkanes, we followed an indirect regression procedure. First, we compiled a dataset of experimental ρ_l values from literature [18], which covers of a wide range of linear and branched alkanes. Next, the values of ρ_l were converted into molar volumes (V_m) using the relationship:

$$V_m = \frac{M}{\rho_l} \tag{7}$$

where M is the molar mass. This transformation allowed us to use V_m as it is directly proportional to molecular size and structure, while ρ_l is a derived property influenced

by both molecular structure and intermolecular packing, introducing less predictable variations [35]. We then applied linear regression to model the V_{m} using the two-level regression. Finally, the predicted V_{m} were converted back to ρ_{l} for direct comparison with the original training dataset. The first- and second-order groups contributions C_i and D_j for V_{m} can be found in Tables 7 and 8. Fig. 10 shows parity plots of the predictive performance of our proposed model for V_{m} and ρ_{l}. For V_{m}, the figure shows an almost perfect alignment with the experimental values, closely following the ideal correlation line $y = x$. Since V_{m} correlates directly with molecular size and structure, it is more suitable for additive modeling based on first- and second-order group contributions. Its linear dependence on group counts enables accurate prediction using regression techniques. In contrast, for the predicted ρ_{l}, which were obtained by converting the predicted values of V_{m}, larger deviations were found. This can be partially attributed to the narrower range of ρ_{l} values (0.66–0.78 g/ml), compared to the broader range of V_{m} (125–300 ml/mol). Since ρ_{l} is inversely proportional to V_{m}, even small prediction errors in V_{m} may lead to amplified deviations in ρ_{l}, especially at higher density values. This aspect should be taken into consideration when evaluating the overall model performance. While the overall trend remains strong, the scatter around the ideal line is visibly larger than in the plot for V_{m}. This deviation arises because density is a derived, nonlinear quantity, inversely proportional to volume, and small errors in are amplified during the transformation. Moreover, ρ_{l} can be affected by complex molecular interactions. Together with the quantitative performance shown in Table 10, these results confirm the advantage of modeling V_{m} as the primary regression target. This approach not only yields more accurate and stable predictions, but also better reflects the physical relationship between molecular structural and thermodynamic properties. Together with the high predictive performance for T_{c}, P_{c} and ω shown earlier, these results validate the applicability of our method to both linear and nonlinear thermodynamic properties.

4. Conclusions

In this work, we proposed a simplified complexity CGM that applies the second-order approximation approach of the Constantinou and Gani method [10] with a sensitivity-guided selection of second-order groups inspired by the Sharma et al. method [13]. By identifying twelve most impactful second-order groups based on sensitivity, we were able to develop a new model that strikes a balance between predictive accuracy and model simplicity. Our method shows a strong predictive performance for both ΔH_f^0 and ΔG_f^0 of alkane isomers for a wide temperature range from 0–1500K. Notably, it retains accuracy comparable to the Sharma et al. method, which uses 69 second-order groups as fitting parameters, while using only 16 parameters. This shows that only a reduced subset of second-order groups can be essential for capturing the key structural variations relevant to thermochemical properties. Beyond linear regression of ΔH_f^0 and ΔG_f^0, we tested the broader applicability of this reduced group set by fitting T_{c}, P_{c}, ω, and ρ_{l} at 298K using nonlinear regression. The results showed excellent agreement with experimental data, e.g., $R^2 > 0.996$ for T_{c}, P_{c}, and ω, confirming the effectiveness of our approach in modeling thermodynamic properties that require nonlinear fitting procedures. These results collectively demonstrate that our methodology is not only efficient but also broadly applicable to both linear and nonlinear regression tasks in group contribution modeling. The high accuracy for diverse property types validates the robustness of our approach, and shows its potential use in industrial applications

where interpretability, scalability, and efficiency are essential. This new methodology has been implemented only to alkanes. Encouraged by the excellent results, future work will expand its implementation to a wide range of pure organic compounds and properties (thermodynamic, transport, environmental-related, safety related, etc.) that would allow the availability of a powerful tool for process optimization and design of molecules with properties of environmental importance.

5. Supporting Information

The Supporting Information consists of SI1.xlsx, SI2.xlsx and SI3.py. All training data are listed in SI1.xlsx together with the first- and second-order model predictions. On SI2.xlsx, the contributions, C_i and D_i, of each group for each property and other fitted parameters, including K, T_0, P_0, a, α and β in Eqs. 7, 8 and 9 are listed in the sheets containing 'contributions' in, and the predictions of each property listed are listed in the sheets containing 'predictions', respectively; in SI2.xlsx, the sheets starting with 'CG94' show the group contributions and properties predictions for CG94 [10], the sheets starting with 'Sharma' show the group contributions and properties predictions for the Sharma et al. method [13], and the sheets starting with 'New method' show the group contributions and properties predictions for our method. The code to convert to the SMILES string and the code to count the groups for CG94 are in SI3.py.

6. Acknowledgments

This work is part of the Advanced Research Center for Chemical Building Blocks, ARC-CBBC, which is cofunded and cofinanced by The Netherlands Organization for Scientific Research (NWO) and The Netherlands Ministry of Economic Affairs and Climate Policy. The authors also acknowledge the use of computational resources of the DelftBlue supercomputer, provided by Delft High Performance Computing Center (https://www.tudelft.nl/dhpc) [36]. The authors acknowledges Dr.ir.Jasper van Baten for providing data on thermochemical properties of alkanes computed using SPLIT software by AmsterCHEM (https://www.amsterchem.com/).

Table 1. First-order group contributions C_i of our method for ΔH_f^0 at different temperature.

Temperature/[K]	CH3	CH2	CH	C
0	−4.473	−15.034	−24.509	−35.827
200	0.619	−0.645	−0.222	1.522
273	3.751	6.149	11.070	18.372
298	4.874	8.523	13.999	22.572
300	4.938	8.735	15.371	24.771
400	9.645	18.607	31.583	48.773
500	14.559	28.771	48.167	73.197
600	19.637	39.129	64.933	97.765
700	24.804	49.654	81.936	122.516
800	29.822	60.104	98.906	147.302
900	35.080	70.677	115.828	171.833
1000	40.217	81.277	132.910	196.507
1100	44.877	91.980	150.716	222.368
1200	50.419	102.514	167.050	245.820
1300	55.296	113.150	184.229	270.247
1400	60.422	123.745	201.228	294.585
1500	63.390	134.719	219.318	321.960

Table 2. Second-order group contributions D_j of our method for $\Delta H_{\mathrm{f}}^{0}$ at different temperature.

Temperature/[K]	CH3(C)	CH3(CH)	CH3(CH2)	CH2(CH)(CH)	CH2(CH)(CH2)	CH2(CH2)(CH2)	CH2(C)(CH3)	CH2(C)(CH2)	CH2(CH)(CH3)	CH2(C)(CH)	CH2(C)(C)	CH2(CH2)(CH3)
0	−0.132	0.677	5.086	−8.676	−4.500	−1.400	−1.819	−8.497	−3.206	−6.441	−1.475	−3.060
200	−0.129	0.734	5.392	−2.840	0.000	−1.254	−1.777	−3.582	−7.751	−1.343	−3.380	−1.819
273	−0.174	0.761	5.518	−8.077	−6.155	−1.369	−1.640	−9.019	−3.731	−8.142	−1.270	−3.521
298	−0.217	0.775	5.550	−8.012	−5.214	−1.387	−1.648	−0.929	−3.780	−2.823	−1.392	−3.567
300	−0.183	0.775	5.550	−8.012	−5.214	−1.387	−1.648	−0.929	−3.780	−2.823	−1.392	−3.567
400	−0.183	0.775	5.550	−8.012	−5.214	−1.387	−1.648	−0.929	−3.780	−2.823	−1.392	−3.733
500	−0.086	0.861	5.759	−3.963	−5.230	−1.453	−2.172	−0.993	−4.927	−1.753	−3.992	−3.992
600	−0.336	0.916	5.885	−8.999	−5.550	−1.480	−1.077	−0.981	−4.256	−10.117	−0.413	−3.920
700	−0.390	1.026	6.053	−8.608	−6.000	−1.460	−0.806	−0.916	−4.428	−10.613	−0.283	−3.976
800	−0.430	1.026	6.053	−8.608	−6.000	−1.460	−0.806	−0.916	−4.428	−10.613	−0.283	−3.976
900	−0.451	1.169	6.123	−9.084	−6.098	−1.918	−0.949	−0.978	−4.918	−11.064	−0.364	−4.054
1000	−0.507	1.287	6.251	−10.004	−6.538	−1.992	−1.046	−1.012	−4.538	−11.591	−0.456	−4.139
1100	−0.475	1.369	6.339	−10.240	−6.500	−1.431	0.092	−1.059	−4.535	−12.140	−0.542	−4.221
1200	−0.483	1.169	6.317	−10.240	−6.500	−1.431	0.092	−1.059	−4.535	−12.140	−0.542	−4.221
1300	−0.523	1.222	6.336	−10.641	−6.174	−1.309	0.369	−1.022	−3.807	−0.824	−4.322	−4.322
1400	−0.336	1.112	6.411	−10.806	−6.718	−1.411	0.370	−1.160	−4.066	−15.403	−1.579	−3.505
1500	−2.247	1.690	−0.669	−4.930	−8.023	−3.029	10.245	−3.875	3.416	−4.992	1.580	−2.781

Table 3. First-order group contributions C_i of our method for ΔG_f^0 at different temperature.

Temperature/[K]	CH3	CH2	CH	C
0	−4.473	−15.034	−24.509	−35.827
200	0.619	−0.645	−0.222	1.522
273	3.751	6.149	11.070	18.372
298	4.874	8.523	13.999	22.572
300	4.938	8.735	15.371	24.771
400	9.645	18.607	31.583	48.773
500	14.559	28.771	48.167	73.197
600	19.637	39.129	64.933	97.765
700	24.804	49.654	81.936	122.516
800	29.822	60.104	98.906	147.302
900	35.080	70.677	115.828	171.833
1000	40.217	81.277	132.910	196.507
1100	44.877	91.980	150.716	222.368
1200	50.419	102.514	167.050	245.820
1300	55.296	113.150	184.229	270.247
1400	60.422	123.745	201.228	294.585
1500	63.390	134.719	219.318	321.960

Table 4. Second-order group contributions D_j of our method for ΔG_f^0 at different temperature.

Temperature/[K]	CH3(C)	CH3(CH)	CH3(CH2)	CH2(CH)(CH)	CH2(CH)(CH2)	CH2(CH2)(CH2)	CH2(C)(CH3)	CH2(C)(CH2)	CH2(CH)(CH3)	CH2(C)(CH)	CH2(C)(C)	CH2(CH2)(CH3)
0	−0.132	0.677	5.086	−8.676	−4.500	−1.400	−1.819	−8.497	−3.206	−6.441	−1.475	−3.060
200	−0.129	0.734	5.392	−2.840	0.000	−1.254	−1.777	−3.582	−7.751	−1.343	−3.380	−1.819
273	−0.174	0.761	5.518	−8.077	−6.155	−1.369	−1.640	−9.019	−3.731	−8.142	−1.270	−3.521
298	−0.217	0.775	5.550	−8.012	−5.214	−1.387	−1.648	−0.929	−3.780	−2.823	−1.392	−3.567
300	−0.183	0.775	5.550	−8.012	−5.214	−1.387	−1.648	−0.929	−3.780	−2.823	−1.392	−3.567
400	−0.183	0.775	5.550	−8.012	−5.214	−1.387	−1.648	−0.929	−3.780	−2.823	−1.392	−3.733
500	−0.086	0.861	5.759	−3.963	−5.230	−1.453	−2.172	−0.993	−4.927	−1.753	−3.992	−3.992
600	−0.336	0.916	5.885	−8.999	−5.550	−1.480	−1.077	−0.981	−4.256	−10.117	−0.413	−3.920
700	−0.390	1.026	6.053	−8.608	−6.000	−1.460	−0.806	−0.916	−4.428	−10.613	−0.283	−3.976
800	−0.430	1.026	6.053	−8.608	−6.000	−1.460	−0.806	−0.916	−4.428	−10.613	−0.283	−3.976
900	−0.451	1.169	6.123	−9.084	−6.098	−1.918	−0.949	−0.978	−4.918	−11.064	−0.364	−4.054
1000	−0.507	1.287	6.251	−10.004	−6.538	−1.992	−1.046	−1.012	−4.538	−11.591	−0.456	−4.139
1100	−0.475	1.369	6.339	−10.240	−6.500	−1.431	0.092	−1.059	−4.535	−12.140	−0.542	−4.221
1200	−0.483	1.169	6.317	−10.240	−6.500	−1.431	0.092	−1.059	−4.535	−12.140	−0.542	−4.221
1300	−0.523	1.222	6.336	−10.641	−6.174	−1.309	0.369	−1.022	−3.807	−0.824	−4.322	−4.322
1400	−0.336	1.112	6.411	−10.806	−6.718	−1.411	0.370	−1.160	−4.066	−15.403	−1.579	−3.505
1500	−2.247	1.690	−0.669	−4.930	−8.023	−3.029	10.245	−3.875	3.416	−4.992	1.580	−2.781

Table 5. Comparison of R^2 values for predicted ΔH_f^0 at various temperatures using three different GCMs: CG94 [10], our method, and the Sharma et al. method [13].

Temperature/[K]	R^2 (CG94)	R^2 (our method)	R^2 (Sharma et al. method)
0	0.943 972	0.976 717	0.999 988
200	0.965 040	0.985 654	0.999 987
273	0.969 607	0.987 911	0.999 993
298	0.970 933	0.988 487	0.999 995
300	0.971 017	0.988 528	0.999 995
400	0.975 165	0.990 045	0.999 997
500	0.977 978	0.990 901	0.999 986
600	0.979 905	0.991 340	0.999 956
700	0.981 066	0.991 520	0.999 935
800	0.981 887	0.991 299	0.999 915
900	0.982 585	0.991 432	0.999 896
1000	0.982 759	0.991 231	0.999 859
1100	0.982 753	0.991 009	0.999 842
1200	0.982 715	0.990 652	0.999 816
1300	0.982 599	0.990 334	0.999 805
1400	0.982 220	0.989 895	0.999 590
1500	0.981 967	0.989 441	0.999 742

Table 6. Comparison of R^2 values for predicted ΔG_f^0 at various temperatures using three different GCMs: CG94 [10], our method, and the Sharma et al. method [13].

Temperature/[K]	R^2 (CG94)	R^2 (our method)	R^2 (Sharma et al. method)
0	0.943 976	0.976 705	0.999 880
200	0.244 684	0.671 566	0.995 275
273	0.791 147	0.919 019	0.997 689
298	0.865 462	0.947 700	0.998 270
300	0.869 337	0.949 164	0.998 313
400	0.959 663	0.983 684	0.999 167
500	0.979 919	0.991 594	0.999 422
600	0.987 419	0.994 528	0.999 543
700	0.991 061	0.995 952	0.999 598
800	0.993 089	0.996 761	0.999 641
900	0.994 359	0.997 251	0.999 675
1000	0.995 183	0.997 582	0.999 695
1100	0.995 812	0.997 809	0.999 690
1200	0.996 180	0.997 989	0.999 719
1300	0.996 533	0.998 127	0.999 730
1400	0.996 438	0.997 936	0.999 590
1500	0.969 542	0.970 780	0.975 832

Table 7. First-order group contribution C_i for T_c, P_c, ω and V_{m} using our method.

Group	T_c	P_c	ω	V_{m}
CH_3	1.571	0.456	0.538	19.725
CH_2	1.681	0.405	0.005	15.942
CH	1.676	−0.623	−0.531	10.714
C	1.967	−1.282	−1.069	4.506

Table 8. Second-order group contribution D_j for T_c, P_c, ω and V_{m} using our method.

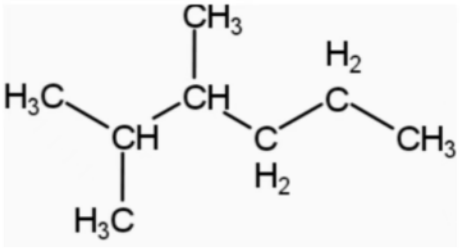
Group	T_c	P_c	ω	V_{m}
CH3(C)	−0.030	0.195	1.124×10^{-4}	0.214
CH3(CH)	0.114	0.187	-1.488×10^{-4}	−0.303
CH3(CH2)	0.223	0.001	-5.382×10^{-4}	−1.320
CH2(CH)(CH)	−0.909	−0.388	1.296×10^{-3}	3.082
CH2(CH)(CH2)	−0.486	−0.390	1.207×10^{-3}	1.976
CH2(CH2)(CH2)	−0.028	−0.391	5.621×10^{-5}	0.554
CH2(C)(CH3)	−0.688	−0.386	1.053×10^{-3}	1.295
CH2(C)(CH2)	0.246	−0.204	-4.237×10^{-4}	−0.426
CH2(CH)(CH3)	−0.096	−0.203	1.036×10^{-4}	0.414
CH2(C)(CH)	−1.109	−0.384	8.907×10^{-4}	3.622
CH2(C)(C)	−1.051	−0.378	2.045×10^{-3}	2.241
CH2(CH2)(CH3)	0.073	−0.210	-2.176×10^{-4}	−1.037

Table 9. Parameters fitted trough non-linear regression using Eqs. 7 8 and 9

T_0/[K]	P_0/[bar]	a	α	β
218.880	7.301	-5.364	2.938	0.575

Table 10. The mean absolute error (MAE), the average relative deviation (ARD) and R^2 for T_c, P_c, and ω using nonlinear regression (our method).

Property	MAE	ARD	R^2
T_c	3.73 K	0.63%	0.9967
P_c	0.17 bar	0.70%	0.9974
ω	0.0091	2.43%	0.9968
V_m	0.95 ml/mol	0.50%	0.9968
ρ_l	0.80 g/ml	0.51%	0.9667



SMILE string:2,3-m-C6

SI3.py

1st order groups	Occurrence (N_i)	2nd order groups	Occurrence (M_j)
CH_3	4	(CH3)2CH	1
CH_2	2	(CH3)3C	0
CH	2	CH(CH3)-CH(CH3)	1
C	0	C(CH3)2-C(CH3)2	0
		CH(CH3)-C(CH3)2	0

Figure 1. An example of using SI3.py to capture the first- and second-order groups for CG94[10]. One can use the SMILE string of a molecule as input to get the number of first- and second-order groups defined in CG94.

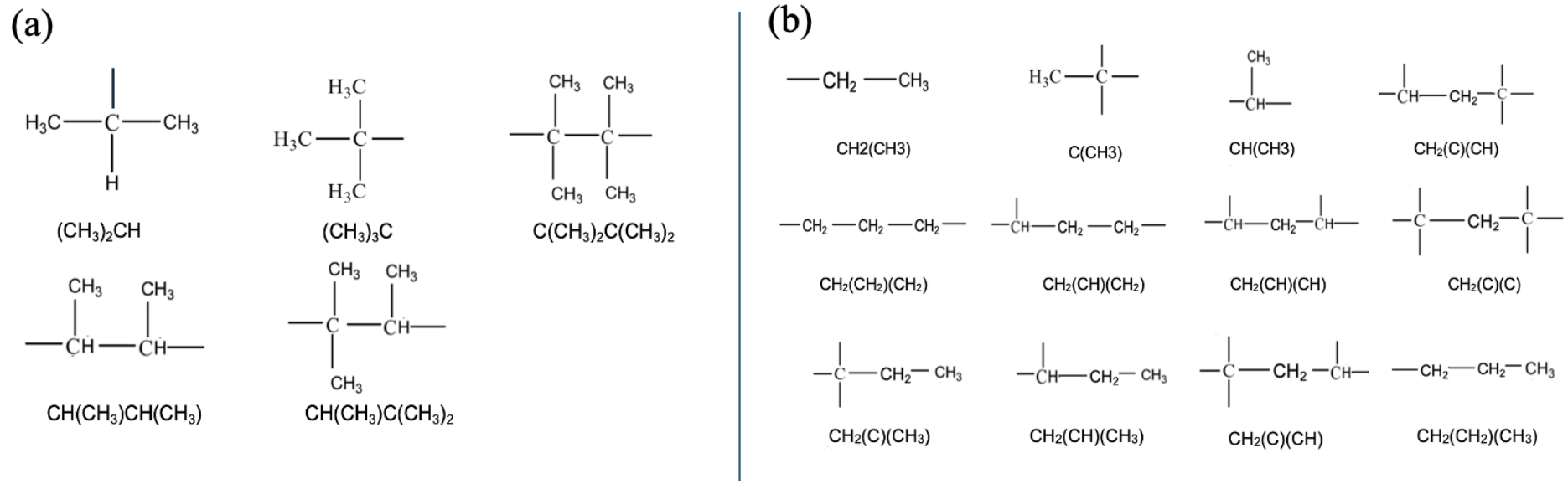


Figure 2. (a) five second-order groups used in CG94 [10] and (b) twelve second-order groups selected through sensitivity analysis. The original CG94 work defined only 5 second-order groups for alkanes, while 12 second-order groups featuring high sensitivity and high occurrence are chosen for our method from the 69 second-order groups defined in the Sharma et al. method [13].

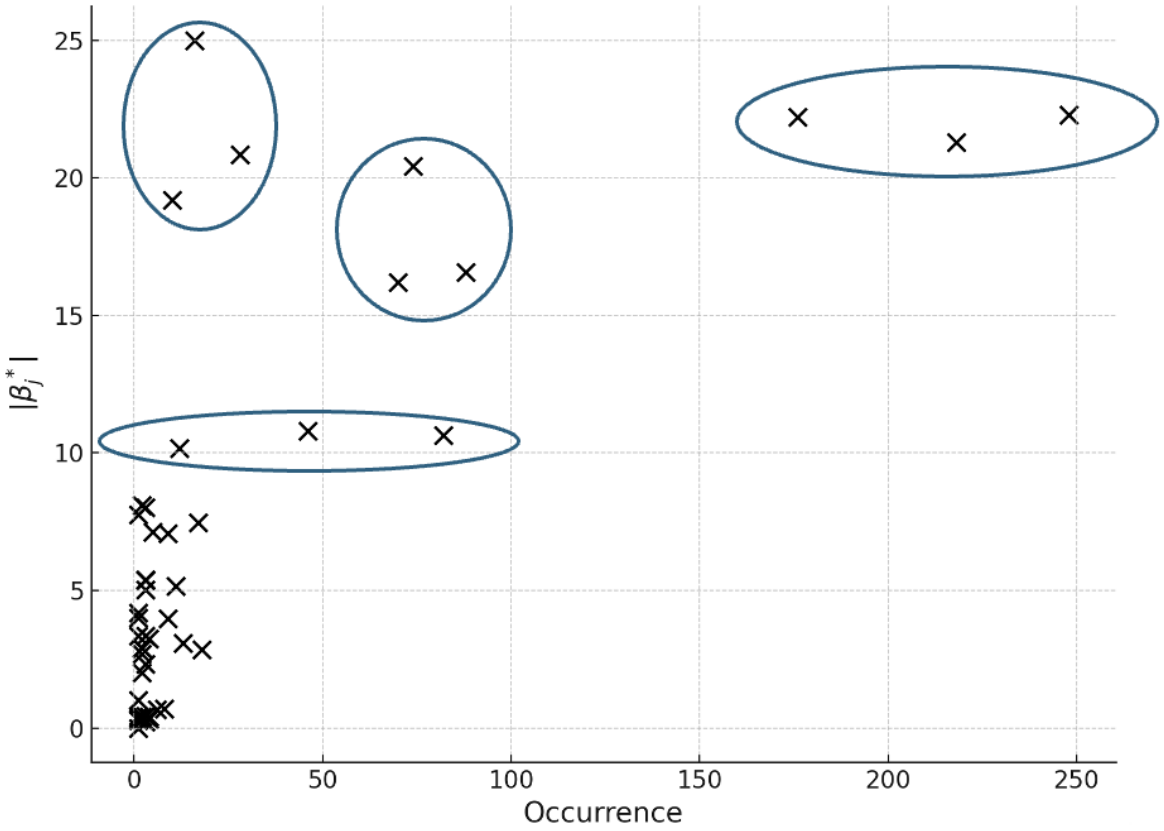


Figure 3. Sensitivity analysis of second-order groups used in the Sharma et al. method [13] for predicting ΔH_f^0 at 500 K. Each point represents a second-order group, with the vertical axis indicating its sensitivity ($|\beta_j^*|$), and the horizontal axis showing its occurrence in the dataset. Groups within the blue ellipses are both highly sensitive and frequently occurring, and were thus selected as the 12 most influential groups for our method to construct a reduced group set for further comparison with CG94 [10] and the Sharma model [13]. Notably, only 46 out of the 69 second-order groups in the Sharma et al. method were detected from all the molecules listed in Scott tables [16].

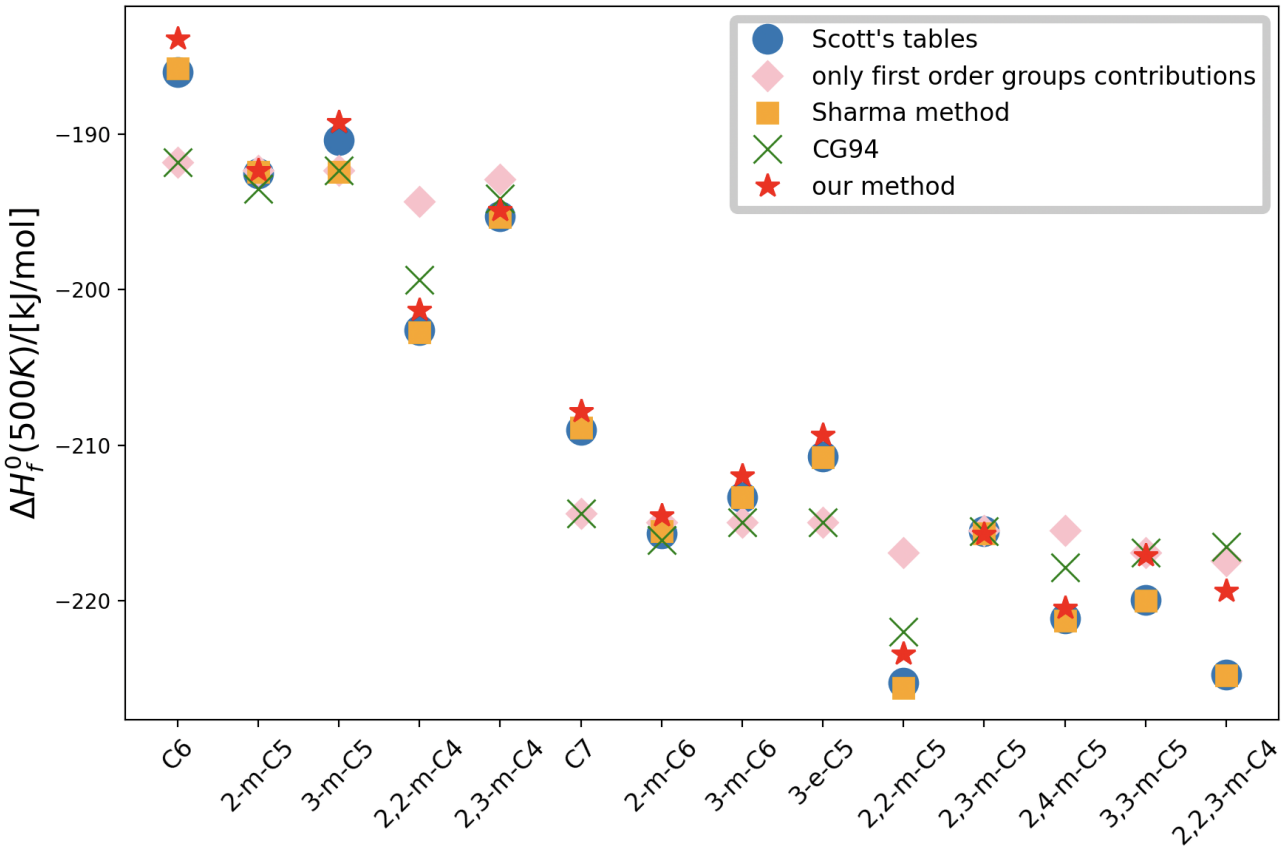


Figure 4. Comparison of predicted values of ΔH_f^0 at 500 K for various iso-alkanes using only first-order group contributions (pink rhomb), Sharma et al. method [13] (yellow squares), CG94 [10] (green crosses), our method (red stars), and the training set from the Scott tables [16] (blue circles). Using only first-order groups provides reasonably accurate predictions for less branched alkanes, the Sharma et al. method shows an excellent agreement with the Scott tables, CG94 a achieves a better accuracy for branched isomers by incorporating local structural corrections via second-order group correction.

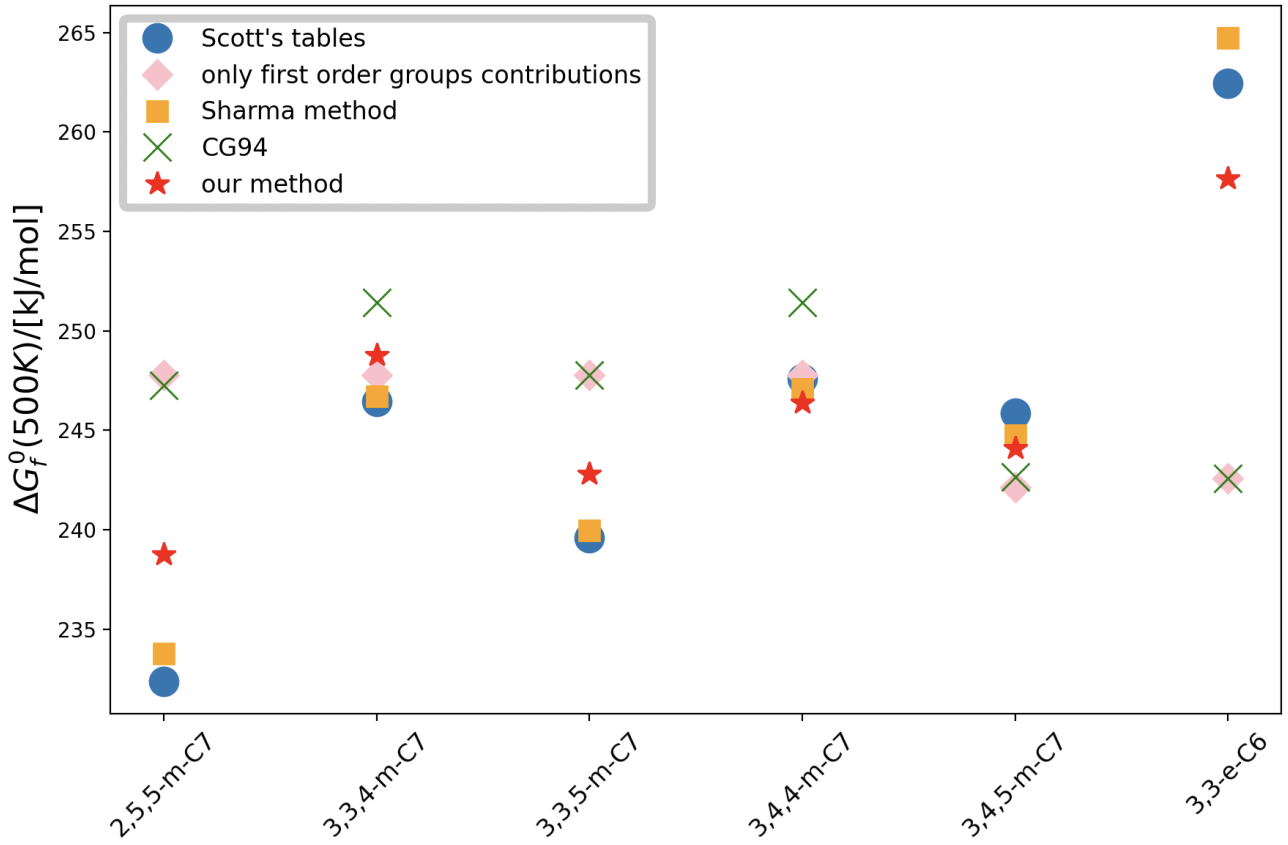


Figure 5. Comparison of predicted values of ΔG_f^0 at 500K for selected branched iso-alkanes for decane using different GCMs: using only first-order group contributions (pink rhomb), CG94 [10] (green crosses), the Sharma et al. method [13] (yellow squares), and our method (red stars), and compared to reference data from Scott thermochemical tables [16] (blue circles).

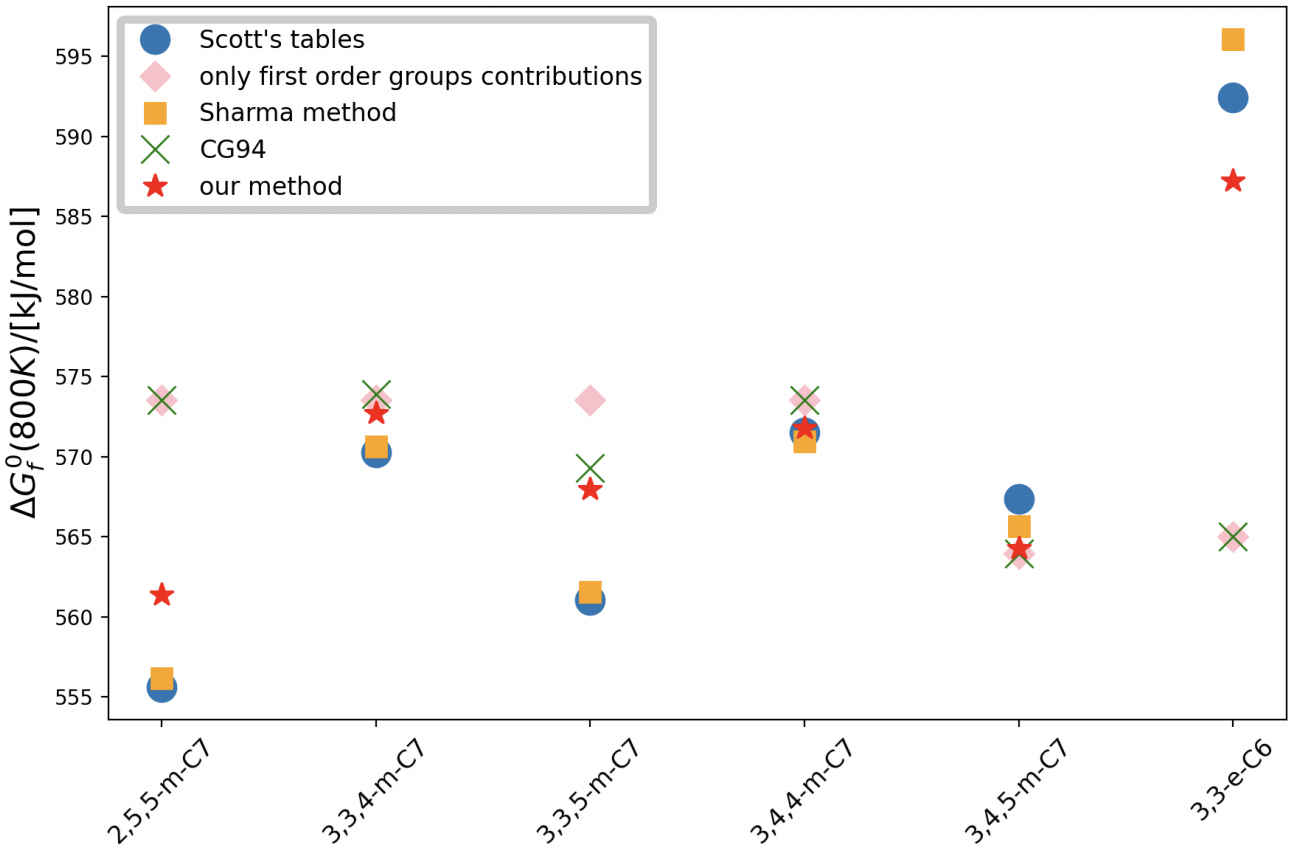


Figure 6. Comparison of predicted values of ΔG_f^0 at 800K for selected branched iso-alkanes for decane using different GCMs: using only first-order group contributions (pink rhomb), CG94 [10] (green crosses), the Sharma et al. method [13] (yellow squares), and our method (red stars), and compared to reference data from Scott thermochemical tables [16] (blue circles).

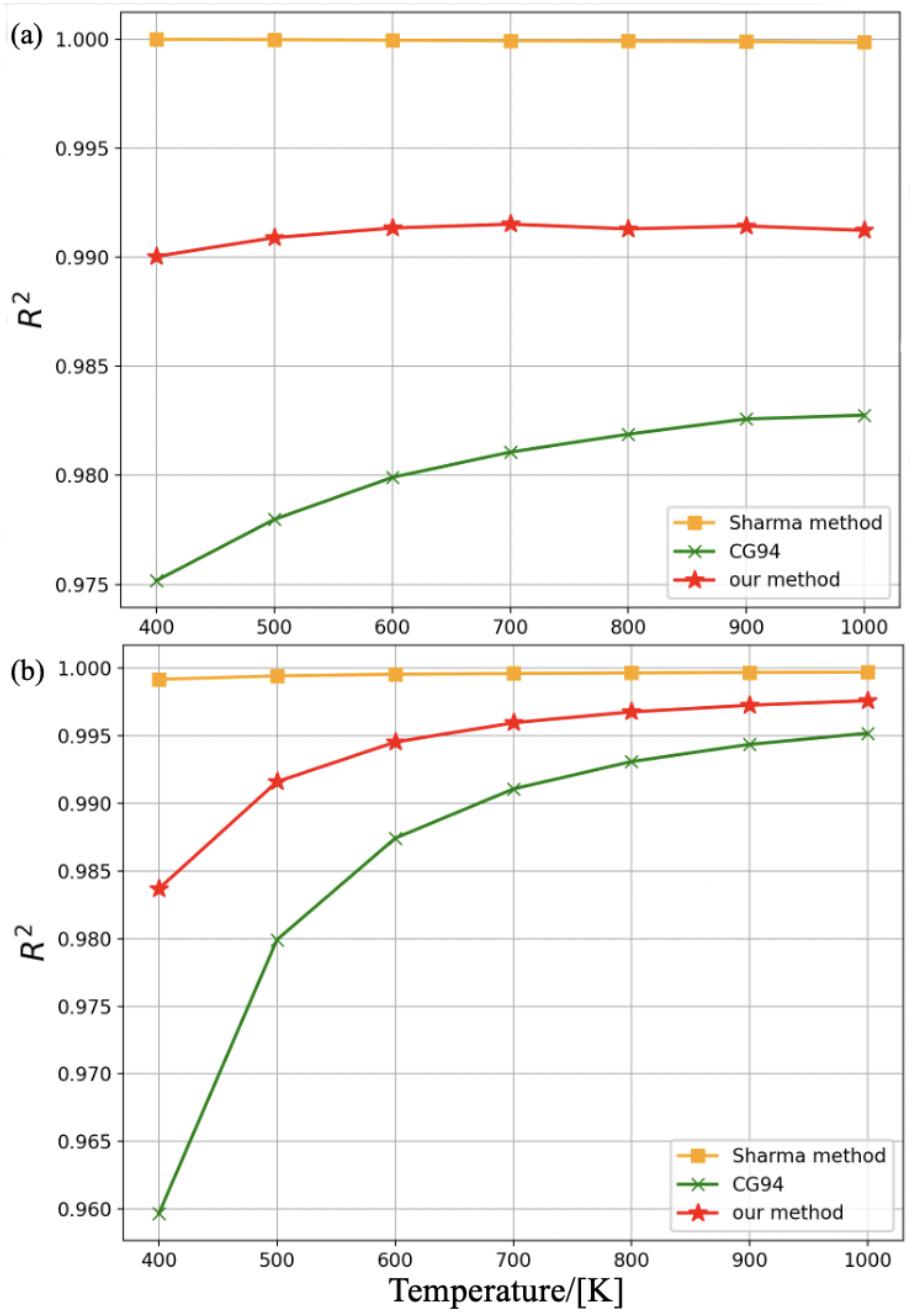


Figure 7. Temperature-dependent coefficient of determination (R^2) for thermochemical property predictions of CG94 (yellow), our method (blue), and the Sharma et al. method (green). (a) R^2 values for ΔH_f^0 predictions for a temperature range of 400–1500K. (b) R^2 values for ΔG_f^0 predictions. The Sharma et al. method maintains consistently high accuracy for all temperature, while our method exhibits strong performance with minor deviations at high temperatures. In contrast, CG94 shows lower R^2 values, particularly at lower temperatures, reflecting its limited structural resolution.

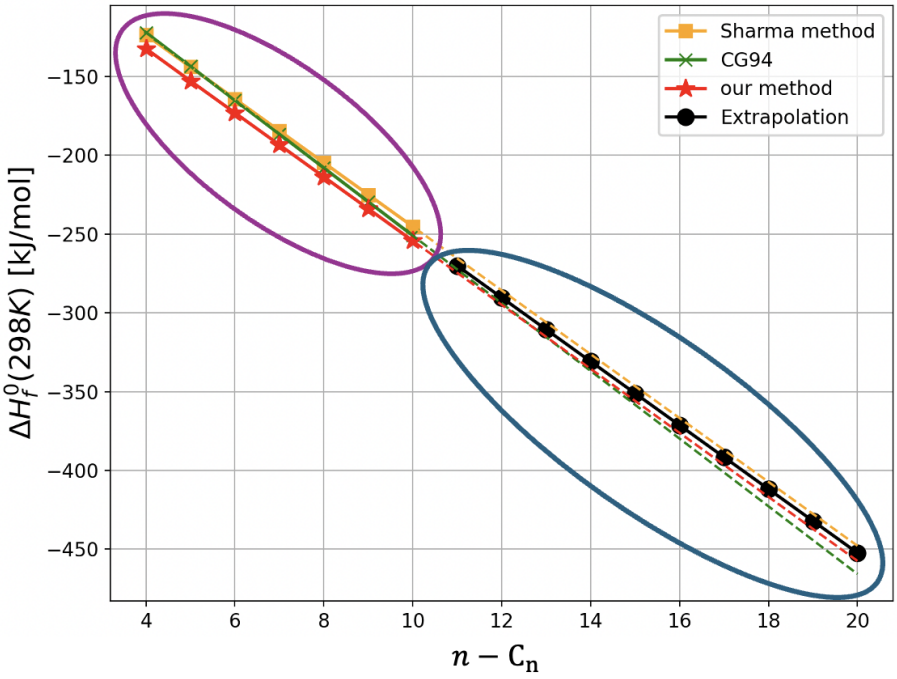


Figure 8. Prediction of $\Delta H_f^0(298\,\mathrm{K})$ for linear alkanes using three GCMs: Sharma et al. method [13] (yellow squares), CG94 [10] (green crosses), and our method (red stars). The black circle line represents extrapolated experimental values of C_{11}–C_{20}. The region on the left (purple ellipse) shows the fitted range C_4–C_{10}, while the right region (blue ellipse) indicates the extrapolation zone. The dashed lines correspond to linear trendlines for each method. Among the three models, the Sharma et al. method shows the best extrapolation performance, followed by our method, while CG94 exhibits the largest deviation from the experimental data.

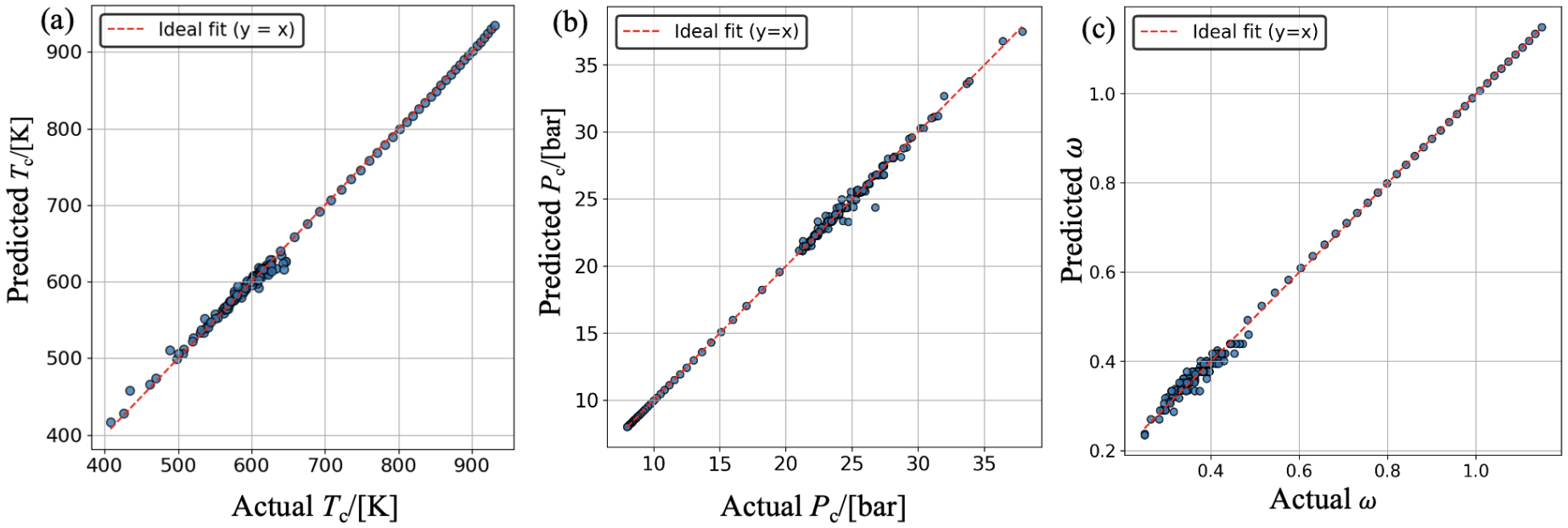


Figure 9. Parity plots comparing predicted and experimental values for (a) critical temperatures (T_c), (b) critical pressures (P_c), and (c) acentric factors (ω) using the proposed nonlinear regression. The red dashed line represents the ideal correlation ($y = x$). All three properties show strong agreement between predicted and actual values, highlighting the accuracy and robustness of the model for a diverse range of hydrocarbon structures.

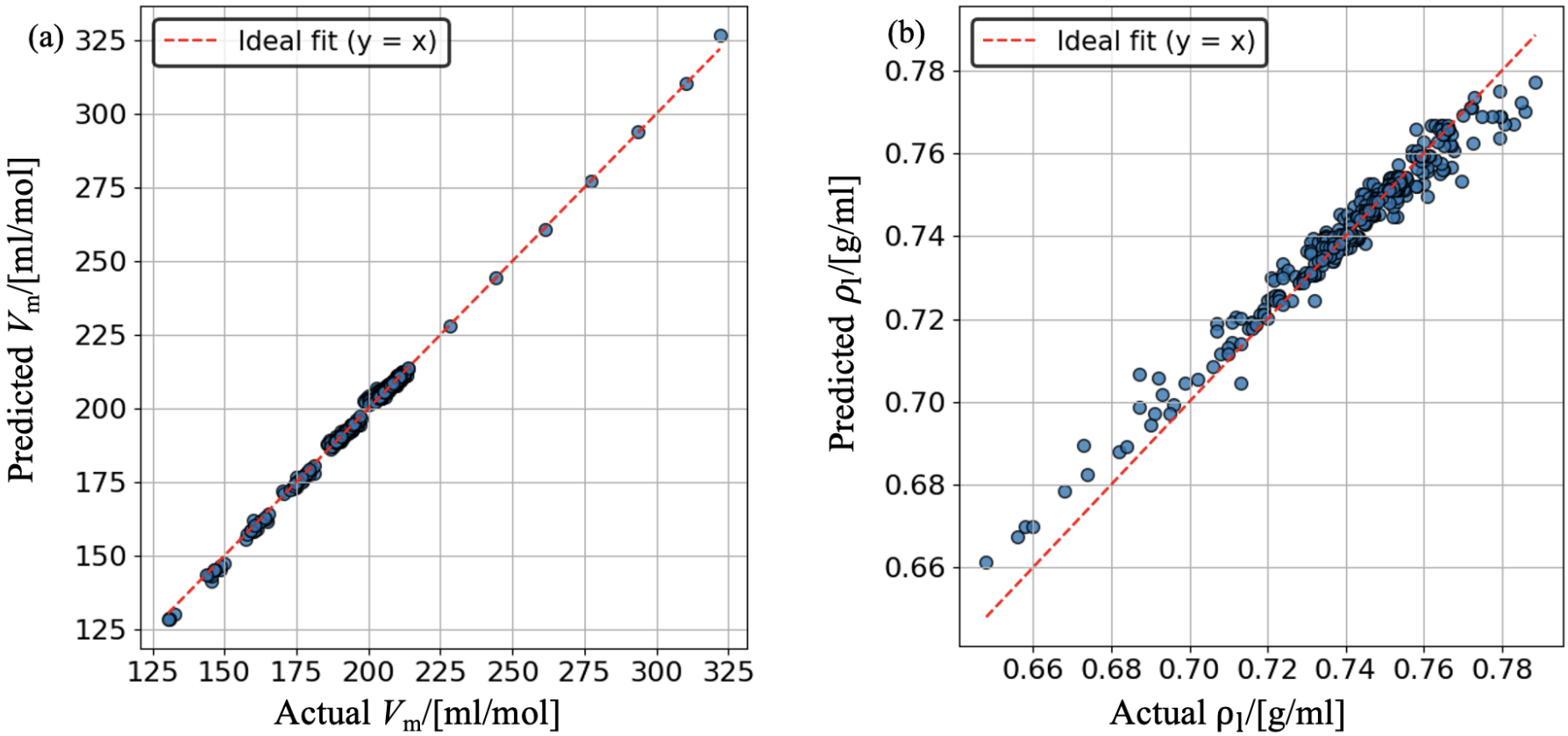


Figure 10. Parity plots comparing predicted and experimental values of (a) V_m and (b) the values of ρ_l. V_m were directly fitted using linear regression, while ρ_l were obtained by converting the predicted V_m values. The excellent agreement in (a) shows the suitability of V_m for linear regression CGMs while the slightly larger deviations in (b) reflect the additional complexity inherent in density predictions.

References

[1] S. van Bavel, S. Verma, E. Negro, and M. Bracht, *Integrating co_2 electrolysis into the gas-to-liquids–power-to-liquids process*, ACS Energy Letters 5 (2020), pp. 2597–2601.

[2] T.M. Letcher (ed.), *Chemical Thermodynamics for Industry*, 1st ed., Royal Society of Chemistry, Cambridge, UK, 2004.

[3] L. Tao, J.J. Jacobson, L. Zhang, M.A. Jackson, D.B. Hodge, C. Kinchin, and M. Wang, *Techno-economic analysis and life-cycle assessment of cellulosic isobutanol and comparison with cellulosic ethanol and n-butanol*, Biofuels, Bioproducts and Biorefining 11 (2017), pp. 965–980.

[4] F.M. Fraser and E.J. Prosen, *Heats of combustion of liquid n-hexadecane, 1-hexadecene, normal-decylbenzene, normal-decylcyclohexane, normal-decylcyclopentane, and the variation of heat of combustion with chain length*, Journal of Research of the National Bureau of Standards 55 (1955), pp. 329–333.

[5] E. Prosen, K. Pitzer, and F. Rossini, *Heats and free energies of formation of the paraffin hydrocarbons, in the gaseous state, to 1500 degree*, National Bureau of Standards 34 (1945), p. 403.

[6] R. Gani, *Group contribution-based property estimation methods: advances and perspectives*, Current Opinion in Chemical Engineering 23 (2019), pp. 184–196.

[7] Z. Li, L. Constantinou, R. Baur, D. Dubbeldam, S. Calero, S. Sharma, M. Rigutto, P. Dey, and T.J.H. Vlugt, *Review of group contribution methods for prediction of thermodynamic properties of long-chain hydrocarbons*, 2025, Molecule Physics. Submitted.

[8] A.L. Lydersen, *Estimation of critical properties of organic compounds by the method of group contributions*, Engineering Experiment Station Report 3, College of Engineering, University of Wisconsin, Madison, Wisconsin, 1955.

[9] K.G. Joback and R.C. Reid, *Estimation of pure-component properties from group-contributions*, Chemical Engineering Communications 57 (1987), pp. 233–243.

[10] L. Constantinou and R. Gani, *New group contribution method for estimating properties of pure compounds*, AIChE Journal 40 (1994), pp. 1697–1707.

[11] J. Abildskov, L. Constantinou, and R. Gani, *Towards the development of a second-order approximation in activity coefficient models based on group contributions*, Fluid Phase Equilibria 118 (1996), pp. 1–12.

[12] K.K. Yalamanchi, V.C. Van Oudenhoven, F. Tutino, M. Monge-Palacios, A. Alshehri, X. Gao, and S.M. Sarathy, *Machine learning to predict standard enthalpy of formation of hydrocarbons*, Journal of Physical Chemistry A 123 (2019), pp. 8305–8313.

[13] S. Sharma, J.J. Sleijfer, J. Op de Beek, S. van der Zeeuw, D. Zorzos, S. Lasala, M.S. Rigutto, E. Zuidema, U. Agarwal, R. Baur, S. Calero, D. Dubbeldam, and T.J.H. Vlugt, *Prediction of thermochemical properties of long-chain alkanes using linear regression: Application to hydroisomerization*, Journal of Physical Chemistry B 128 (2024), pp. 9619–9629.

[14] S. Hwang and J. Kang, *Group contribution-based graph convolution network: Pure property estimation model*, International Journal of Thermophysics 43 (2022), pp. 9–42.

[15] J. Gmehling, *Present status and potential of group contribution methods for process development*, Journal of Chemical Thermodynamics 41 (2009), pp. 731–747.

[16] D.W. Scott, *Correlation of the chemical thermodynamic properties of alkane hydrocarbons*, Journal of Chemical Physics 60 (1974), pp. 3144–3165.

[17] C.L. Yaws, *Yaws' Handbook of Thermodynamic and Physical Properties of Chemical Compounds*, 1st ed., Knovel, Houston, USA, 2003.

[18] C.L. Yaws, *Thermophysical Properties of Chemicals and Hydrocarbons*, 1st ed., William Andrew Inc., Norwich, USA, 2008.

[19] C.H. Achen, *Interpreting and Using Regression*, 1st ed., Sage Publications, Beverly Hills, CA, 1982.

[20] J.W. Johnson and J.M. LeBreton, *History and use of relative importance indices in organizational research*, Organizational Research Methods 7 (2004), pp. 238–257.

[21] R. Azen and D.V. Budescu, *The dominance analysis approach for comparing predictors in multiple regression*, Psychological Methods 8 (2003), pp. 129–148.
[22] E. Stefanis and C. Panayiotou, *Prediction of hansen solubility parameters with a new group-contribution method*, International Journal of Thermophysics 29 (2008), pp. 568–585.
[23] L. Constantinou, R. Gani, and J.P. O'Connell, *Estimation of the acentric factor and liquid molar volume at 298K through a new group contribution method*, Fluid Phase Equilibria 103 (1995), pp. 11–22.
[24] J. Marrero and R. Gani, *Group-contribution based estimation of pure component properties*, Industrial & Engineering Chemistry Research 40 (2001), pp. 5256–5267.
[25] L. Constantinou and V. Vassiliades, *Computer-aided molecular design using a linear group contribution method for the prediction of pure component properties: Applications to solvent selection*, in *Chemical Product Design: Towards a Perspective through Case Studies*, Elsevier, 2007, p. 34.
[26] A. Tihic, N. von Solms, M.L. Michelsen, G.M. Kontogeorgis, and L. Constantinou, *Analysis and applications of a group contribution sPC-SAFT equation of state*, Fluid Phase Equilibria 281 (2009), pp. 60–69.
[27] A. Groniewsky and B. Hégely, *Extension of Constantinou and Gani's group contribution method for vapor pressure and acentric factor estimation through automated group conversion*, Fluid Phase Equilibria 577 (2024), p. 113990.
[28] Y. Nannoolal, J. Rarey, and D. Ramjugernath, *Estimation of pure component properties part 2. estimation of critical property data by group contribution*, Fluid Phase Equilibria 252 (2007), pp. 1–27.
[29] R. Rajesh, R. Morales-Rodríguez, and R. Gani, *Group-contribution methods for the prediction of thermodynamic properties: Recent developments, limitations, and future improvements*, Industrial & Engineering Chemistry Research 61 (2022), pp. 17469–17493.
[30] N.M. Kuznetsov and S.M. Frolov, *Heat capacities and enthalpies of normal alkanes in an ideal gas state*, Energies 14 (2021), p. 2641.
[31] Y. Nannoolal, J. Rarey, D. Ramjugernath, and W. Cordes, *Estimation of pure component properties. part 1. estimation of the normal boiling point of non-electrolyte organic compounds via group contributions and group interactions*, Fluid Phase Equilibria 226 (2004), pp. 45–63.
[32] X. Wen and Y. Qiang, *A new group contribution method for estimating critical properties of organic compounds*, Industrial & Engineering Chemistry Research 40 (2001), pp. 20–32.
[33] W. Wakeham, G. Cholakov, and R. Stateva, *Liquid density and critical properties of hydrocarbons estimated from molecular structure*, Journal of Chemical and Engineering Data 47 (2002), p. 70.
[34] V. Villazón-León, A. Bonilla-Petriciolet, J. Tapia-Picazo, J. Segovia-Hernández, and M. Corazza, *A review of group contribution models to calculate thermodynamic properties of ionic liquids for process systems engineering*, Chemical Engineering Research and Design 185 (2022), pp. 458–480.
[35] C. Ihmels and J. Gmehling, *Extension and revision of the group contribution method gcvol for the prediction of pure compound liquid densities*, Industrial & Engineering Chemistry Research 42 (2002).
[36] Delft High Performance Computing Centre (DHPC), *DelftBlue Supercomputer (Phase 2)* (2024). Accessed on Jul 26, 2024.