Group Contribution-Based Graph Convolution Network: Pure Property Estimation Model

Sun Yoo Hwang1 · Jeong Won Kang1,2

Received: 6 June 2022 / Accepted: 16 June 2022 / Published online: 14 July 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
Properties data for chemical compounds are essential information for the design and operation of chemical processes. Experimental values are reported in the literature, but that are too scarce compared with exploding demand for data. When the data are not available, various estimation methods are employed. The group contribution method is one of the standards and simple techniques used today. However, these methods have inherent inaccuracy due to the simplified representation of the molecular structure. More advanced methods are emerging, including improved molecular representations and handling experimental data. However, such processes also suffer from a lack of valid data for adjusting many parameters. We suggest a compromise between a complex machine learning algorithm and a linear group contribution method in this contribution. Instead of representing a molecule using a graph of atoms, we employed bulkier blocks—a graph of functional groups. The new approach dramatically reduced the number of adjustable parameters for machine learning. The result shows higher accuracy than the conventional methods. The whole process was also examined in various aspects—incorporating uncertainties in the data, the robustness of the fitting process, and detecting outlier data.
Graphical Abstract

Keywords Artificial neural network · Graph convolution networks · Group contribution method · Machine learning · Thermodynamic property estimation

1 Introduction

Thermodynamic properties are essential and most basic information for chemical industries. However, the experimental values are often unavailable because experimental efforts cannot meet the demand for data for various chemical compounds. Therefore, the estimation methods are widely employed in the early stage of the investigation due to their convenience.

Group contribution methods have been developed by researchers such as Joback [1], Lydersen [2], Klincewicz [3], and Constantinou [4]. They assumed that the properties of the compounds depend only on their structure. Therefore, the functional groups were parametrized into linear combinations and used with non-linear functions. These methods provide quick estimation without any experimental processes. However, many of these methods failed to represent the compound uniquely, failing to distinguish the properties of isomers [5].

Constantinou and Gani [6] suggested the second-order groups with more complex structures and rules. As a result, it was possible to cover the different properties of isomers. Later, this technique was extended to the third-order groups by Marrero and Gani [7]. Moreover, Marrero and Pardillo [8] considered the interactions between groups and bonds, allowing more discrimination of isomers. However, these attempts typically include more parameters and complex implementation.

Recently, machine learning and digital image processing methods have been applied for property estimations. Goll [9], Gharagheizi [10], and Olayiwola [11] showed improved results by replacing the multi-linear regression with fully...
connected networks. However, this method did not include the connectivity of
groups. Lu [12], Kearnes [13], Sanchez-Lengeling [14], and others [15] showed suc-
cessful estimation of quantum mechanics calculation results with molecular graph
convolution network. However, they used about 130,000 to 1,000,000 data for train-
ing adjustable parameters, which is impossible to cover in thermophysical experi-
mental data.

Such limitations can be overcome by compromising the molecular graph and
the group contribution framework. Instead of using a fully connected graph of
atomic contributions, we employed a connected graph of bulky functional groups to
reduce the number of adjustable parameters. By using such a technique, adjustable
parameters can be effectively trained using a few thousands of data with reasonable
accuracies.

This study proposed a new model called the Group Contribution Graph Convo-
lution Neural Network (GC-gcn) model for estimating pure thermodynamic prop-
erties. The model maintains a conventional group contribution method framework,
while the group contributions are updated according to its adjacent group by graph
convolution network.

The fully connected networks are used for machine learning applications for
property estimation. The experimental data are from the NIST TDE SOURCE data-
base [16]. The capability of our model was tested by comparison with the commonly
used group contribution methods.

2 Theory

2.1 Group Contribution Method

The group contribution method is a widely used qualitative structure–property rela-
tions method. The procedure assumes that a functional group contributes to the
same properties for all the molecules. Some small but complex molecules can be
treated as independent groups for some group contribution methods. For example,
water (H2O) and methanol (CH3OH) are treated as a single group in UNIFAC [17].
However, due to the over-simplifying of the molecular structure, many methods
have questionable accuracy for complex molecules, including isomers. The higher-
order groups are suggested to overcome these limitations. However, such higher-
order methods are only effective for particular molecules and difficult to generalize
the rules of defining higher-order terms.

The general term for group contribution method for property $X$ is:

$$X = f(X_0 + \sum_j \omega_j \sum_i n_{j, i} R_{j, i})$$

(1)

where $f$ is a simple function, $X_0$ is a constant, $n_{j, i}$ is the number of $j$-th order of
groups $i$ in the molecule, $R_{j, i}$ is the contribution of $j$-th order of group $i$, and $\omega_j$ is a
weight for $j$-th order group.
2.2 Group Contribution Graph Convolution Network

With the advances in machine learning algorithms, many researchers tried to analyze property–structure relationships using innovative methods. Among them, converting molecules into a graph format and analyzing them through a graph convolution neural network was the most successful. Several applications are suggested [12–14, 18], but most require a lot of training data for adjusting parameters. In the case of thermophysical properties, the amount of measured data is relatively small, and they typically contain significant uncertainties.

To reduce the number of parameters, we simplified the graph of atomic constituents into the graph of group contributions. The group contribution graph network can be trained by a machine learning algorithm. The abbreviation of the new method is GC-gcn (Group contribution Graph Convolution Neural Network).

2.3 Molecule into a Graph of Group Contributions

A component is fragmented into individual groups, the same as other group contribution methods. Then, it is represented in a molecular graph $G(V, E)$, which is an undirected graph consisting of a vertex set $V(G)$ and an edge set $E(G)$. Each element $i \in V$ represents the consisting group of the compound, and it is labeled $I$ to $N$, $V(G) = \{v_1, v_2, v_3 \ldots v_N\}$. Each element $\{i, j\} \in E(\text{where} v_i, v_j \in V)$ represents the edge connecting vertices $v_i$ and $v_j$. The list of groups used in this study is summarized in Table 1. The table shows that the number of composing groups (48) is minimal compared with conventional group contribution models.

2.4 Group Feature Matrix

From the molecular graph $G(V, E)$, we obtain the group feature matrix $G^0$. The rows of the $G^0$ indicate the vertexes, and the columns indicate the group IDs. $G^0_{ij}$ equal to 1 means that the group ID of the $v_i$ is $j$, and the other elements are zero. For consistency of model inputs, we fixed the size of the group feature matrix $G^0$ to $(N_{\text{max}} \times N_G)$, where $N_{\text{max}}$ represents the maximum number of groups that consists of a compound and $N_G$ represents the number of groups. The compounds with more than $N_{\text{max}}$ groups are neglected. If the compound has fewer than $N_{\text{max}}$ groups, the vacant elements are filled with zeros.

2.5 Connectivity Matrix

The adjacency matrix $A$ was obtained from the molecular graph $G(V, E)$. $A_{ij} = 1$ if $\{i, j\} \in E$ and $A_{ij} = 0$ otherwise. However, we add an identity matrix if the adjacency matrix does not have information about the vertex itself.

$$A^I = A + I$$ (2)
Because the functions of the graph convolution network were designed to have values between $-1$ and $1$, we normalized the $\mathbf{A}^I$ with Symmetrically Laplacian and obtained the connectivity matrix $\mathbf{C}$. The process can be expressed as:

$$\mathbf{C} = \mathbf{D}^{-1/2} \mathbf{A}^I \mathbf{D}^{-1/2}$$  \hspace{1cm} (3)

where $\mathbf{D}$ is a degree matrix of the $\mathbf{A}^I$.

### 2.6 An Example of Group Contribution Graph Interpretation of a Molecule

This section shows a graph interpretation of 1-propanol, for example. The maximum number of groups $N_{max}$ was set to 5, and the groups used in this example are shown Table 2.

First, the compound is fragmented into groups and numbered. This process is shown in Fig. 1.

Then, the group feature matrix $\mathbf{G}^0$ can be represented as:

| ID | GROUP | INDEX | ID | GROUP | INDEX |
|----|-------|-------|----|-------|-------|
| 1  | $-\text{CH}_3$ | 25    | 25 | $-I$  | HALOGEN |
| 2  | $=\text{CH}_2$ | 26    | 26 | $-\text{OH}$ |  |
| 3  | $\equiv\text{CH}$ | 27    | 27 | $-\text{O}$ |  |
| 4  | $-\text{CH}_2$ | 28    | 28 | $-\text{O}$ | RING |
| 5  | $>\text{CH}$ | 29    | 29 | $-\text{O}$ | AROMATIC |
| 6  | $>\text{C}< $ | 30    | 30 | $=\text{O}$ |  |
| 7  | $=\text{CH}< $ | 31    | 31 | $-\text{NH}_2$ |  |
| 8  | $=\text{C}< $ | 32    | 32 | $>\text{NH}$ |  |
| 9  | $=\text{C}$ | 33    | 33 | $>\text{NH}$ | RING |
| 10 | $\equiv\text{C}$ | 34    | 34 | $>\text{NH}$ | AROMATIC |
| 11 | $-\text{CH}_2$ | RING  | 35 | $>\text{N}$ |  |
| 12 | $>\text{CH}$ | RING  | 36 | $>\text{N}$ | RING |
| 13 | $>\text{C}< $ | RING  | 37 | $>\text{N}$ | AROMATIC |
| 14 | $=\text{CH}$ | RING  | 38 | $-\text{N}$ |  |
| 15 | $=\text{C}< $ | RING  | 39 | $-\text{N}$ | RING |
| 16 | $=\text{C}$ | RING  | 40 | $-\text{N}$ | AROMATIC |
| 17 | $\equiv\text{C}$ | RING  | 41 | $\equiv\text{N}$ |  |
| 18 | $-\text{CH}_2$ | AROMATIC | 42 | $-\text{SH}$ |  |
| 19 | $>\text{CH}$ | AROMATIC | 43 | $-\text{S}$ |  |
| 20 | $=\text{CH}$ | AROMATIC | 44 | $-\text{S}$ | RING |
| 21 | $=\text{C}< $ | AROMATIC | 45 | $-\text{S}$ | AROMATIC |
| 21 | $-\text{F}$ | HALOGEN | 46 | $=\text{S}$ |  |
| 23 | $-\text{Cl}$ | HALOGEN | 47 | $-\text{O}[+]$ |  |
| 24 | $-\text{Br}$ | HALOGEN | 48 | $=\text{N}[+]$ | < |
where rows represent the group index of 1-propanol and columns represent its group ID. The adjacency matrix $A$ is represented as:

\[
\begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}
\]

where rows and columns represent the group index, and it is filled with 1 if they are connected. For example, the groups 1 and 2 are connected, so the $A_{1,2}$ and $A_{2,1}$ are 1. From (3) and (4), the connectivity matrix $C$ is calculated as:

\[
\begin{pmatrix}
0.50 & 0.41 & 0.00 & 0.00 \\
0.41 & 0.33 & 0.33 & 0.00 \\
0.00 & 0.33 & 0.33 & 0.41 \\
0.00 & 0.00 & 0.41 & 0.50
\end{pmatrix}
\]

where rows and columns both represent the group index, and it is filled with 1 if they are connected. For example, the groups 1 and 2 are connected, so the $A_{1,2}$ and $A_{2,1}$ are 1. From (3) and (4), the connectivity matrix $C$ is calculated as:

2.7 Graph Convolution

After the conversion, the graph convolution is performed next. In this step, the information of the groups is updated according to their surroundings. It starts with multiplying the group feature matrix $G^k$ and connectivity matrix $C$, obtaining $H^{k+1}$.

\[
H^{k+1} = C \times G^k
\]  

(4)
The group feature matrix $G^k$ is updated to $G^{k+1}$ with $H^{k+1}$ by the following equation.

\[
G^{k+1} = \text{ReLU}(H^{k+1} \times W^{k+1} + b^{k+1})
\]  

where $W^k$ is a weight and $b^k$ is a bias for the layer $k$. ReLU is the Rectified Linear Unit function which gives non-linearity. This procedure is repeated several times. After the graph convolution process is completed, $G^k$ represents the contribution of $v_i$. These terms are summed up column-wise, just as group contribution methods. In addition, we can use higher-order method developed recently [19] with more complexity. However, due to the limitation of experimental data for the training, we should keep the number of adjustable parameters minimum. Therefore, we concluded that the summation of contribution, slightly modified by the adjacent groups, will be enough for our purpose of property estimation. By this process, we could obtain molecular feature vector $M$.

### 2.8 An Example of Graph Convolution Network of Proposed Model

An example of graph convolution is described for 1-propanol in this section. The length of the $M$ is set to 6 and held graph convolution only once. The parameters used in this example are just for illustrative purposes, not actual fitted results. From the previous example of a graph representation of 1-propanol, we can obtain $H^0$ by Eq. [5] as:

\[
\begin{pmatrix}
0.50 & 0.41 & 0.00 & 0.00 \\
0.41 & 0.67 & 0.00 & 0.00 \\
0.00 & 0.67 & 0.00 & 0.41 \\
0.00 & 0.41 & 0.00 & 0.50
\end{pmatrix}
\]

$H^0$ contains the information for the neighbor groups. The third column refers to $>\text{CH}_2$ in the ring, so it is 0 for all rows. Because $-\text{CH}_3$ does not exist near group 3 of 1-propanol, $H_{3,1}^0$ is 0. On the other hand, since $>\text{CH}_2$ exists one more than $-\text{OH}$ near group 3, $H_{3,2}^0$ has a higher value than $H_{3,4}^0$. If we assume $W^1$ and $b^1$ are:

\[
\begin{pmatrix}
0.05 & 0.22 & 0.10 & 0.04 & -0.08 & 0.15 \\
-0.06 & 0.39 & 0.46 & -0.12 & 0.29 & 0.03 \\
0.07 & 0.43 & -0.43 & -0.41 & -0.48 & 0.33 \\
0.28 & 0.37 & 0.48 & 0.30 & -0.04 & 0.28
\end{pmatrix}
\]

We can obtain $G^1$ by (6) as:

\[
\begin{pmatrix}
0.12 & 0.39 & 0.36 & 0.09 & 0.20 & 0.20 \\
0.62 & 0.99 & 0.99 & 0.58 & 0.80 & 0.72 \\
0.22 & 0.56 & 0.65 & 0.19 & 0.32 & 0.28 \\
1.06 & 1.29 & 1.37 & 1.05 & 1.04 & 1.10
\end{pmatrix}
\]

Then by summing all the values of columns in $G^1$, we could obtain $M$ as:
$M$ is treated with fully connected layers for the regression. The mathematical representation of a property $X$ is given as follows:

$$
X = \sum_{i \in M} ReLU(W_2 ReLU(W_1 M + b_1) + b_2)
$$

(6)

where $M$ is the one-dimensional molecular vector that came from the read-out layer, $W_1, W_2$ are weights and $b_1, b_2$ are bias, and $ReLU$ is the Rectified Linear Unit function. Figure 2 shows the outline of the calculation procedure.

### 3 Datasets and Models

#### 3.1 Reduction of Parameters

The replacement of atomic segments in molecular graph with bulkier functional groups is expected to reduce the number of parameters. Although many property estimation models [12–14, 18] used graph convolution networks, no method was designed for thermophysical properties. Therefore, we also implemented a new simple Molecular Graph Convolution Network (M-gcn) for the illustrative comparison. The specific details can be found in Appendix A. $N_{\text{max}}$ is set to 25, and the length of the group feature is set to 150.

Table 3 shows the number of adjustable parameters for the two models. The number of parameters was dramatically decreased for GC-gcn model. Since the M-gcn uses multiple graphs to distinguish the bonds between atoms, a graph convolution network with a more complex architecture [20] must be used. In the case of the proposed method (GC-gcn), the bonds between atoms are lumped into a group description.
3.2 Datasets

The structural information required for compounds is constituent groups and their connectivity. The data were obtained by analyzing about 50,000 PubChem bulk SMILES databases [21]. The pure-component physical properties data were from NIST TDE SOURCE database [16]. The number of properties data used in this study is summarized in Table 4.

3.3 Consideration of Uncertainties in the Data

The advantage of using the NIST SOURCE database [16] as a training set is that each data point has critically evaluated uncertainty [22]. The conventional cost functions—such as average absolute error (AAE) or means square error (MSE)—can be modified to reflect the uncertainty of each point.

\[
\text{Cost(Loss)Function} = \frac{1}{N} \sum (X_{i}^{\text{pred}} - X_{i}^{\exp})^2 e^{-3\sigma_i^2}, \quad \sigma_i = \frac{X_{i}^{\text{unc}}}{X_{i}^{\exp}} \tag{7}
\]

where \( N \) is the number of data, \( X_{i}^{\text{pred}} \) and \( X_{i}^{\exp} \) are the estimated result and experimental data \( i \), \( \sigma_i \) is an uncertainty ratio of data \( i \), and \( X_{i}^{\text{unc}} \) is an uncertainty of data \( i \).

The cost function is slightly modified from the conventional ratio-based approaches. The weight of the square error decreases as the value of \( \sigma \) increases. Data with high \( \sigma \) are penalized because of a high uncertainty ratio. On the other hand, when the \( \sigma \) decreases near zero, the exponential term approaches one. The modified scheme prevents exceptionally high weight.

Table 3 The number of adjustable parameters of the models

| Model                  | Molecular Graph Convolution Network (M-gcn) | Proposed Method (GC-gcn) |
|------------------------|-----------------------------------------------|----------------------------|
| Total Parameters       | 51,291,073                                    | 61,121                     |
| Size of Input          | 1200                                          | 1200                       |
| Total Parameters/Size of Input | 42,743                                        | 51                         |

Table 4 The number of pure-component thermophysical properties used in this study

| Property                  | Symbol | Unit | No. of data |
|---------------------------|--------|------|-------------|
| Normal Boiling Point      | \( T_b \) | K    | 16,192      |
| Normal Melting Point      | \( T_m \) | K    | 6,583       |
| Critical Temperature      | \( T_c \) | K    | 15,771      |
| Critical Pressure         | \( P_c \) | kPa  | 11,801      |
| Critical Volume           | \( V_c \) | l/mol | 555         |
| Flash Point               | \( T_f \) | K    | 896         |
3.4 Model Training and Evaluation

The proposed model (GC-gcn) was constructed using the open-source toolkit RDkit, Keras, Numpy. The composed network may fail to converge or fall into local minimum when the properties data are insufficient. Such behavior can be prevented by initializing the parameters with a well-known prediction model. For such reason, the model was first trained from the predicted data produced by Joback & Reid group contribution method [1]. After the initialization, the model was trained and verified with actual experimental datasets. Five-fold cross-validation has been used for determining the training rates. Adam [23], with 0.001 learning rates, was used as the optimizer. The parameter fitting processes with different random initializations (including train-test splitting) were conducted ten times for the reliability test.

4 Result and Discussion

4.1 Result of Estimation

The result of estimation for the proposed model (GC-gcn) is summarized and compared with the well-known conventional group contribution models in Table 5. The results are analyzed by average absolute error (AAE) and average relative error (ARE) between estimated and experimented data.

\[
AAE = \frac{1}{N} \left| X_{\text{Predicted}} - X_{\text{Experimental}} \right|
\]  

Table 5 Comparison property estimation results

| No. of data | GCM* | Present Work |
|-------------|------|--------------|
| \(T_b\)     | 15,080 | AAE (K) | 17.9 | 7.96 |
|             |       | ARE (%)  | 3.83 | 1.76 |
| \(T_c\)     | 763   | AAE (K) | 22.3 | 7.79 |
|             |       | ARE (%)  | 3.25 | 1.11 |
| \(P_c\)     | 508   | AAE (kPa) | 213 | 178 |
|             |       | ARE (%)  | 7.57 | 6.98 |
| \(V_c\)     | 1743  | AAE (L/mol) | 1.75×10^{-2} | 8.02×10^{-3} |
|             |       | ARE (%)  | 5.17 | 2.66 |
| \(T_m\)     | 1081  | AAE (K) | 41.8 | 20.8 |
|             |       | ARE (%)  | 14.0 | 7.38 |
| \(T_f\)     | 854   | AAE (K) | 13.7 | 10.4 |
|             |       | ARE (%)  | 3.98 | 3.17 |

*GCM: Marrero&Gani group contribution method [7] for \(T_b, T_c, P_c, V_c, T_m,\) and Rowley&Wilding group contribution method [24] for \(T_f\)
where $N$ is the number of data.

For comparisons, we used Marrero and Gani method [7] for most properties and Rowley and Wilding method [24] for the flashpoints. The group contribution methods cannot predict all the properties due to limitations in the representation of the molecules. In contrast, the proposed model can cover all the molecules in the database. For fair comparisons, we only compared the properties of molecules when the conventional method can predict the properties.

As shown in Table 5, the proposed model prediction is much better for all properties considered in this work ($T_b$, $T_c$, $P_c$, $V_c$, $T_m$, and $T_f$). For the boiling point and the critical point, the improvements are evident. On the other hand, for flashpoints and melting points, the qualities of data points are not very good, and the improvements are substantial. Figure 3 shows a plot of experimental values vs. predicted values for two models. The plotted points are more concentrated near the crossing-line for the GC-gcn model than the group contribution models.

### 4.2 Properties of Isomers

Distinguishing the properties of isomers is a challenging task for the group contribution methods. Several methods [6, 7] introduced the second or the third order to characterize the properties. However, such methods introduce additional complexity, and it is difficult to generalize the scheme for representing molecules. The proposed method uses the connectivity information contained group contribution graph and successfully distinguishes the difference in isomer properties.

The results are shown in Table 6, 7, 8, and 9 for n-boiling points. The proposed GC-gcn model shows a correct trend with experimental data, while the conventional group contribution cannot distinguish the difference in boiling points.

### 4.3 Robustness in fitting

Due to the limited availability of experimental data points, robustness to outliers is critical for property estimation models. Some weird data points where data are scarce may influence the whole estimation network. In this section, we would like to present the robustness of the proposed model compared with a more comprehensive molecular version of the graph convolution network (M-gcn). The M-gcn model is described in Appendix A.

Figure 4 and Table 10 show the estimation results of methyl branched hydrocarbon at carbon numbers 14 to 18 using GC-gcn and M-gcn. Interestingly, the previous version (9.0) of NIST SOURCE database [16] contains a suspected outlier(CASRN: 2882–96-4, Name: 3-methyl pentadecane, Boling point: 539.45 K), but it was corrected (548.16 K) in version10.1 [25]. In addition, Yaws [26] reported some data points, but they are not evident as experimental
or estimated values. Two models (GC-gcn and M-gcn) are trained, including the suspected data point, and the results are pretty different.

As shown in Fig. 4 and Table 10, the M-gcn overfit the outlier, and the prediction matches the outlier data point exactly. Such behavior is due to an excessive number of adjustable parameters. However, GC-gcn shows robust behavior even though the training set contains the outlier data point. Furthermore, the whole
prediction results match the data points reported by Yaws [26]. This result shows the robustness of our model compared with a model with more parameters.

### Table 6 Comparison of estimated boiling points for isomers example 1

| CAS          | Name                    | SMILES                  |
|--------------|-------------------------|-------------------------|
| 1067–08-9    | 3-Ethyl-3-methylpentane | CCC(C)(C)CC             |
| 563–16-6     | 3,3-Dimethylhexane     | CCCC(C)(C)CC            |

| Exp. $T_b$ (K) | MG $T_b$ (K) | Proposed Method $T_b$ (K) |
|----------------|--------------|----------------------------|
| 391.41         | 385.12       | 392.55                     |
| 379.62         | 379.62       | 386.12                     |

Exp: experimental data; MG: Marrero&Gani group contribution method [7]

### Table 7 Comparison of estimated boiling points for isomers example 2

| CAS          | Name                    | SMILES                  |
|--------------|-------------------------|-------------------------|
| 2453–00-1    | 1,3-Dimethylcyclopentane | CC1CCC(C1)C            |
| 2452–99-5    | 1,2-Dimethylcyclopentane | CC1CCCCC1C             |

| Exp. $T_b$ (K) | MG $T_b$ (K) | Proposed Method $T_b$ (K) |
|----------------|--------------|----------------------------|
| 364.05         | 372.75       | 367.45                     |
| 366.47         | 366.47       | 373.93                     |

Exp: experimental data, MG Marrero&Gani group contribution method [7]

### 4.4 Detection of outliers

There are various reasons for outliers of experimental properties databases – simple typos, careless experimental procedures, mistakes in database construction, etc. As shown in Sect. 4.4, the proposed model is resistant to outliers. Thus, our new model can detect possible outliers in the extensive database. A similar process has been implemented in the construction of databases, and we only show a small set of
outliers in the flashpoints. The points deviate excessively from the estimation values are screened and classified as outliers.

Table 11 shows the flashpoints of 1-pentene and 1-hexene. The experimental values are reported in different data sources [27, 28], but the flash point of longer chain molecules has a lower flash point. One of the data points is definitely an outlier. From the estimation result, we can conclude that the flash point of the 1-pentene is suspected.
similar situation can be found for 1-alkanols [29–31], shown in Table 12. In this case, the flash point of 1-pentadecanol is suspected.

5 Conclusion

This study suggested a new architecture for thermodynamic properties estimation, a hybrid form of the group contribution method, and a molecular graph convolution network. The application of graph convolution network has led to successful unique group representation and isomer discrimination of a molecule. Compared with conventional group contribution methods, the proposed method shows a more accurate estimation and broader applicability of molecular structures. In addition, simplification using bulkier groups has reduced the number of parameters to fit while maintaining the connectivity among the groups. Finally, the robustness of the model has been examined by comparison with the molecular graph convolution network. The proposed model can be further extended to other properties when the experimental data are available.

5.1 Implementation

A web page was constructed to use the new Group Contribution graph convolution network (GC-gcn) model conveniently. The result of this work can be accessed at https://www.chemical.org/products/gcgcn/online. Figure 5 shows a
| nC | Chemical Name    | SMILES                        | Experimental $T_b$ (K) | M-gcn $^*$ | Proposed Method       |
|----|-----------------|--------------------------------|------------------------|------------|------------------------|
|    |                 |                                |                        | Prediction (K) | Absolute Error (K)   | Prediction (K) | Absolute Error (K) |
| 13 | 3-methyl dodecane | CCCCCCCCCCCCCC(C)CC           | 503.45                 | 490.21     | 13.24                  | 498.02         | 5.43               |
| 14 | 2-methyl tridecane | CCCCCCCCCCCCCC(C)C           | 521.05                 | 507.86     | 13.19                  | 519.53         | 1.52               |
| 15 | 2-methyl tetradecane | CCCCCCCCCCCCCC(C)C           | 538.35                 | 521.88     | 16.47                  | 535.56         | 2.79               |
| 16 | 3-methyl pentadecane | CCCCCCCCCCCCCC(C)CC          | 555.15                 | 539.51     | 15.64                  | 552.69         | 2.46               |
| 17 | 2-methyl hexadecane | CCCCCCCCCCCCCC(C)C           | 570.15                 | 554.14     | 16.01                  | 566.13         | 4.02               |
| 18 | 3-methyl heptadecane | CCC(C)CCCCCCCCCCCCC           | 586.15                 | 571.16     | 14.99                  | 582.81         | 3.34               |

$^*$The Yaws Handbook of Physical Properties for Hydrocarbons [26]

$^{**}$Molecular graph convolution network
Table 11  Comparison of experimental data with estimated result using GC-gcn model, example 1

| CAS     | 109–67-1 | 592–41-6 |
|---------|----------|----------|
| Name    | 1-pentene | 1-hexene |
| SMILES  | C=CCCCC  | C=CCCCCC |
| Exp. $T_f$ (K) | 255.15 [27] | 247.15 [28] |
| Estimated $T_f$ (K) | 220.52 | 241.89 |

Table 12  Comparison of experimental data with estimated result using GC-gcn model, example 2

| CAS     | 112–72-1 | 629–76-5 | 36,653–82-4 |
|---------|----------|----------|-------------|
| Name    | 1-Tetradecanol | 1-Pentadecanol | 1-Hexadecanol |
| SMILES  | CCCCCCCCCCCCCCO | CCCCCCCCCCCCCCCCO | CCCCCCCCCCCCCCCCOCO |
| Exp. $T_f$ (K) | 413.15 [29] | 383.15 [30] | 427.15 [31] |
| Estimated $T_f$ (K) | 401.76 | 409.28 | 416.85 |

Fig. 5  An example of estimation result using the present work (GC-gcn) on online platform
result example of a calculation result on an online platform. The input requires a SMILES representation of the molecule. The uncertainty of the model estimation result is also shown. The uncertainty of the machine learning prediction will be discussed in a different material.

Appendix A

Description of the Molecular Graph Convolution Network (M-gcn) used in this work.

Similar to the proposed model GC-gcn, the compound is represented in molecular graph $G'(d', V', E')$. $d'$ represents the dimension of the graph, $V'$ represents the vertex, and $E'$ represents the edge. A vertex should not need to be connected to another vertex but must be connected once in the entire dimension $d'$. The atom feature (or called as node feature) is represented at $G'(V')$, while the edge feature is represented at $G'(d', E')$.

The rows of $V'$ indicate the vertex, and the columns indicate the atom types of the vertex. For consistency of model inputs, the shape of the atom feature matrix is fixed to $(N_{\text{max}} \times N_A)$, where $N_{\text{max}}$ represents the maximum number of atoms and $N_A$ represents the number of atom descriptors. Dimension $d'$ represents the type of connection between vertices. The description of atoms and bonds is summarized in Table 15. To consider the atom itself, the identity matrix is generally added to $G'(d', E')$. The functions of multi-dimensional graph convolution networks are also designed to have values between -1 and 1. So matrix must be normalized with symmetrical Laplacian.

$$B' = D'^{-1/2}(G'(d', E') + I)D'^{-1/2}$$  \hspace{1cm} (11)

where $B'$ is called the bond feature matrix, $D'$ is the degree matrix of $G'(d', E') + I$, and $I$ is the identity matrix.

| Table 15 | Descriptors of the molecule used in M-gcn |
| --- | --- |
| Feature | Description | Size |
| Atom | C, N, O, F, S, Cl, Br, I, | 8 |
| Bond $(d')$ | single, double, triple, aromatic, or is in ring | 5 |

| Table 16 | Simplified description used in M-gcn example |
| --- | --- |
| Feature | Description | Size |
| Atom | C, N, O (one-hot) | 3 |
| Bond | single, double, triple, aromatic, or is in ring | 5 |
An Example of a Molecular Graph Interpretation of a Molecule—Glycolonitrile

The maximum number of groups \( N_{\text{max}} \) was set to 4, and the descriptor we used in this example is shown in Table 16.

First, atoms are numbered. The process is shown in Fig. 6.

From Fig. 6, we can obtain the atom feature matrix as:

\[
V' = V'
\]

\[
\begin{bmatrix}
    C & O & N \\
    \{1\} & 0 & 0 & 1 \\
    \{2\} & 1 & 0 & 0 \\
    \{3\} & 1 & 0 & 0 \\
    \{4\} & 0 & 1 & 0 \\
\end{bmatrix}
\]

\( V' \) is also can be expressed as \( V'^0 \), where 0 indicates that any processing has been held. The bond feature matrix \( B'_{(d', j, k)} \), where \( d' \) is the dimension, \( j \) and \( k \) are the indexes of the atom, can be obtained by Eq. 11 as:

\[
G'(1, E') = \begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 \\
0 & 0 & 1 & 0 \\
\end{pmatrix}, \quad B' = \begin{pmatrix}
0.50 & 0 & 0 & 0 \\
0 & 0.50 & 0.41 & 0 \\
0 & 0.41 & 0.33 & 0.41 \\
0 & 0 & 0.41 & 0.50 \\
\end{pmatrix}
\]

\[
G'(3, E') = \begin{pmatrix}
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\end{pmatrix}, \quad B' = \begin{pmatrix}
0.5 & 0.5 & 0 & 0 \\
0.5 & 0.5 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
\end{pmatrix}
\]

Because there is no double bonds and rings in glycolonitrile, the dimensions 2, 4 and 5 of \( G'(d, E') \) are zero matrix.

\[
G'(2, E') = G'(4, E') = G'(5, E') = \begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\end{pmatrix}
\]
Multi-Dimensional Graph Convolution

In the multi-dimensional graph convolution process, or message passing phase, the contributions of atoms are updated according to their neighbor atoms and bond types. Unlike a graph convolution network, this procedure consists of two steps. First, the atoms are updated to each dimension. The result of this process is called the message passing function \( H^{k+1} \), which can be expressed as follows:

\[
H^{k+1}(i) = \text{ReLU}(B'(i) \times V^k \times W^{k+1} + b^{k+1})
\]

(12)

where \( W^k \) is a weight and \( b^k \) is a bias for the layer \( k \). \( \text{ReLU} \) is the Rectified Linear Unit function. In this work, we assume that there is no interaction across the graph dimensions. Then, the contribution of atoms \( V^k \) is updated as:

\[
V^{k+1} = \text{ReLU}(W^{k+1} \times \text{Concat}_{i=1}^{D} H^{k+1}(i))
\]

(13)

where \( \text{Concat} \) is concatenating the all the matrix, \( W^{k} \) is a weight, \( \text{ReLU} \) is the Rectified Linear Unit function.

Acknowledgements This work was supported by National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) [grant numbers NRF-2021R1A5A6002853 and NRF-2019M3E6A1064876]

Author Contributions SYH contributed to method implementation, data preparation, and machine learning and wrote the main manuscript. JWK contributed to supervision, manuscript preparation, and manuscript review.

Funding None.

Declarations

Competing interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. K.G. Joback, R.C. Reid, Chem. Eng. Commun. 57, 233 (1987)
2. A.L. Lydersen, Univ. Wisconsin Coll. Eng., Eng. Exp. Stn. Rep. 3, 1 (1955).
3. K.M. Klincewicz, R.C. Reid, AIChE J. 30, 137 (1984)
4. L. Constantinou, S.E. Prickett, M.L. Mavrovouniotis, Ind. Eng. Chem. Res. 32, 1734 (1993)
5. R. Gani, Curr. Opin. Chem. Eng. 23, 184 (2019)
6. L. Constantinou, R. Gani, AIChE J. 40, 1697 (1994)
7. J. Marrero, R. Gani, Fluid Phase Equilib. **183–184**, 183 (2001)
8. J. Marrero-Morejón, E. Pardillo-Fontdevila, AIChE J. **45**, 615 (1999)
9. E.S. Goll, P.C. Jurs, J. Chem. Inf. Comput. Sci. **39**, 1081 (1999)
10. F. Gharagheizi, R.F. Alamdari, M.T. Angaji, Energy Fuels **22**, 1628 (2008)
11. F. Yusuf, T. Olayiwola, C. Afagwu, Fluid Phase Equilib. **531**, 112898 (2021)
12. C. Lu, Q. Liu, C. Wang, Z. Huang, P. Lin, L. He, Proc. Conf. AAAI Artif. Intell. **1**, 1052 (2019)
13. S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley, J. Comput. Aided. Mol. Des. **30**, 595 (2016)
14. B. Sanchez-Lengeling, J. N. Wei, B. K. Lee, R.C. Gerkin, A. Aspuru-Guzik, A. B. Wuitschko, arXiv preprint arXiv:1910.10685 (2019).
15. O. Wieder, S. Kohlbacher, M. Kuenemann, A.aron, P. Ducrot, T. Seidel, T. Langer, Drug Discov. Today Technol. **37**, 1 (2020)
16. M. D. Frenkel, R. D. Chirico, V. Diky, K. Kroenlein, C. Muzny, A. F. Kazakov, J. Kang, Thermo-Data Engine (TDE) Version 9.0 (Pure Compounds, Binary Mixtures, Ternary Mixtures, and Chemical Reactions); NIST Standard Reference Database 103b (NIST, 2014), https://www.nist.gov/publications/thermodata-engine-tde-version-90-pure-compounds-binary-mixtures- ternary-mixtures-and (Accessed June 16, 2022)
17. A. Fredenslund, *Vapor-liquid equilibria using UNIFAC: a group-contribution method* (Elsevier, Amsterdam, 2012), p. 40
18. J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Proc. Mach. Learn. Res. **1**, 1263 (2017)
19. O. Vinyals, S. Bengio, M. Kudlur, arXiv preprint arXiv:1511.06391 (2015)
20. Y. Ma, S. Wang, C.C. Aggarwal, D. Yin, J. Tang, Proc. (Conf. Data Mining, SIAM Int, 2019). https://doi.org/10.1137/1.9781611975673.74
21. S. Kim et al., Nucleic Acids Res. **47**, D1102 (2019). https://doi.org/10.1093/nar/gky1033
22. L. Kirkup, R.B. Frenkel, *An Introduction to Uncertainty in Measurement: Using the GUM (Guide to the Expression of Uncertainty in Measurement)* (Cambridge University Press, Cambridge, 2006), p. 97
23. D. P. Kingma, J. Ba, arXiv preprint arXiv:1412.6980 (2014) https://doi.org/10.48550/arXiv.1412.6980
24. J.R. Rowley, R.L. Rowley, W.V. Wilding, Proc. Saf. Prog. **29**(4), 353 (2010)
25. V. Diky, C. Muzny, A. Smolyanitsky, A. Bazyleva, R. Chirico, J. Magee, Y. Paulechka, A. Kazakov, S. Townsend, E. Lemmon, M. D. Frenkel, K. Kroenlein, ThermoData Engine (TDE) Version 10.1 (Pure Compounds, Binary Mixtures, Ternary Mixtures, and Chemical Reactions); NIST Standard Reference Database 103b (NIST, 2016). https://www.nist.gov/publications/thermodata-engine-tde-version-101-pure-compounds-binary-mixtures-ternary-mixtures-and. Accessed June 15, 2022
26. C.L. Yaws, *Thermophysical Properties of Chemicals and Hydrocarbons* (William Andrew, New York, 2008), pp. 75–84
27. J.M. Kuchta, *Investigation of Fire and Explosion Accidents in the Chemical, Mining, and Fuel-Related Industries-a Manual* (Bulletin (Bureau of Mines, Washington DC, 1986)
28. J. Bond, *Sources of Ignition: Flammability Characteristics of Chemicals and Products* (Elsevier, Amsterdam, 2017)
29. J.S. Brecher, Chimia **52**, 658 (1998). https://doi.org/10.2533/chimia.1998.658
30. W.I. Milwaukee, *Catalog Handbook of Fine Chemicals* (Aldrich Chemical Company, Burlington, 1990)
31. D. Kerr, *Fire Protection Guide to Hazardous Materials*, 10th edn. (National Fire Protection Association, Quincy, 1991)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Authors and Affiliations

Sun Yoo Hwang\textsuperscript{1} · Jeong Won Kang\textsuperscript{1,2}

\textsuperscript{1} Jeong Won Kang
\texttt{jwkang@korea.ac.kr}

\textsuperscript{1} Department of Chemical Biological Engineering, Korea University, 145 Anam-ro, Sungbuk-gu, Seoul 02841, Republic of Korea

\textsuperscript{2} Graduate School of Energy and Environment, Korea University, 145 Anam-ro, Sungbuk-gu, Seoul 02841, Republic of Korea