

Extension of Graph Convolution Network for Sigma-Profile Prediction to Silicon-Containing Compounds

Ji Eun Rhee

Korea University

Beom Chan Ryu

Korea University

Sun Yoo Hwang

Korea University

Seung Yeop Baek

SK Inc. AX

Seon Min Kim

SK Inc. AX

Ha Young Kim

SK Inc. AX

Jeong Won Kang

`jwkang@korea.ac.kr`

Korea University

Research Article

Keywords: Thermophysical property prediction, Sigma-profile, COSMO-SAC, Silicon-containing compounds, Machine learning, Vapor–liquid equilibrium

Posted Date: September 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10966846/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

1 **Title page**

2 **Title**

3 Extension of Graph Convolution Network for Sigma-Profile Prediction to Silicon-Containing
4 Compounds

5

6 **Author information**

7 Ji Eun Rhee¹, Beom Chan Ryu¹, Sun Yoo Hwang^{1,2}, Seung Yeop Baek², Seon Min Kim², Ha
8 Young Kim², Jeong Won Kang^{1*}

9 ¹ Department of Chemical and Biological Engineering, Korea University, Seoul 02841,
10 Republic of Korea

11 ² SK Inc. AX, Gyeonggi-do 13558, Republic of Korea

12 * *Corresponding author.* (jwkang@korea.ac.kr)

13

14

15

Abstract

Silicon-containing compounds are widely used in electronic materials, lubricants, coatings, sealants, refrigerants, and many other industrial applications. However, the limited availability of experimental thermodynamic data and the high computational cost of quantum-chemical calculations make accurate predictions of their thermodynamic properties challenging. In this study, an integrated machine-learning framework is proposed to predict sigma-profiles directly from molecular structures using a Group Contribution-based Graph Convolution Network (GC-GCN) and to perform thermodynamic calculations using the conductor-like screening model with segment activity coefficient (COSMO-SAC) model. To improve the framework's applicability to silicon-containing chemical systems, the GC-GCN training database was expanded to include 546 silicon-containing compounds, bringing the total to 4,102 compounds. In addition, an extended version of the COSMO-SAC model, COSMO-SAC-COOH, was adopted to account for carboxylic-acid interactions in thermodynamic calculations explicitly. The predicted sigma-profiles showed good agreement with Density Functional Theory (DFT)-based COSMO descriptors and, through the calculated activity coefficients, reproduced vapor–liquid equilibrium behavior across systems exhibiting diverse phase-equilibrium characteristics, yielding average squared error in pressure ($RRMSE_p$) and mole fraction ($RRMSE_y$) values of 4.56% and 0.02, respectively, relative to experimental data. The proposed framework enables reliable thermodynamic predictions directly without performing quantum-chemical calculations or generating molecular structure files, providing a practical and computationally efficient tool for predicting thermodynamic and phase-equilibrium properties of a broad range of compounds, including silicon-containing compounds.

Keywords

Thermophysical property prediction, Sigma-profile, COSMO-SAC, Silicon-containing compounds, Machine learning, Vapor–liquid equilibrium

Acknowledgments

This work was supported by the Technology Innovation Program (RS-2024-00431821) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea). Also, this work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NTIS 2340047109; IRIS RS-2025-25438650).

Research funding

This research did not receive funding.

Competing interests

All evaluated property values, together with their uncertainties and quality classifications, are available in the KDB platform (<https://www.mdlkdb.com>).

CRedit authorship contribution statement

Ji Eun Rhee: Software, Data curation, Investigation, Formal analysis, Visualization, Validation, Writing – original draft. **Beom Chan Ryu**: Methodology, Validation. **Sun Yoo Hwang**: Conceptualization, Methodology. **Seung Yeop Baek**: Validation, Data curation. **Seon Min Kim**: Validation, Formal analysis. **Ha Young Kim**: Validation, Formal analysis, Project administration. **Jeong Won Kang**: Supervision, Writing – review & editing, Funding acquisition.

1. Introduction

Silicon-containing compounds are of significant interest in chemical and process industries due to their broad applications in electronic materials, lubricants, coatings, sealants, refrigerants, and other functional chemical systems [1]. These applications frequently involve vapor–liquid transitions and phase-equilibrium phenomena during manufacturing, separation, purification, and high-temperature processing [2]. Despite their broad industrial applications, experimental thermodynamic data for silicon-containing compounds remain scarce [3]. Consequently, accurate prediction of thermodynamic and vapor–liquid equilibrium properties has become increasingly important.

Activity-coefficient models are widely used in chemical engineering to predict thermodynamic and phase-equilibrium properties of complex mixtures. Among the widely used activity-coefficient models, COSMO-based models such as COSMO-RS [4] and COSMO-SAC [5] predict thermodynamic properties from molecular structures using quantum-chemical COSMO calculations. Since the initial development of COSMO-SAC by Lin and Sandler [5], various extensions have been introduced to enhance prediction accuracy across diverse thermodynamic systems and phase-equilibrium conditions [6-8]. The COSMO-SAC framework relies on sigma-profiles and cavity volumes obtained from Density Functional Theory (DFT)-based COSMO calculations, which require substantial computational resources, specialized software, expertise, and considerable computation time; the COSMO calculation accounts for most of the overall computational cost [9]. These limitations become a major bottleneck for large-scale molecular screening, particularly for structurally diverse silicon-containing compounds.

To overcome these limitations, various surrogate models, such as GC-based methods [10, 11], have been proposed to predict COSMO descriptors without expensive quantum-chemical

1 calculations. More recently, advances in machine learning (ML) have led to ML-based
2 approaches that effectively predict COSMO descriptors directly from molecular structures.
3 Representative examples include high-dimensional neural networks [12], Transformer-
4 convolutional neural networks [13], and Message-passing neural network based multi-task
5 deep-learning framework [14]. Although these approaches demonstrated improved predictive
6 accuracy, they were developed using relatively limited COSMO datasets.

7 Graph-based neural-network approaches are particularly effective, as molecular structural
8 features and intermolecular interactions are naturally represented as graph data. Hwang et al.
9 [15] introduced a Group Contribution-based Graph Convolution Network (GC-GCN) model
10 for predicting pure-component properties of organic compounds. Previous COSMO-SAC-ML
11 [16] studies have shown that GC-GCN [15]-predicted sigma-profiles can reproduce phase-
12 equilibrium properties with accuracy comparable to conventional COSMO-SAC methods. The
13 model was trained on an expanded dataset of 3,556 sigma-profiles, representing a larger
14 training set than those used in earlier ML-based COSMO studies. In that framework,
15 thermodynamic calculations were performed using the COSMO-SAC 2010 [17] model.
16 However, the training data remains largely limited to conventional organic compounds, leaving
17 silicon-containing compounds underrepresented in available databases and training datasets.

18 This study extends the previously proposed COSMO-SAC-ML [16] framework in two key
19 areas. The COSMO-SAC-COOH model [8], which explicitly incorporates COOH-segment
20 interactions, was adopted for thermodynamic calculations. Additionally, the chemical space
21 was broadened to include silicon-containing compounds. A comprehensive sigma-profile
22 database comprising 4,102 compounds, including 546 silicon-containing compounds, was
23 developed to enhance the GC-GCN [15]-based prediction framework. The model's
24 performance was assessed by comparing vapor–liquid equilibrium (VLE) predictions obtained

- 1 from activity coefficients calculated using the predicted sigma-profiles with those obtained
- 2 from conventional quantum-chemical COSMO-SAC calculations.
- 3

2. Theoretical Background

2.1 Activity-coefficient model based on COSMO-SAC

Activity-coefficient (γ_i) models describe the non-ideal behavior of liquid mixtures and are widely used for predicting phase-equilibrium properties. Among them, the conductor-like screening model with segment activity coefficient (COSMO-SAC), developed by Lin and Sandler [5], predicts thermodynamic properties from molecular structures based on quantum-chemical surface charge distributions. In this study, the COSMO-SAC-COOH model [8], an extended version of COSMO-SAC that explicitly accounts for carboxylic acid interactions, was employed throughout the work. The model represents the molecular surface using four sigma-profile categories: non-hydrogen-bonding (*nhb*), hydroxyl-group hydrogen-bonding (*OH*), other hydrogen-bonding (*OT*), and carboxylic-acid hydrogen-bonding (*COOH*) segments.

The sigma-profile, $p_i(\sigma)$, is defined as the probability distribution of molecular surface segments with a screening charge density σ and is defined as Eq. 1.

$$p_i(\sigma) = \frac{A_i(\sigma)}{A_i} \quad (1)$$

where $A_i(\sigma)$ is the cavity surface area corresponding to the screening charge density σ , and A_i is the total cavity surface area of component i . σ is the average screening charge density of a surface segment and is discretized into 51 grid intervals ranging from -0.025 to $0.025 \text{ e}/\text{\AA}^2$ in practical COSMO-SAC calculations. Fig. 1 shows a representative sigma-profile of water.

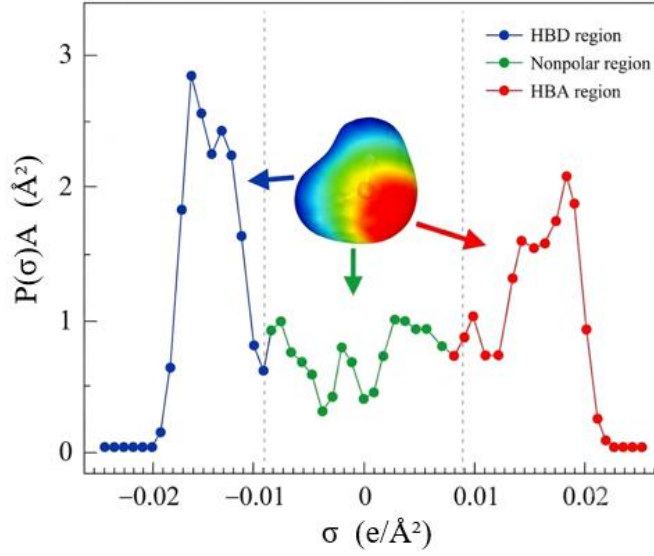


Fig. 1 Representative sigma-profile of water

The sigma-profile of component i in the COSMO-SAC-COOH formulation is expressed as Eq.2.

$$p_i(\sigma) = p_i^{nhb}(\sigma) + p_i^{OH}(\sigma) + p_i^{OT}(\sigma) + p_i^{COOH}(\sigma) \quad (2)$$

where p_i^{nhb} , p_i^{OH} , p_i^{OT} , and p_i^{COOH} denote the non-hydrogen-bonding, hydroxyl-group hydrogen-bonding, other hydrogen-bonding, and carboxylic-acid hydrogen-bonding sigma-profiles, respectively.

The activity coefficient, γ_i , is expressed as the sum of combinatorial and residual contributions.

$$\ln \gamma_i = \ln \gamma_i^{comb} + \ln \gamma_i^{res} \quad (3)$$

where the combinatorial contribution accounts for differences in molecular size and shape, while the residual contribution represents intermolecular interactions between molecular surface segments.

The combinatorial contribution is calculated using the modified Staverman–Guggenheim equation.

$$\ln \gamma_i^{comb} = \ln \left(\frac{\phi_i}{x_i} \right) + \frac{z}{2} q_i \ln \left(\frac{\theta_i/x_i}{\phi_i/x_i} \right) + l_i - \frac{\phi_i}{x_i} \sum_j x_j l_j \quad (4)$$

$$\frac{\theta_i}{x_i} = \frac{q_i}{\sum_j x_j q_j} \quad (5)$$

$$\frac{\phi_i}{x_i} = \frac{r_i}{\sum_j x_j r_j} \quad (6)$$

$$l_i = \frac{z}{2} (r_i - q_i) - (r_i - 1) \quad (7)$$

1 where r_i and q_i are the normalized cavity volume and normalized cavity surface area of
2 component i , respectively.

3 The residual activity coefficient accounts for electrostatic and hydrogen-bonding
4 interactions between molecular surface segments

$$\ln \gamma_i^{res} = \frac{A_i}{a_{eff}} \sum_s^{nhb, OH, OT, COOH} \sum_{\sigma_m} p_i^s(\sigma_m^s) [\ln \Gamma_s^s(\sigma_m^s) - \ln \Gamma_i^s(\sigma_m^s)] \quad (8)$$

5 where $\Gamma(\sigma)$ denotes the segment activity coefficient at screening charge density σ .
6 Subscripts S and i represent the mixture and pure-component states, respectively, and a_{eff}
7 is the effective segment area. The segment activity coefficient is calculated iteratively as Eq. 9.

$$\ln \Gamma_j^t(\sigma_m^t) = - \ln \left\{ \sum_s^{nhb, OH, OT, COOH} \sum_{\sigma_n} p_j^s(\sigma_n^s) \Gamma_j^s(\sigma_n^s) \exp \left[\frac{-\Delta W(\sigma_m^t, \sigma_n^s)}{RT} \right] \right\} \quad (9)$$

8 where superscripts denote the sigma-profile type (nhb , OH , OT , or $COOH$), and subscripts
9 denote the screening charge density indices. The sigma-profile of the mixture is obtained as the
10 area-weighted average of the pure-component sigma-profiles.

$$p_s(\sigma) = \frac{\sum_i x_i A_i p_i(\sigma)}{\sum_i x_i A_i} \quad (10)$$

11 The interaction between two surface segments is described by the segment exchange energy.

$$\Delta W(\sigma_n^s, \sigma_m^t) = C_{ES}(\sigma_m^t + \sigma_n^s)^2 - C_{hb}(\sigma_m^t - \sigma_n^s)^2 \quad (11)$$

1 where C_{ES} is an electrostatic constant and C_{hb} is the hydrogen-bonding interaction parameter.

$$C_{hb}(\sigma_m^t, \sigma_n^s) = \begin{cases} C_{OH-OH} & \text{if } s = t = OH \text{ and } \sigma_m^t, \sigma_n^s < 0 \\ C_{OH-OT} & \text{if } s = OH, t = OT \text{ and } \sigma_m^t, \sigma_n^s < 0 \\ C_{OH-COOH} & \text{if } s = OH, t = COOH \text{ and } \sigma_m^t, \sigma_n^s < 0 \\ C_{COOH-COOH} & \text{if } s = t = COOH \text{ and } \sigma_m^t, \sigma_n^s < 0 \\ C_{COOH-OT} & \text{if } s = COOH, t = OT \text{ and } \sigma_m^t, \sigma_n^s < 0 \\ C_{OT-OT} & \text{if } s = t = OT \text{ and } \sigma_m^t, \sigma_n^s < 0 \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

2 All model parameters used in the COSMO-SAC-COOH calculations were adopted from the
3 original COSMO-SAC-COOH model proposed by Kang et al. [8].

4 **2.2 Machine Learning Model for COSMO-SAC Sigma-Profile Prediction**

5 The machine-learning model employed in this study is based on the Group Contribution-
6 based Graph Convolution Network (GC-GCN) proposed by Hwang et al. [15]. In the GC-GCN
7 model, each molecule is represented as a graph of functional groups, where functional groups
8 are treated as graph nodes and their chemical connectivity defines the graph edges. This group-
9 level representation reduces the number of trainable parameters while preserving the structural
10 information required for molecular property prediction, making it suitable for relatively small
11 training datasets. The graph representation consists of a group feature matrix G^0 , generated by
12 one-hot encoding of functional-group identities, and a connectivity matrix C , obtained from
13 the symmetrically normalized adjacency matrix.

14 Graph convolution is subsequently performed to update the representation of each functional
15 group by incorporating information from neighboring groups. Neighboring information is first
16 aggregated according to Eq. 13.

$$H^{k+1} = C \times G^k \quad (13)$$

17 and the group representations are then updated as Eq. 14.

$$G^{k+1} = \phi(H^{k+1} \times W^{k+1} + b^{k+1}) \quad (14)$$

where W and b denote the trainable weight matrix and bias vector, respectively, and $\phi(\cdot)$ denotes the activation function, which is implemented as Leaky Rectified Linear Unit (LeakyReLU) in this study. Repeating this procedure through multiple graph convolution layers enables each functional group to capture its local chemical environment.

After graph convolution, a read-out operation performs node-wise summation to generate a fixed-length molecular feature vector M . This molecular representation is subsequently processed by fully connected neural network (FCN) layers to predict the target molecular property.

$$X = W_{out}\phi(W_3\phi(W_2\phi(W_1M + b_1) + b_2) + b_3) + b_{out} \quad (15)$$

where W_1 , W_2 , and W_3 are the trainable weight matrices of the three fully connected hidden layers; b_1 , b_2 , and b_3 are the corresponding bias vectors; W_{out} and b_{out} denote the weight matrix and bias vector of the output layer, respectively; and $\phi(\cdot)$ denotes the LeakyReLU activation function. The FCN predicts the target COSMO descriptor from the molecular feature vector. The predicted descriptors are subsequently used as inputs to the COSMO-SAC model.

2.3 Vapor–Liquid Equilibrium Calculation

Vapor–liquid equilibrium (VLE) describes the equilibrium state in which liquid and vapor-phases coexist under a given temperature and pressure. VLE is calculated by combining activity coefficients with the saturation pressures of pure components using the modified Raoult's law [18].

$$y_iP = x_i\gamma_iP_i^{sat} \quad (16)$$

where x_i and y_i denote the liquid- and vapor-phase mole fractions of component i , respectively, P is the total pressure, γ_i is the activity coefficient obtained from the COSMO-

1 SAC model, and P_i^{sat} is the saturation pressure of the pure component.

2 The saturation pressure of a pure component is commonly represented by empirical vapor-
3 pressure correlations. Among them, the Antoine equation is expressed as Eq. 17 [19].

$$\log_{10} P^{\text{sat}} = A - \frac{B}{T + C} \quad (17)$$

4 where P^{sat} is the saturation pressure, T is the temperature, and A , B , and C are substance-
5 specific Antoine parameters. Alternatively, the Wagner equation is expressed as Eq. 18 [19].

$$\ln\left(\frac{P^{\text{sat}}}{P_c}\right) = \frac{A\tau + B\tau^{1.5} + C\tau^{2.5} + D\tau^5}{T/T_c} \quad (18)$$

$$\tau = 1 - \frac{T}{T_c} \quad (19)$$

6 where P_c and T_c denote the critical pressure and critical temperature, respectively, and A , B ,
7 C , and D are substance-specific Wagner parameters.

8

3. Computational Details

3.1 Model Architecture

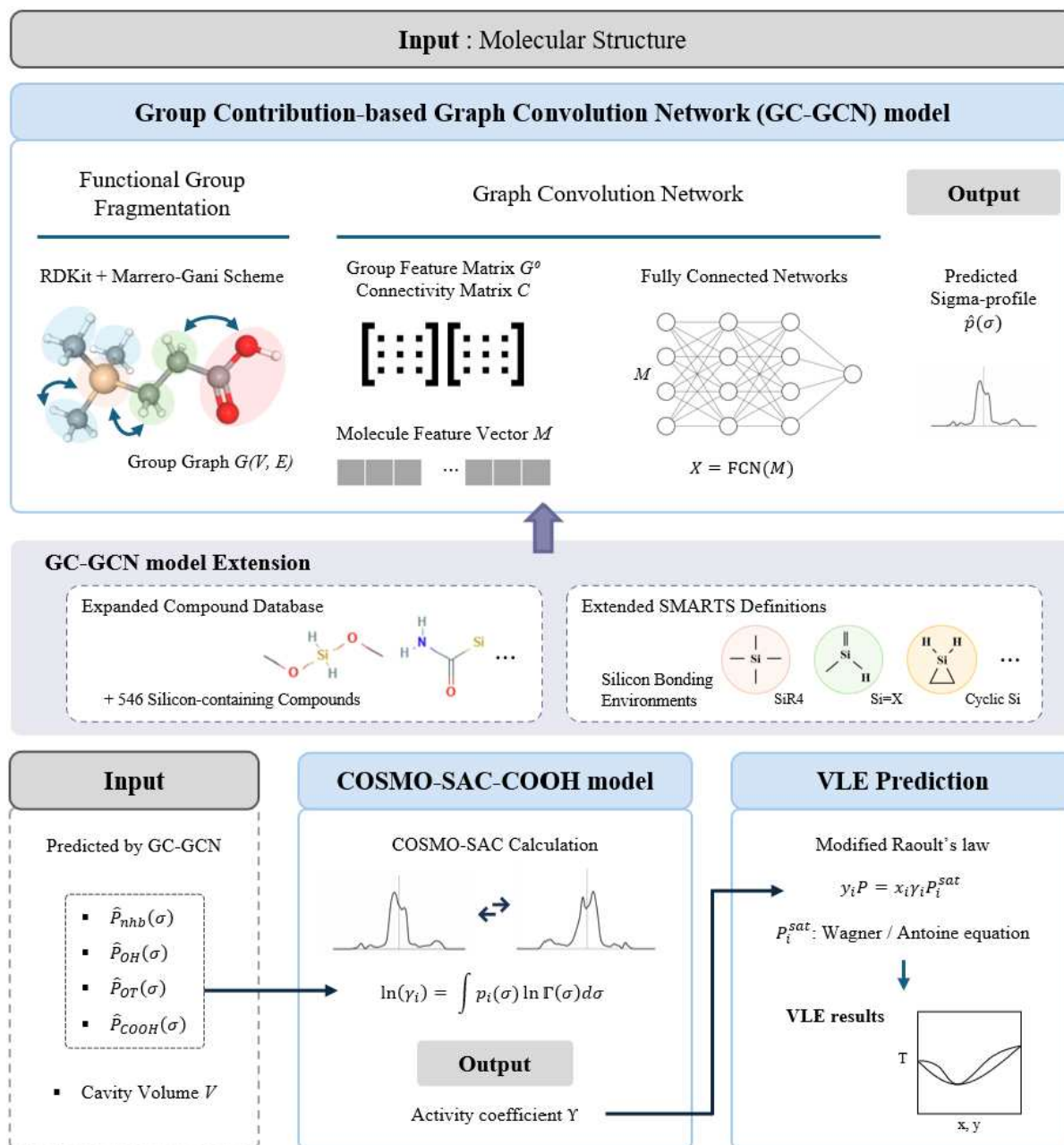


Fig. 2 Overall architecture of the proposed machine-learning-based thermodynamic prediction framework

The proposed framework is based on the previously reported COSMO-SAC-ML framework developed by Ryu et al. [16]. Fig. 2 illustrates the overall architecture of the proposed

framework. Molecular structures were represented as Simplified Molecular Input Line Entry System (SMILES) [20] strings and processed with RDKit [21] to construct molecular graphs. Functional groups identified according to the Marrero–Gani group contribution scheme [22] serve as graph nodes, with their chemical connectivity defining the graph edges; each functional group is represented as a node feature vector. To improve the representation of silicon-containing compounds, 9 additional Si-centered functional groups were introduced, increasing the total number of functional groups to 57, with the corresponding group definitions presented in Table 1. This molecular graph representation is used as the input to the GC-GCN model described in Section 2.2.

Table 1 List of additional silicon-containing functional groups

ID	Group	SMARTS	Index
49	–SiH3	[!R;SiX4H3]	
50	>SiH2	[!R;SiX4H2]	
51	>SiH–	[!R;SiX4H1]	
52	>Si<	[!R;SiX4H0]	
53	=SiH2	[!R;SiX3H2;\$(!R;SiX3H2)(=*)]	
54	=SiH–	[!R;SiX3H1;\$(!R;SiX3H1)(=*)]	
55	=Si<	[!R;SiX3H0;\$(!R;SiX3H0)(=*)]	
56	>SiH2	[R;SiX4H2]	RING
57	>SiH–	[R;SiX4H1]	RING

The generated graph representation is processed by the GC-GCN model described in Section 2.2. Through successive graph convolution layers, each functional-group representation is iteratively updated by aggregating information from neighboring groups, enabling the model

to capture both the chemical identity of individual functional groups and their local structural environments. A summation read-out operation is then performed to generate a fixed-length molecular feature vector, which is subsequently passed through fully connected neural network (FCN) layers to predict the target molecular descriptors.

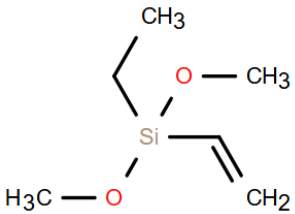
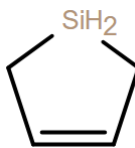
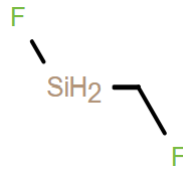
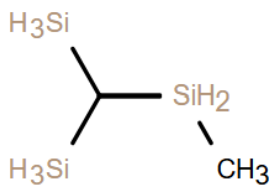
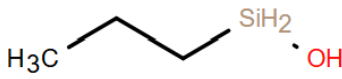
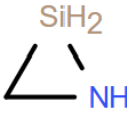
The trained framework consists of five independent GC-GCN models that predict cavity volume V and four sigma-profile components ($p^{nhb}(\sigma)$, $p^{OH}(\sigma)$, $p^{OT}(\sigma)$, and $p^{COOH}(\sigma)$). Each GC-GCN model consists of two graph convolution layers, followed by a summation-based readout layer and three fully connected hidden layers with a hidden dimension of 140. LeakyReLU activation functions and He-normal initialization were employed throughout the hidden layers. Each model independently predicts a single target property directly from the molecular structure, and the predicted outputs are subsequently combined to reconstruct the complete set of COSMO descriptors for the molecule. These predicted COSMO descriptors are then supplied to the COSMO-SAC-COOH model, where the molecular activity coefficients are calculated according to the thermodynamic formulation described in Section 2.1. For vapor–liquid equilibrium calculations, the calculated activity coefficients are combined with pure-component saturation pressures obtained using either the Antoine or Wagner equation, as described in Section 2.3, depending on the availability of vapor-pressure parameters. The resulting framework enables isothermal and isobaric VLE calculations without performing DFT-based COSMO calculations for the target molecules.

3.2 Dataset Construction

The training dataset for the GC-GCN model was assembled from both publicly available databases and internally generated datasets. Open-source data for 2,387 compounds were collected from the VT-2005 [9] and UD databases [7]. An additional 1,169 compounds generated in our previous study [16] were incorporated into the training dataset.

1 In this study, COSMO descriptors for 546 silicon-containing compounds were newly
2 generated using the DFT-based COSMO calculation procedure proposed by Mullins et al. [9].
3 Representative examples of these newly generated compounds are listed in Table 2. The final
4 dataset comprised 4,102 compounds and was used to train the GC-GCN models for predicting
5 the cavity volume V and the four sigma-profile components: the non-hydrogen-bonding sigma-
6 profile ($p^{nhb}(\sigma)$), the hydroxyl-group hydrogen-bonding sigma-profile ($p^{OH}(\sigma)$), the other
7 hydrogen-bonding sigma-profile ($p^{OT}(\sigma)$), and the carboxylic-acid hydrogen-bonding sigma-
8 profile ($p^{COOH}(\sigma)$).

Table 2 Examples of newly generated silicon-containing compounds

IUPAC name	Structure	SMILES
ethenyl-ethyl-dimethoxysilane		<chem>CC[Si](C=C)(OC)OC</chem>
2,5-dihydro-1 <i>H</i> -silole		<chem>C1C=CC[SiH2]1</chem>
fluoro(fluoromethyl)silane		<chem>C(F)[SiH2]F</chem>
disilylmethyl(methyl)silane		<chem>C[SiH2]C([SiH3])[SiH3]</chem>
hydroxy(propyl)silane		<chem>CCC[SiH2]O</chem>
azasiliridine		<chem>C1N[SiH2]1</chem>

3.3 Model Training and Evaluation

The GC-GCN models were implemented using RDKit, Keras, and NumPy. Separate models were trained to predict the cavity volume and each sigma-profile component required for the COSMO-SAC-COOH model. Model parameters were optimized using the Adam [23] optimizer with a learning rate of 0.001, and the mean squared error (MSE) was employed as

the loss function.

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (20)$$

where N is the number of samples used to calculate the loss, y_i and $\hat{y}_i$ denote the reference and predicted values, respectively.

To evaluate the generalization performance of the proposed models, the complete dataset was randomly divided into a training set comprising 90% of the compounds and an independent test set comprising the remaining 10%. The same data partition was used for all target properties. A five-fold cross-validation procedure was performed exclusively on the training set to determine the training epoch for each target property. In each fold, the model was trained for up to 10,000 epochs using full-batch optimization, with early stopping (patience = 50) based on the validation loss. The model parameters corresponding to the minimum validation loss were restored, and the corresponding convergence epoch was recorded. The average convergence epoch across the five folds was adopted as the final training epoch. Using the selected epoch, the model was retrained on the complete training set without a validation split and subsequently evaluated on the independent test set.

The predictive performance of the trained models was evaluated using the Root Mean Square Error (RMSE) for the sigma-profile components and the Relative Root Mean Square Error (RRMSE) for the cavity volume. Since the cavity volume strongly depends on molecular size, the RRMSE was adopted to normalize the prediction error relative to the reference cavity volume. The RMSE and RRMSE are given by Eq. 21 and 22, respectively.

$$\text{RMSE} (\text{\AA}^2) = \frac{1}{N} \sum_{k=1}^N \frac{1}{n} \sum_{i=1}^n \sqrt{[(p(\sigma)A_i)^{GCGN} - (p(\sigma)A_i)^{COSMO}]^2} \quad (21)$$

$$RRMSE (\%) = 100 \times \frac{1}{N} \times \sum_{k=1}^N \sqrt{\left(\frac{V^{GCGCN} - V^{COSMO}}{V^{COSMO}} \right)^2} \quad (22)$$

where N is the number of molecules in the test set, n is the number of screening charge density grids (51 in this study), $p(\sigma)$ is the sigma profile, and A_i is the cavity surface area obtained from the COSMO calculation. The superscripts $GCGCN$ and $COSMO$ denote the values predicted by the proposed model and those obtained from the original COSMO calculations, respectively.

The accuracy of the VLE calculations was quantitatively assessed using the experimental data for the selected binary systems summarized in Table 3. For each experimental data point, the calculated bubble-point property (pressure or temperature) and vapor-phase composition were compared with the corresponding experimental values. The overall deviations were evaluated in terms of the Relative Root Mean Square Error of pressure ($RRMSE_p$) and the Root Mean Square Error of vapor-phase composition ($RMSE_y$), as defined in Eq. 23 and 24.

$$RRMSE_p (\%) = 100 \times \frac{1}{N} \times \sum_{k=1}^N \sqrt{\left(\frac{p^{cal} - p^{exp}}{p^{exp}} \right)^2} \quad (23)$$

$$RMSE_y = \frac{1}{N} \times \sum_{k=1}^N \sqrt{(y^{cal} - y^{exp})^2} \quad (24)$$

where N denotes the total number of experimental VLE data points, and the superscripts cal and exp represent the calculated and experimental values, respectively.

1 **Table 3** Experimental binary VLE systems used for model evaluation

Dataset	Compounds 1	Compounds 2	References
1	Silicon tetrachloride	Chlorotrimethylsilane	[24]
2	Ethanol	Tetraethoxysilane	[25]
3	Acetone	Tetraethylsilane	[26]
4	Silicon tetrachloride	Acetonitrile	[27]
5	Carbon tetrachloride	Octamethylcyclotetrasiloxane	[28]
6	Methyl acetate	Ethyl L-(+)-lactate	[29]
7	1-Phenylethanol	Acetophenone	[30]
8	2-Ethylhexanal	n-Butanol	[31]
9	Dipropyl ketone	Di-n-butyl ether	[32]
10	Ethyl acetate	Hexamethyldisiloxane	[33]
11	n-Heptane	Furfural	[34]
12	Ethyl acetate	3-Methyl-1-butanol	[35]
13	Isopentyl acetate	Ethanol	[36]
14	Cyclohexyl acetate	Acetic acid	[37]
15	Ethyl acetate	Furfural	[38]

2

3

4 Results and Discussion

4.1 Prediction Accuracy

The predictive performance of the proposed GC-GCN models on the independent test set is summarized in Table 4. The cavity volume was evaluated using the RRMSE, while the sigma-profile components were evaluated using the RMSE. Overall, the proposed models achieved low prediction errors for both the cavity volume and all sigma-profile components, demonstrating that the graph-based molecular representation effectively captures the structural information required to predict COSMO-SAC descriptors directly from molecular structures.

Compared with the GC-GCN model used in the previous COSMO-SAC-ML [16] framework, the proposed framework maintained a comparable level of predictive accuracy despite the substantial expansion of the training dataset and prediction targets. The number of compounds in the training dataset increased from 3,556 to 4,102 through the incorporation of 546 additional silicon-containing compounds. Furthermore, unlike the previous model, which predicted only the *nhb*, *OH*, and *OT* sigma-profile components, the present model additionally predicts the *COOH* sigma-profile required for the COSMO-SAC-COOH model.

Although the prediction errors of the *nhb* and *OH* sigma-profile components increased slightly, the proposed model was trained on a substantially larger and more chemically diverse dataset that additionally includes silicon-containing compounds and predicts an additional *COOH* sigma-profile component. These extensions increase the complexity of the prediction task and may have contributed to the modest increase in the prediction errors of the *nhb* and *OH* components. In contrast, the *OT* sigma-profile was predicted with improved accuracy, and the cavity-volume prediction exhibited a comparable RRMSE (0.9032%) to that of the previous model (0.896%). Overall, despite the increased complexity of the prediction task, the proposed

framework maintains predictive performance comparable to that of the original GC-GCN model while extending both the chemical space and descriptor coverage.

Table 4 Comparison of the proposed GC-GCN model and the previous GC-GCN model

	This work	Previous GC-GCN [16]
Training Dataset		
Number of compounds	4,102	3,556
Silicon-containing compounds	546	0
Prediction performance (RMSE or RRMSE)		
<i>nhb</i> sigma-profile ($\text{\AA}^2$)	1.14	1.01
<i>OH</i> sigma-profile ($\text{\AA}^2$)	0.09	0.08
<i>OT</i> sigma-profile ($\text{\AA}^2$)	0.17	0.18
<i>COOH</i> sigma-profile ($\text{\AA}^2$)	0.12	—
Cavity volume (%)	0.90	0.90

4.2 Validation of Sigma-Profile Prediction

To further validate the proposed model, the predicted σ -profiles were compared with the corresponding reference profiles obtained from quantum-chemical COSMO calculations for several example compounds. Because the σ -profile serves as the fundamental molecular descriptor in the COSMO-SAC framework, accurately reproducing its distribution is essential for reliable thermodynamic property prediction.

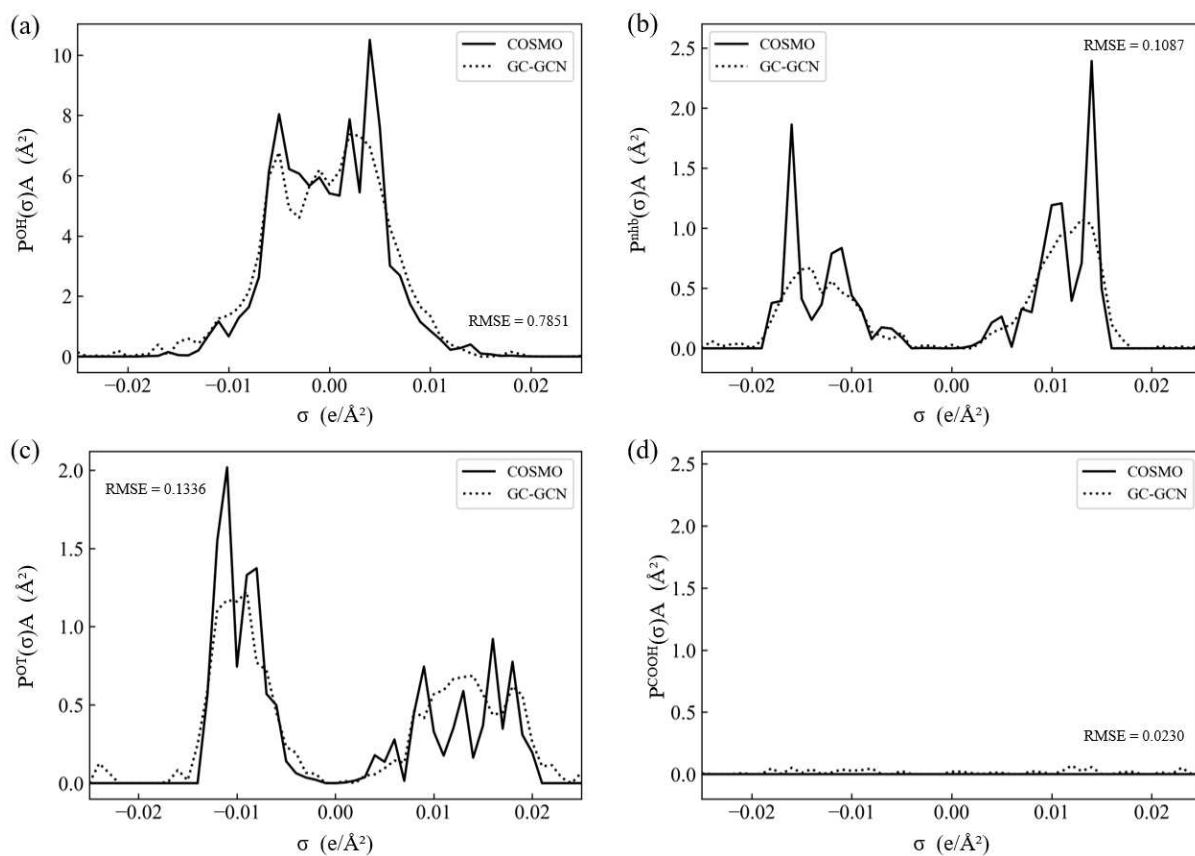


Fig. 3 Comparison of COSMO-calculated and GC-GCN-predicted sigma-profiles for [amino(hydroxy)silyl]ethene: (a) *nhb*, (b) *OH*, (c) *OT*, and (d) *COOH*

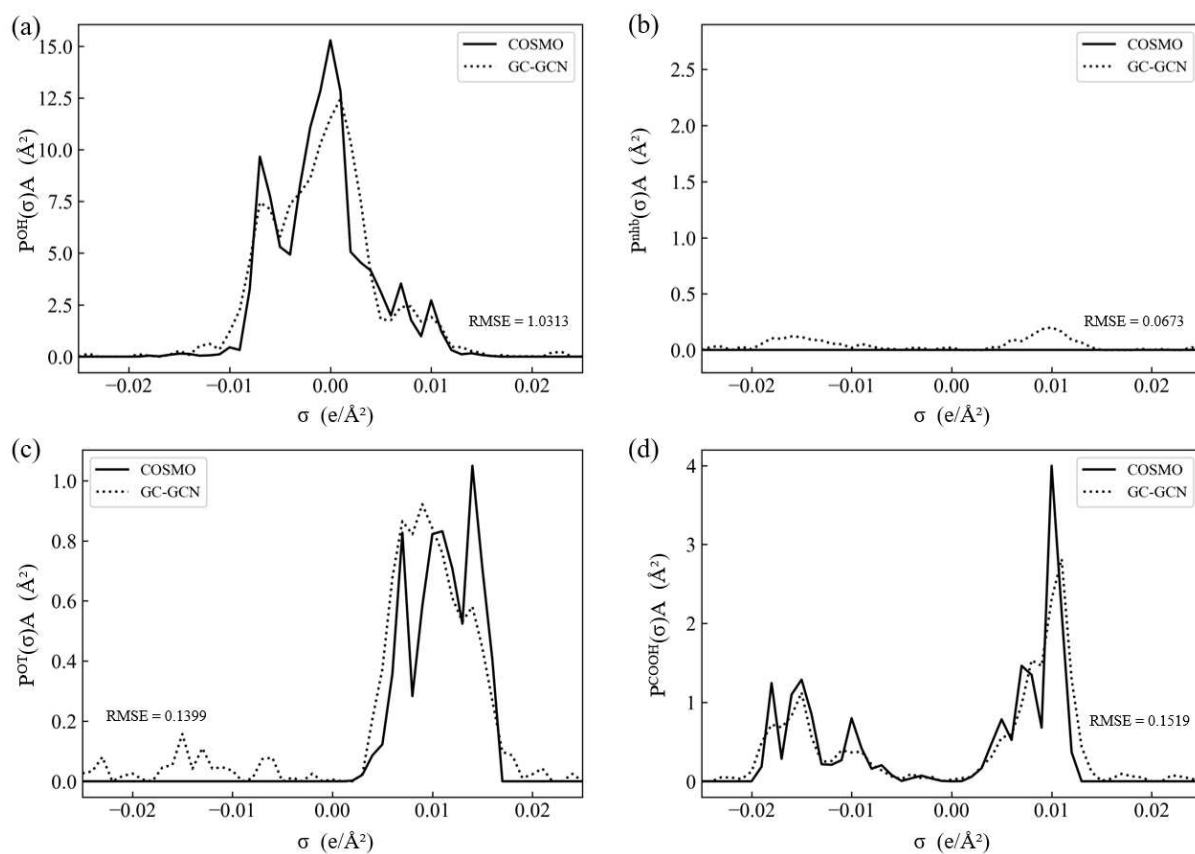


Fig. 4 Comparison of COSMO-calculated and GC-GCN-predicted sigma-profiles for pyridine-3-carboxylic acid: (a) *nhb*, (b) *OH*, (c) *OT*, and (d) *COOH*

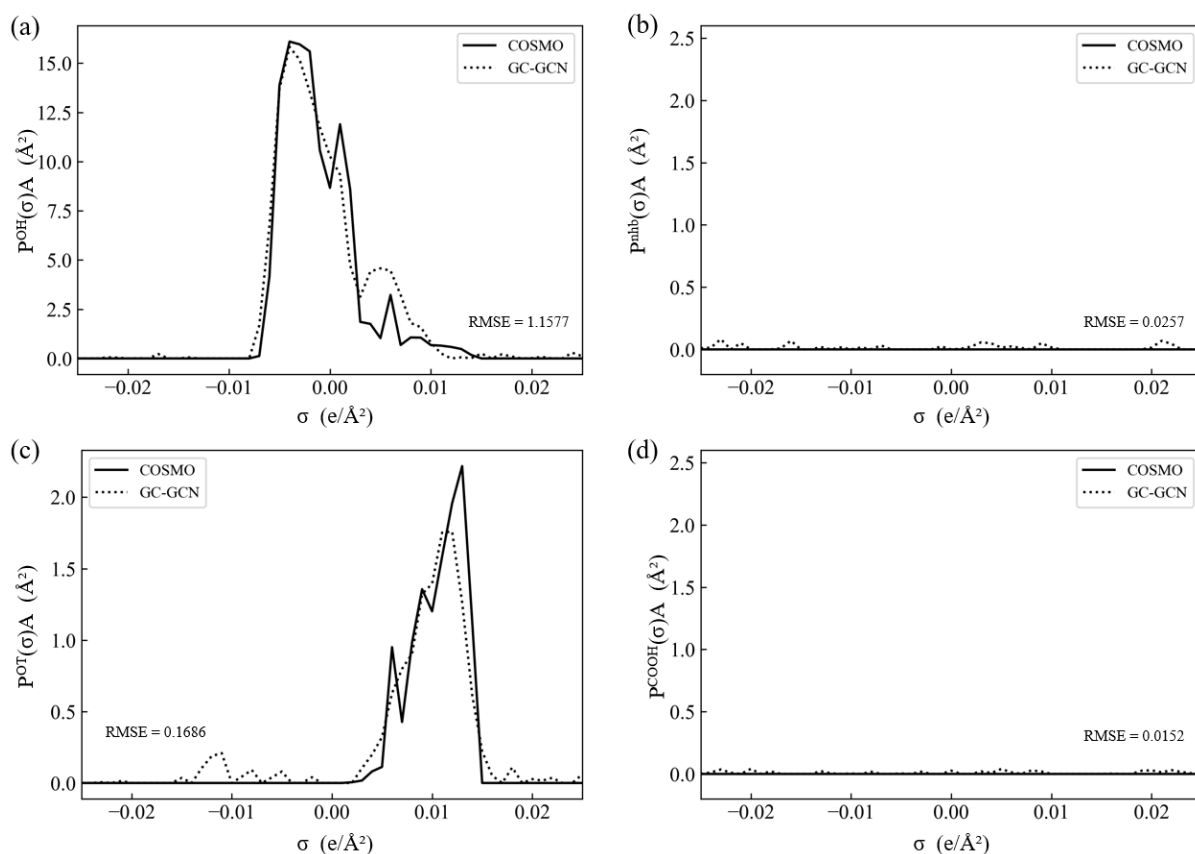


Fig. 5 Comparison of COSMO-calculated and GC-GCN-predicted sigma-profiles for dimethoxysilane: (a) n_{hb} , (b) OH , (c) OT , and (d) $COOH$

Figs. 3–5 present comparisons between the predicted and reference σ -profiles for each compound in the n_{hb} , OH , OT , and $COOH$ interaction channels. Overall, the proposed model successfully reproduced the major features of the reference σ -profiles, including the overall distribution patterns and the locations of the dominant peaks. Figs. 3 and 5 present representative silicon-containing compounds, whereas Fig. 4 shows a representative carboxylic-acid-containing compound to demonstrate the predictive performance of the proposed framework for the newly incorporated silicon-containing compounds and the newly introduced $COOH$ sigma-profile. Although slight discrepancies were observed in peak intensity and local profile shape, particularly in regions with sharp peaks, the overall agreement remained good across all interaction channels. In particular, the $COOH$ sigma-profile, which is

1 newly introduced in this study, was also predicted consistently despite its relatively sparse
2 distribution. These results indicate that the proposed GC-GCN model can reproduce the
3 characteristic charge-density distributions required by the COSMO-SAC-COOH model,
4 suggesting that the predicted σ -profiles are suitable for subsequent thermodynamic property
5 calculations.

6

4.3 Vapor-Liquid Equilibrium Prediction

To evaluate the applicability of the predicted sigma-profiles in thermodynamic calculations, vapor–liquid equilibrium (VLE) calculations were performed using the GC-GCN-predicted sigma-profiles within the COSMO-SAC-COOH framework. Fig. 6 shows selected VLE results from the evaluated binary systems. The selected systems exhibit diverse VLE behaviors, allowing the performance of the proposed framework to be examined under different phase-equilibrium conditions. The experimental VLE data used for comparison were obtained from the literature [24, 25, 27]. Given the limited availability of experimental VLE data for additional silicon-containing compounds suitable for independent validation, the silicon-containing systems were evaluated using validation-fold predictions obtained from cross-validation: for each such compound, VLE was calculated from the sigma-profile predicted by the fold model in which that compound had been held out, i.e., an out-of-sample prediction with respect to that fold. UNIFAC [39] predictions were not attempted for the silicon-containing systems in this study, as the group-interaction parameters covering the relevant Si-containing groups and their interactions with other functional groups have not yet been sufficiently established.

Across the 15 binary systems listed in Table 3, VLE calculations using the original COSMO-generated sigma-profiles yielded average $RRMSE_P$ and $RMSE_y$ values of 3.99% and 0.02, respectively, whereas those using the ML-predicted sigma-profiles yielded corresponding values of 4.56% and 0.02.

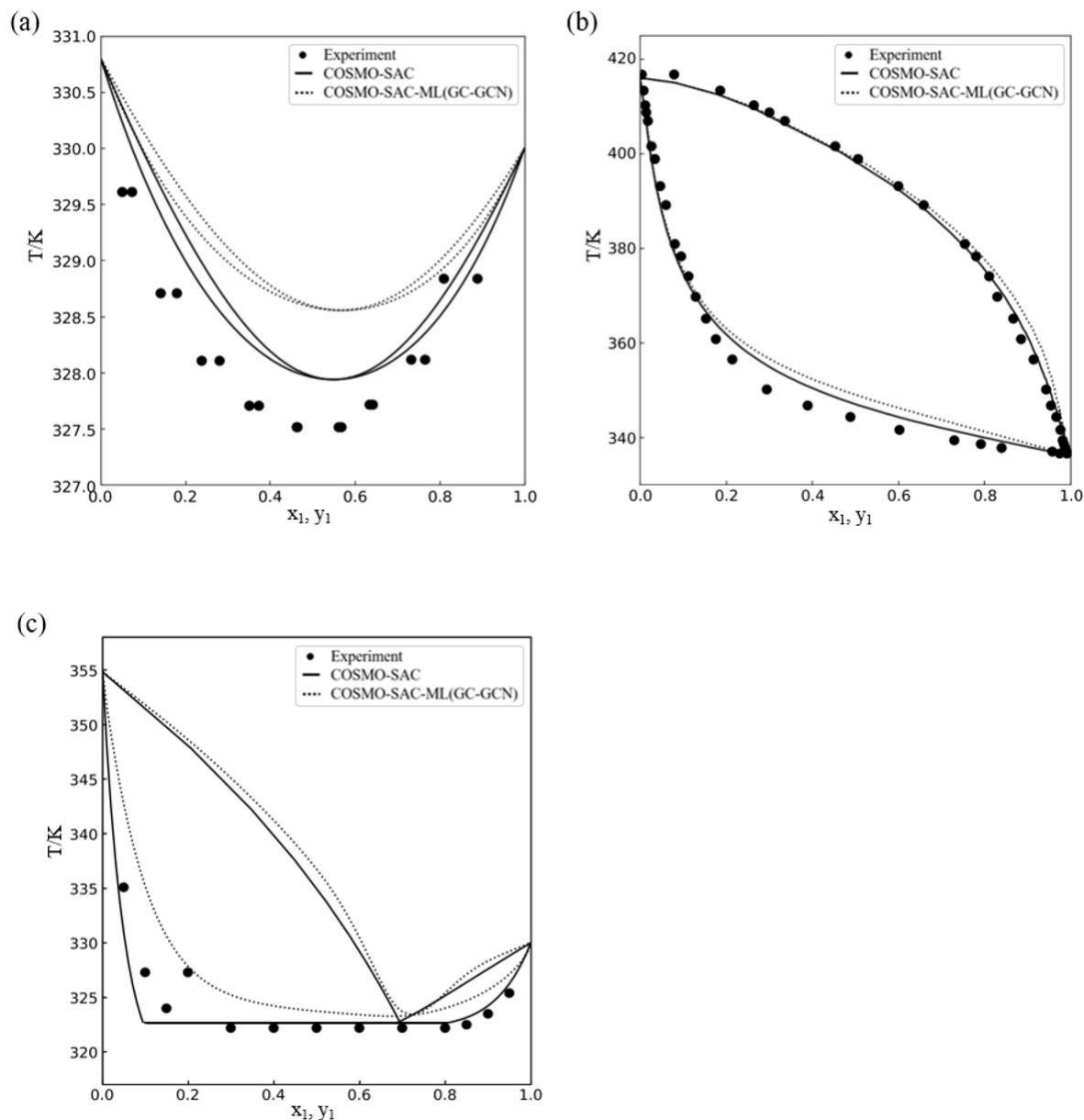


Fig. 6 Comparison of experimental data and model predictions for binary VLE of (a) silicon tetrachloride (1) + chlorotrimethylsilane (2) at 101.325 kPa, (b) ethanol (1) + tetraethoxysilane (2) at 53.32 kPa, and (c) silicon tetrachloride (1) + acetonitrile (2) at 101.325 kPa

Across the evaluated systems, the proposed framework successfully reproduced the overall shapes of the bubble- and dew-point curves across the composition range. The results obtained using the predicted sigma-profiles were in close agreement with those calculated using the original COSMO-generated sigma-profiles and reproduced the overall trends observed in the

experimental data. In particular, the non-ideal system demonstrates that the predicted sigma-profiles preserve the molecular interaction characteristics required to describe pronounced deviations from ideal-solution behavior. Although minor deviations were observed in some composition regions, the overall phase-equilibrium behavior was successfully reproduced. These results indicate that the proposed framework can reproduce overall VLE behavior without requiring computationally expensive quantum-chemical COSMO calculations.

5. Conclusion

This study presents an integrated framework that predicts sigma-profiles directly from a molecular graph representation using a GC-GCN model and applies the predicted descriptors within the COSMO-SAC-COOH model for thermodynamic calculations. Compared with the previous framework, the model expands the chemical space by incorporating 546 silicon-containing compounds into the GC-GCN training database, thereby extending its applicability to silicon-containing chemical systems. Furthermore, adopting the COSMO-SAC-COOH model enables prediction of sigma-profile descriptors that explicitly account for carboxylic-acid interactions.

The predicted sigma-profiles showed good agreement with the DFT-based COSMO descriptors, achieving RMSE values of 1.14, 0.09, 0.17, and 0.12 Å² for the *nhb*, *OH*, *OT*, and *COOH* components, respectively, and an RRMSE of 0.90% for the cavity volume. The framework successfully reproduced diverse vapor–liquid equilibrium behaviors through activity coefficients calculated from the predicted sigma-profiles, yielding average $RRMSE_p$ and $RMSE_y$ values of 4.56% and 0.02, respectively. These results suggest that reliable thermodynamic predictions can be achieved directly from ML-predicted sigma-profiles, enabling rapid evaluation of previously unexamined compounds from their

1 molecular structures without quantum-chemical calculations or additional molecular
2 structure files.

3 Overall, the proposed framework provides a practical and efficient screening tool for
4 predicting the thermodynamic behavior of a broader range of molecular systems, including
5 silicon-containing compounds. Furthermore, it enables large-scale molecular screening and
6 supports the design and evaluation of novel compounds across diverse industrial sectors
7 where accurate thermodynamic and phase-equilibrium predictions are essential.

8

References

1. M. Andriot, S.H. Chao, A. Colas, S. Cray, F. de Buyl, J.V. DeGroot, Jr., A. Dupont, T. Easton, J.L. Garaud, E. Gerlach, F. Gubbels, M. Jungk, S. Leadley, J.P. Lecomte, B. Lenoble, R. Meeks, A. Mountney, G. Shearer, S. Stassen, C. Stevens, X. Thomas, A.T. Wolf, In *Inorganic Polymers*, ed. by R. De Jaeger, M. Gleria (Nova Science Publishers, Hauppauge, NY, 2007), p. 61–161
2. J.M. Prausnitz, R.N. Lichtenthaler, E.G. De Azevedo, *Molecular thermodynamics of fluid-phase equilibria*. 3rd edn. (Prentice Hall PTR, Upper Saddle River, 1999)
3. A. Burcat, E. Goos, *Int. J. Chem. Kinet.* **50** (9), 633 (2018) <https://doi.org/10.1002/kin.21188>
4. A. Klamt, *J. Phys. Chem.* **99** (7), 2224 (1995) <https://doi.org/10.1021/j100007a062>
5. S.-T. Lin, S.I. Sandler, *Ind. Eng. Chem. Res.* **41** (5), 899 (2002) <https://doi.org/10.1021/ie001047w>
6. R.d.P. Soares, L.F. Baladão, P.B. Staudt, *Fluid Ph. Equilib.* **488**, 13 (2019) <https://doi.org/10.1016/j.fluid.2019.01.015>
7. I.H. Bell, E. Mickoleit, C.-M. Hsieh, S.-T. Lin, J. Vrabec, C. Breitkopf, A. Jäger, *J. Chem. Theory Comput.* **16** (4), 2635 (2020) <https://doi.org/10.1021/acs.jctc.9b01016>
8. S.S. Kang, J. Lee, J.W. Kang, *Fluid Ph. Equilib.* **521**, 112673 (2020) <https://doi.org/10.1016/j.fluid.2020.112673>
9. E. Mullins, R. Oldland, Y. Liu, S. Wang, S.I. Sandler, C.-C. Chen, M. Zwolak, K.C. Seavey, *Ind. Eng. Chem. Res.* **45** (12), 4389 (2006) <https://doi.org/10.1021/ie060370h>
10. T. Mu, J. Rarey, J. Gmehling, *AIChE J.* **53** (12), 3231 (2007) <https://doi.org/10.1002/aic.11338>
11. T. Mu, J. Rarey, J. Gmehling, *AIChE J.* **55** (12), 3298 (2009) <https://doi.org/10.1002/aic.11933>
12. Q. Liu, L. Zhang, K. Tang, L. Liu, J. Du, Q. Meng, R. Gani, *AIChE J.* **67** (2), e17110 (2021) <https://doi.org/10.1002/aic.17110>
13. G. Chen, Z. Song, Z. Qi, *Chem. Eng. Sci.* **246**, 117002 (2021) <https://doi.org/10.1016/j.ces.2021.117002>
14. J. Zhang, Q. Wang, W. Shen, *Chem. Eng. Sci.* **254**, 117624 (2022) <https://doi.org/10.1016/j.ces.2022.117624>
15. S.Y. Hwang, J.W. Kang, *Int. J. Thermophys.* **43** (9), 136 (2022) <https://doi.org/10.1007/s10765-022-03060-7>
16. B.C. Ryu, S.Y. Hwang, S.S. Kang, J.W. Kang, D. Yang, *Int. J. Thermophys.* **44** (4), 49 (2023) <https://doi.org/10.1007/s10765-022-03141-7>
17. C.-M. Hsieh, S.I. Sandler, S.-T. Lin, *Fluid Ph. Equilib.* **297** (1), 90 (2010) <https://doi.org/10.1016/j.fluid.2010.06.011>
18. J.M. Smith, H.C. Van Ness, M.M. Abbott, M.T. Swihart, *Introduction to chemical engineering thermodynamics*. 9th edn. (McGraw-Hill Education, New York, 2022)
19. R.C. Reid, J.M. Prausnitz, B.E. Poling, *The properties of gases and liquids*. 4th edn. (McGraw-Hill Book Company, New York, 1987)
20. D. Weininger, *J. Chem. Inf. Comput. Sci.* **28** (1), 31 (1988) <https://doi.org/10.1021/ci00057a005>
21. G. Landrum et al., rdkit/rdkit: 2025_09_2 (Q3 2025) Release, version Release_2025_09_2 (Zenodo, 2025) <https://doi.org/10.5281/zenodo.17495409>
22. J. Marrero, R. Gani, *Fluid Ph. Equilib.* **183-184**, 183 (2001) [https://doi.org/10.1016/S0378-3812\(01\)00431-9](https://doi.org/10.1016/S0378-3812(01)00431-9)
23. D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015 (2015)
24. T. Rugina, L. Sacarescu, M. Gaspar, *Rev. Chim. (Bucharest)* **39** (10), 869 (1988)
25. J.-f. Huang, Y.-f. Lee, L.-s. Lee, *J. Chem. Eng. Data* **49** (5), 1175 (2004) <https://doi.org/10.1021/je0302248>
26. N. Bjerrum, E. Jozefowicz, Z. Phys. Chem. Abt. A **159**, 194 (1932) <https://doi.org/10.1515/zpch-1932-15918>
27. T.P. Korablina, M.V. Barinova, A.I. Kurakina, *Plast. Massy* (12), 18 (1963)
28. K. Marsh, *Trans. Faraday Soc.* **64**, 883 (1968) <https://doi.org/10.1039/TF9686400883>
29. J.M. Resa, J.M. Goenaga, A.I. Sánchez-Ruiz, M. Iglesias, *J. Chem. Eng. Data* **51** (4), 1294

- 1 (2006) <https://doi.org/10.1021/je060052p>
2 30. A.F. Frolov, I.V. Vinogradova, M.A. Loginova, Uch. Zap. Yaroslav. Tekhnol. Inst. (28), 110
3 (1972)
4 31. Y. Zhang, J. Fu, J. Zhang, J. Chem. Ind. Eng. (China) **43** (1), 98 (1992)
5 32. H.S. Wu, S.I. Sandler, J. Chem. Eng. Data **33** (3), 316 (1988)
6 <https://doi.org/10.1021/je00053a027>
7 33. W. Zhang, N. Meng, R. Sun, C. Li, J. Chem. Eng. Data **56** (12), 5078 (2011)
8 <https://doi.org/10.1021/je2008445>
9 34. F. Garner, R. Hall, J. Inst. Pet. **41**, 1 (1955)
10 35. M.T. Sanz, J. Gmehling, Fluid Ph. Equilib. **267** (2), 158 (2008)
11 <https://doi.org/10.1016/j.fluid.2008.03.002>
12 36. S.E. Kharin, V.M. Pereygin, A.G. Volkov, Izv. Vyssh. Uchebn. Zaved., Pishch. Tekhnol. (6),
13 119 (1971)
14 37. D.F. Othmer, Ind. Eng. Chem. **35** (5), 614 (1943) <https://doi.org/10.1021/ie50401a018>
15 38. M.G. Myles, R. Wingard, Ind. Eng. Chem. **53** (3), 219 (1961)
16 <https://doi.org/10.1021/ie50615a029>
17 39. A. Fredenslund, J. Gmehling, P. Rasmussen, *Vapor-liquid equilibria using UNIFAC: a group-*
18 *contribution method*. (Elsevier, Amsterdam, 1977)