Systems & Control Transactions

Research Article - Peer Reviewed Conference Proceeding
ESCAPE 35 - European Symposium on Computer Aided Process Engineering
Ghent, Belgium. 6-9 July 2025
Jan F.M. Van Impe, Grégoire Léonard, Satyajeet S. Bhonsale,
Monika E. Polańska, Filip Logist (Eds.)

PSE PRESS

A Component Property Modeling Framework Utilizing Molecular Similarity for Accurate Predictions and Uncertainty Quantification

Youquan Xu[a], Zhijiang Shao[a], Anjan K. Tula[a,*]

[a] State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China
* Corresponding Author: anjantula@zju.edu.cn.

ABSTRACT

A key step in developing high-performance industrial products lies in the design of their constituent molecules. Computer-aided molecular design (CAMD) has garnered significant attention for its potential to accelerate and improve the design process. The mainstream method involves using property prediction models to predict the properties of potential molecules and selecting the best candidates based on these predictions. However, prediction errors are inevitable, introducing unreliability into the design. To address this issue, this paper proposes a novel component property modeling framework based on a molecular similarity coefficient. By calculating the similarity between a target molecule and those in an existing database, the framework selects the most similar molecules to form a tailored training dataset. The similarity coefficient also quantifies the reliability of the property predictions. In tests across various properties, this framework not only provides a quantifiable evaluation of prediction reliability but also improves the prediction accuracy of molecules with high reliability, which has the potential to enhance the integrity of molecular design.

Keywords: Property prediction, Molecular design, Similarity coefficient

1. INTRODUCTION

Developing chemical products used as working media in industrial settings is a crucial aspect of the industrial process. One of the most important elements in this development is the design of molecular structures for these products. Traditional design methods, which rely on expert experience, are often time-consuming, labor-intensive, and prone to overlooking potentially excellent molecules. As a result, computer-aided molecular design (CAMD) is gaining increasing attention. This approach utilizes molecular property data in conjunction with efficient optimization algorithms to identify product molecules that exhibit optimal performance.

When designing a new product molecule, it is essential to predict the properties of unknown molecules based on existing molecular property data. This process is referred to as the property prediction model. The property prediction model primarily utilizes molecular structure information along with known data about molecular structures and their properties. It employs data fitting methods to establish the mapping relationship between molecular structures and their properties. Among the most classical molecular property prediction models are the Group Contribution (GC) method[1], Quantitative Structure-Property Relationship (QSPR) modeling[2], and ab initio quantum-mechanics-based methods[3]. With advancements in artificial intelligence, many new property prediction models and variations of classical models have been developed[4]. For instance, Zhang et al.[5] developed an accurate and interpretable QSPR molecular property prediction model based on deep neural networks. Alshehri et al.[6] introduced a new generation of molecular property prediction models based on Gaussian processes, which achieved higher accuracy compared to simpler GC-based models.

However, these models are not strictly mechanistic relationships between molecular structure and property,

so prediction errors are inevitable. In the process of molecular design using the property prediction model, the basis of molecular selection is the predicted value of molecular properties. Therefore, the prediction error of the property prediction model will directly affect the process of molecular screening, and easily lead to a serious mismatch between the properties of the designed molecules and the values of the experimentally verified properties. In other words, the reliability of the property prediction largely determines the reliability of the molecular design. At present, there is no quantitative method for the reliability of property prediction.

To solve this problem, this paper innovatively proposes a property prediction modeling framework based on molecular similarity. Firstly, this framework customizes a training set for the target molecules according to molecular similarity. The properties of target molecules are predicted using this training set, and the similarities between the molecules in the training set and the target molecules are used to provide a quantitative index of prediction reliability for target molecules.

2. MOLECULAR SIMILARITY COEFFICIENT

If the reliability of predicted property values can be assessed prior to experimental validation, the only information that can be utilized is that derived from the training molecules used in modeling these properties. Specifically, the adequacy of the information that training molecules provide for the target molecules is crucial. The greater the similarity between two molecules, the stronger the correlation in their properties. This similarity allows us to leverage the relationship between training and target molecules to gauge how adequately the training molecules can inform predictions about the target molecules. Consequently, it becomes possible to quantitatively measure the reliability of property predictions for the target molecules. Therefore, this paper introduces a novel method for calculating the molecular similarity coefficient.

Here, the group contribution method is used to describe the structure of a molecule. For molecules M_1 and M_2, their structures can be described as follows,

$$M_1: [n_{1\,1}, n_{1\,2}, n_{1\,3}, n_{1\,4}, \cdots, n_{1\,i}, \cdots, n_{1\,424}]$$

$$M_2: [n_{2\,1}, n_{2\,2}, n_{2\,3}, n_{2\,4}, \cdots, n_{2\,i}, \cdots, n_{2\,424}]$$

The 424 dimensions represent 424 groups which are used to form molecules [7,8]. $n_{1\,i}$ and $n_{2\,i}$ represent the number of the i^{th} group inside the molecules. If a molecule doesn't have this group, then the value is just 0.

First, a classical mathematical method is introduced, that is, Jaccard similarity coefficient. It is an efficient tool to calculate the similarity between two sets. Its overall logic is to calculate the ratio of the number of elements in the intersection and union of two sets, and the general formula is as follows:

$$JSC(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Jaccard similarity coefficient ranges in [0,1]. Specially, when A is identical to B, then $JSC(A,B) = 1$. If two sets have no common elements, then $JSC(A,B) = 0$. If Jaccard similarity coefficient is directly applied to calculate the similarity between M_1 and M_2, the result is as follows.

$$\mathrm{JSC}(M_1, M_2) = \frac{\sum_1^{424} Minimum(n_{1\,i}, n_{2\,i})}{\sum_1^{424} Maximum(n_{1\,i}, n_{2\,i})}$$

However, this result only considers a molecule as a mathematical set containing the groups forming the molecule. In other words, it never considers the chemical features of a molecule, mainly the type of compound and the size of molecule here. As for the type of compound, the similarity of two molecules belonging to the same type of compound should be higher than that belonging to two different types of compounds. For example, based on the Jaccard similarity coefficient, the similarity of $CH_3(CH_2)_6SO_2CH_3$ and $CH_3(CH_2)_6CH_3$ is 0.89 while the similarity of $CH_3(CH_2)_6SO_2CH_3$ and $CH_3(CH_2)_4SO_2CH_3$ is 0.78. However, chemically speaking, the first set of molecules belong to two types of compounds and the second set of molecules belong to the same type of compound. Therefore, this leads to a phenomenon that the similarity of two molecules belonging to different types is higher than that belonging to the same type, which is unacceptable. As for the size of molecule, the difference of two molecules is determined by the absolute difference of the structure not the relative difference of the structure. For example, for molecules described by $[1,1,0,0,\cdots]$ and $[1,1,1,0,\cdots]$, their similarity is 0.67 while the similarity of molecules $[1,1,1,1,1,1,1,1,0,0,\cdots]$ and $[1,1,1,1,1,1,1,1,1,0,\cdots]$ is 0.89. The structure differences in these two sets of molecules are both one group, but their difference of similarity is so large. This is because of the difference of the molecular size. For small size of molecules, the effect of one different group is more significant than large size of molecules. However, the information one molecules can provide for another molecule in the first set is also relatively valuable like the second set of molecules. So, the difference on the similarity due to the molecular size is too large and needs to be reduced. Additionally, if it is not reduced, only large molecules can find similar molecules and small molecules are hard to find similar molecules which is unacceptable as well.

Based on the above analysis, this paper proposed a molecular similarity coefficient calculation method which considers the molecule as a chemical concept. First, consider the type of compound. The type of compound is determined by the type of group inside its molecule. If there

are some groups whose number is 0 in one molecule but not 0 in another molecule, then these molecules belong to two different types. If two molecules are composed of the same types of groups but only the number is different, then these molecules belong to the same type and their similarity increases as the number becomes closer. Therefore, a warping function here is used to transform the molecular group vector, as follows,

$$y = \log(x + 1)$$

The warping function needs to satisfy that the derivative is maximum when the input is near 0, and gradually becomes smaller as the input becomes larger, which conforms to the above analysis of molecular type discrimination. Then molecules M_1 and M_2 are transformed as follows,

$$M_1^{warping}: [\log(n_{1\,1}+1), \cdots, \log(n_{1\,i}+1), \cdots, \log(n_{1\,424}+1)]$$

$$M_2^{warping}: [\log(n_{2\,1}+1), \cdots, \log(n_{2\,i}+1), \cdots, \log(n_{2\,424}+1)]$$

Their intersection and union are:

$$Intersection^{warping}: [a_1, a_2, \cdots, a_i, \cdots, a_{424}], a_i = min(log(n_{1\,i}+1), log(n_{2\,i}+1))$$

$$Union^{warping}: [b_1, b_2, \cdots, b_i, \cdots, b_{424}], b_i = max(log(n_{1\,i}+1), log(n_{2\,i}+1))$$

Then, consider the molecular size. Since the ratio of the elements number of the intersection and the union is a relative comparison but the difference of the elements number is the absolute comparison, the division operation of the intersection and union elements needs to be converted into a subtraction operation. Therefore, for M_1 and M_2 after warping transformation, the following method is used to calculate their similarity,

$$\mathrm{MSC}_\alpha(M_1, M_2) = \frac{\alpha^{\sum_1^{424} a_i}}{\alpha^{\sum_1^{424} b_i}} = \alpha^{\sum_1^{424} a_i - \sum_1^{424} b_i}$$

where α is a parameter greater than 1. The number of intersection and union elements in the formula appear in the exponents in the numerator and denominator respectively, so the division is converted to subtraction. Finally, in order to convert the value range of the molecular similarity coefficient to [0,1], it is necessary to normalize it, as shown below.

$$\mathrm{MSC}_\alpha(M_1, M_2) = \frac{\alpha^{\sum_1^{424} a_i} - \alpha^{\min(\sum_1^{424} a_i)}}{\alpha^{\sum_1^{424} b_i} - \alpha^{\min(\sum_1^{424} a_i)}} = \frac{\alpha^{\sum_1^{424} a_i} - 1}{\alpha^{\sum_1^{424} b_i} - 1}$$

Here, $\alpha^{\min(\sum_1^{424} a_i)}$ is a normalized item, which is 1 obviously when the intersection elements are all 0.

3. MODELING FRAMEWORK BASED ON MOLECULAR SIMILARITY

When the formula of molecular similarity is proposed, a modeling framework based on it can be further proposed. The dataflow is shown in Figure 1. For a target molecule and a known database containing m molecules, the similarity between the target molecule and all the molecules in the dataset should be calculated one by one to obtain m similarities, namely, MSC_1, MSC_2, $\cdots$, MSC_m. Then, the n molecules having the highest similarities will be selected to form a subset, which contains the molecules which can provide the most valuable information for the property prediction of the target molecule. Then this subset is used as the training dataset to train the model and use the target molecule as the input to get the predicted value of its property. Simultaneously, the similarities of the customized training molecules to the target molecule can be used to form a quantitative index that reflects the prediction reliability of the target molecule. In this paper, the maximum similarity of all the training molecules is employed as the reliability index defined by R. Consequently, the prediction result using this modeling framework is composed of two parts, that is, the predicted property and the reliability index R, which are used together to guide the molecular design. Additionally, the customized training set can provide much more pertinent and valuable information for the target molecule to improve the prediction accuracy of the target molecules with high reliability R.

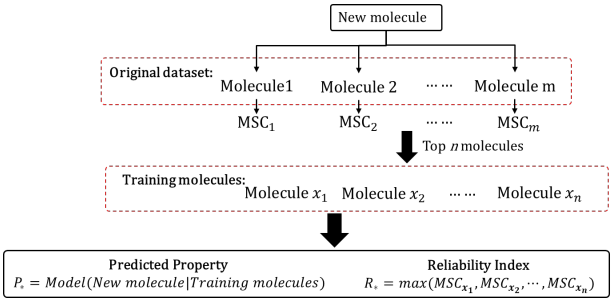

Figure 1. The dataflow of the modeling framework.

4. TEST AND DISCUSSION

4.1 Test method

To verify whether the property prediction framework proposed in this paper possesses the characteristics outlined, a series of tests were conducted. The modeling framework does not limit the types of models used, so three property prediction models were employed for the tests: a simple Group Contribution (GC) model[9], a Support Vector Regression (SVR) model[6], and an improved Gaussian Process Regression (GPR) model[10].

For each model, a test group utilized the proposed modeling framework, while a control group did not, allowing for a comparison to assess the framework's effectiveness. To ensure the test results are objective and reliable, nine different properties were selected for testing, as shown in Table 1. The data for each property was

divided into a candidate training set and a testing set, using an 80:20 ratio. Customized training molecules were chosen from the candidate training set.

Table 1: The information of 9 testing properties

	Property	Symbol	Data volume
1	Critical volume (mL/mol)	Vc	636
2	Critical pressure (bar)	Pc	651
3	Auto ignition temperature (K)	Ait	477
4	Gibbs energy of formation at 298 K (kJ/mol)	Gf	616
5	Standard enthalpy of formulation (kJ/mol)	Hf	787
6	Hildebrandt solubility parameter at 298 K ($MPa^{1/2}$)	Hsolp	1102
7	Enthalpy of formation at 298 K (kJ/mol)	Hv	326
8	Liquid molar volume at 298 K (mL/mol)	Lmv	855
9	Critical temperature (K)	Tc	706

To determine the optimal number of molecules to achieve the best prediction accuracy, the test group was further separated into two subgroups: using 10 molecules and 50 molecules for training. In contrast, the control group used all the molecules from the candidate training dataset for training. Finally, the predicted values and reliability indices for the molecules in the testing set were recorded.

4.2 Result analysis

4.2.1 Verify the quantification of the prediction reliability

In order to verify that the reliability index R can really be used as a quantitatively measure for the prediction reliability, two cases are displayed first in Figure 2.

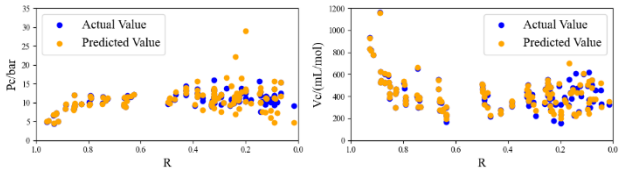

Figure 2. The results of Pc and Vc.

It can be seen that as R decreases from 1 to 0, the gap becomes much larger gradually. Further, the test results of all properties using the proposed modeling framework are displayed in Figure 3. Here, to eliminate the interference of some outlier data points, the following formula is used to calculate the Average Prediction Error (APE) of test molecules with reliability greater than u,

$$APE(R \geq u) = \frac{\sum_{R_i \geq u} error_i}{\sum_{R_i \geq u} 1}$$

All the test results are displayed with reliability R as the horizontal coordinate and APE as the vertical coordinate. From these results, as R decreases, APE always has an increasing tendency. Besides, when fewer molecules are used for training, APE increases more rapidly. From the result analysis above, the reliability index R based on the proposed molecular similarity coefficient can be used as a quantitative index for the reliability of property prediction without model limitation.

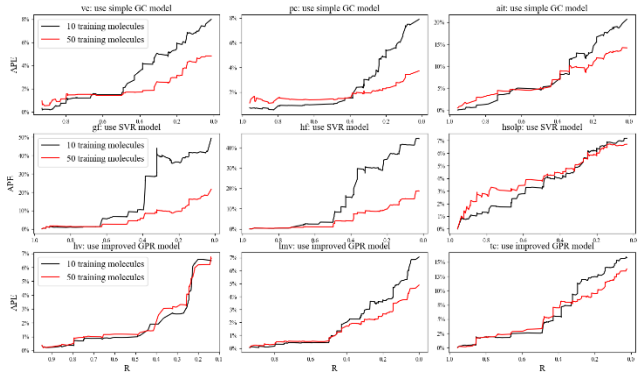

Figure 3. The results of 9 properties (list 3 properties for each model: Vc, Pc and Ait use simple GC model; Gf, Hf and Hsolp use SVR model; Hv, Lmv and Tc use improved GPR model).

4.2.2 Verify the improvement of the prediction accuracy

Then, in order to verify the prediction accuracy improvement of the proposed modeling framework, the results of all the properties are summarized according to different number of training molecules, namely, 10 training molecules, 50 training molecules and all the candidate training molecules. The average prediction error under different level of reliabilities is shown for different number of training molecules. The results using simple GC model, SVR model and improved GPR model are listed in Table 2, Table 3 and Table 4.

For each line representing the same level of reliability, we compare the prediction errors among three columns that utilize different numbers of training molecules. The green color scale indicates the case with the lowest prediction error, followed by a light green scale, and finally, the white scale, which represents the highest prediction error. From these tables, it is evident that regardless of the model used, when dealing with molecules that have high reliability (R), applying a subset of molecules from the candidate training dataset—specifically, the proposed modeling framework—results in significantly better prediction accuracy compared to the routine method that uses all candidate training molecules. Only when the reliability (R) is at a low level does the prediction accuracy of using all candidate training molecules surpass that of the proposed modeling framework.

The primary reason for this result is that when molecules with high similarity to the target molecules are identified, using a subset of these molecules can provide significantly more relevant information for predicting the properties of the target molecules. In contrast, molecules

Table 2: The result of simple GC model

Reliability level	10 training molecules	50 training molecules	All candidate training molecules
R≥0.9	0.67%	1.22%	2.70%
R≥0.85	0.81%	1.56%	3.00%
R≥0.8	0.89%	1.72%	3.09%
R≥0.7	1.30%	1.96%	3.26%
R≥0.6	1.93%	2.14%	3.43%
R≥0.5	2.05%	2.29%	3.47%
R≥0.4	4.70%	3.45%	4.48%
R≥0.3	9.62%	5.87%	7.00%
R≥0.2	11.04%	6.82%	7.40%
R≥0.1	13.77%	8.94%	8.48%
R≥0	15.56%	10.17%	9.61%

Table 3: The result of SVR model

Reliability level	10 training molecules	50 training molecules	All candidate training molecules
R≥0.9	0.63%	1.19%	2.74%
R≥0.85	0.81%	1.78%	2.89%
R≥0.8	1.29%	1.73%	2.86%
R≥0.7	2.22%	2.62%	3.06%
R≥0.6	3.77%	3.48%	3.38%
R≥0.5	4.10%	3.57%	3.48%
R≥0.4	7.09%	4.91%	4.34%
R≥0.3	13.19%	6.68%	6.11%
R≥0.2	14.06%	7.26%	6.58%
R≥0.1	16.13%	9.30%	7.82%
R≥0	17.83%	10.92%	8.92%

Table 4: The result of improved GPR model

Reliability level	10 training molecules	50 training molecules	All candidate training molecules
R≥0.9	0.59%	0.80%	1.01%
R≥0.85	0.69%	0.88%	1.08%
R≥0.8	0.76%	0.92%	1.21%
R≥0.7	1.18%	1.48%	1.72%
R≥0.6	1.80%	1.93%	2.09%
R≥0.5	1.92%	2.05%	2.20%
R≥0.4	5.37%	4.22%	4.07%
R≥0.3	10.92%	5.95%	5.68%
R≥0.2	12.48%	7.32%	6.62%
R≥0.1	14.86%	9.60%	8.08%
R≥0	16.00%	11.44%	9.50%

that are too dissimilar from the target molecules tend to offer little valuable information and may even lead to misleading conclusions; therefore, they have been excluded. This process enhances prediction accuracy. As the similarity among molecules in the training set decreases, the value and relevance of the information also diminish. If a small number of training molecules are still used, the prediction accuracy will decline due to insufficient data. Consequently, it is essential to increase the amount of data. When the reliability (R) approaches 0, it becomes challenging to find similar molecules for the target molecules within the candidate training dataset. At this point, maximizing the amount of data is crucial—using all available candidate training molecules is necessary to achieve the best prediction accuracy. However, even when all data is utilized, the informational value of these molecules regarding the target molecules remains minimal, resulting in significant prediction errors and low reliability, as demonstrated in the tables.

5. DATA CONSISTENCY TEST

Based on the molecular similarity coefficient, a methodology for data consistency testing can be established. If molecules with high MSC serve as the training dataset but the predicted property value of the target molecule deviates significantly from the actual value, it suggests that errors may exist in the property data of the training molecules or the target molecule itself.

For instance, in a property database of Tc, the prediction error for $CH_3(CH_2)_{16}COOH$ reached 10.55% while its prediction reliability was 0.9437. Table 5 details the 10 training molecules for further analysis.

Table 5: The training data before correction

No	MSC	Tc (real)/K	CH_3	CH_2	COOH
Target molecule	1	799	1	16	1
Molecule 1	0.943662	474.2	1	17	1
Molecule 2	0.940299	590	1	15	1
Molecule 3	0.893333	809	1	18	1
Molecule 4	0.880597	785	1	14	1
Molecule 5	0.820896	693	1	13	1
Molecule 6	0.761194	568	1	12	1
Molecule 7	0.701493	591.15	1	11	1
Molecule 8	0.641791	618	1	10	1
Molecule 9	0.58209	654	1	9	1
Molecule 10	0.522388	456.86	1	8	1

It can be noted that these molecules are of the same compound type, with the primary difference being the number of CH_2 groups. According to physicochemical principles, the property Tc of these molecules generally varies consistently with the number of carbon atoms (C). Consequently, the relationship between Tc and the number of carbon atoms before correction has been plotted in Figure 4. The Tc values of these molecules do not exhibit a consistent relationship with the number of carbon atoms (C). This observation suggests a significant possibility of errors in the Tc data for either the training molecules or the target molecule. To address this, we verified the actual Tc values of these molecules by ICAS[8] and discovered mistakes in the training data. The corrected data is shown in Table 6.

Table 6: The training data after correction

No	MSC	Tc (real)/K	CH_3	CH_2	COOH
Target molecule	1	799	1	16	1
Molecule 1	0.943662	812[a]	1	17	1
Molecule 2	0.940299	793[a]	1	15	1
Molecule 3	0.893333	821[a]	1	18	1
Molecule 4	0.880597	785	1	14	1
Molecule 5	0.820896	775[a]	1	13	1
Molecule 6	0.761194	765[a]	1	12	1
Molecule 7	0.701493	754[a]	1	11	1
Molecule 8	0.641791	734[a]	1	10	1
Molecule 9	0.58209	732[a]	1	9	1
Molecule 10	0.522388	722.1[a]	1	8	1

a: checked in ICAS (Integrated Computer Aided System).

Similarly, the Tc values and the number of C after correction are plotted in Figure 4. It illustrates that after correction, the Tc values and the number of C atoms

exhibit a consistently increasing trend. This suggests that the dataset has likely been corrected at this stage. The corrected dataset is then utilized to predict Tc of the target molecule, revealing a prediction error of only 0.42%. This result further confirms that the errors in the original dataset have been addressed.

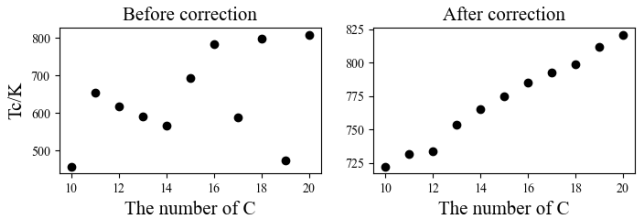

Figure 4. The data of Tc with the number of C.

6. CONCLUSION

This paper introduces a modeling framework based on the molecular similarity coefficient. This framework utilizes the similarity between molecules in a candidate training dataset and target molecules to create a customized training dataset specifically for the target molecules. Testing was conducted on nine properties. The results confirm that this modeling framework provides a quantitative measure of prediction reliability and significantly enhances the prediction accuracy for molecules with a high reliability index. For molecular design, the framework proposed in this paper offers a novel two-fold evaluation system. It emphasizes that the molecular design process should assess not only the predicted properties but also the reliability of those predictions. However, the analysis presented in this paper indicates that the prediction performance of the proposed framework is somewhat sensitive to data errors because using a smaller amount of data for model training amplifies the impact of mistakes in the training dataset. Therefore, the data consistency testing method proposed in this study should be further developed, and the modeling framework should be implemented alongside this testing method to achieve more reliable, accurate, and stable prediction performance.

ACKNOWLEDGMENTS

Financial support from the Natural Science Foundation of China (No. 22150410338) is gratefully acknowledged.

REFERENCE

1. Gani R. Group contribution-based property estimation methods: advances and perspectives. Curr Opin Chem Eng 2019;23:184–96. https://doi.org/10.1016/j.coche.2019.04.007.
2. Le T, Epa VC, Burden FR, Winkler DA. Quantitative structure-property relationship modeling of diverse materials properties. Chem Rev 2012;112:2889–919. https://doi.org/10.1021/cr200066h.
3. Wen S, Nanda K, Huang Y, Beran GJO. Practical quantum mechanics-based fragment methods for predicting molecular crystal properties. Physical Chemistry Chemical Physics 2012;14:7578–90. https://doi.org/10.1039/c2cp23949c.
4. Jirasek F, Hasse H. Machine Learning of Thermophysical Properties. Fluid Phase Equilib 2021;549. https://doi.org/10.1016/j.fluid.2021.113206.
5. Zhang J, Wang Q, Su Y, Jin S, Ren J, Eden M, et al. An accurate and interpretable deep learning model for environmental properties prediction using hybrid molecular representations. AIChE Journal 2022;68. https://doi.org/10.1002/aic.17634.
6. Alshehri AS, Tula AK, You F, Gani R. Next generation pure component property estimation models: With and without machine learning techniques. AIChE Journal 2022;68. https://doi.org/10.1002/aic.17469.
7. Joback KG, Reid RC. Estimation of Pure-Component Properties from Group-Contributions. Chem Eng Commun 1987;57:233–43. https://doi.org/10.1080/00986448708960487.
8. Gani R, Hytoft G, Jaksland C, Jensen AK. An integrated computer aided system for integrated design of chemical processes. Comput Chem Eng 1997;21:1135–1146. https://doi.org/10.1016/S0098-1354(96)00324-9.
9. Hukkerikar AS, Sarup B, Ten Kate A, Abildskov J, Sin G, Gani R. Group-contribution + (GC +) based estimation of properties of pure components: Improved property estimation and uncertainty analysis. Fluid Phase Equilib 2012;321:25–43. https://doi.org/10.1016/j.fluid.2012.02.010.
10. Cao X, Gong M, Tula A, Chen X, Gani R, Venkatasubramanian V. An Improved Machine Learning Model for Pure Component Property Estimation. Engineering 2024;39:61–73. https://doi.org/10.1016/j.eng.2023.08.024