ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

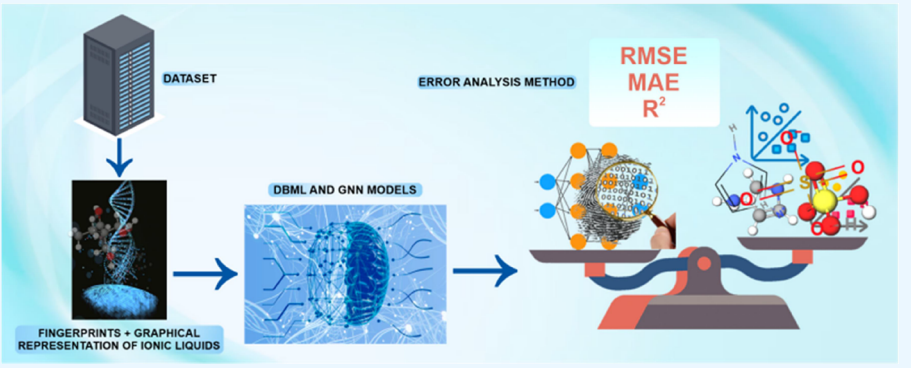
Prediction and Interpretability of Melting Points of Ionic Liquids Using Graph Neural Networks

Haijun Feng,* Lanlan Qin, Bingxuan Zhang, and Jian Zhou

Cite This: *ACS Omega* 2024, 9, 16016−16025

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



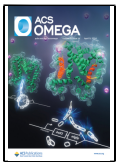
ABSTRACT: Ionic liquids (ILs) have wide and promising applications in fields such as chemical engineering, energy, and the environment. However, the melting points (MPs) of ILs are one of the most crucial properties affecting their applications. The MPs of ILs are affected by various factors, and tuning these in a laboratory is time-consuming and costly. Therefore, an accurate and efficient method is required to predict the desired MPs in the design of novel targeted ILs. In this study, three descriptor-based machine learning (DBML) models and eight graph neural network (GNN) models were proposed to predict the MPs of ILs. Fingerprints and molecular graphs were used to represent molecules for the DBML and GNNs, respectively. The GNN models demonstrated performance superior to that of the DBML models. Among all of the examined models, the graph convolutional model exhibited the best performance with high accuracy (root-mean-squared error = 37.06, mean absolute error = 28.79, and correlation coefficient = 0.76). Benefiting from molecular graph representation, we built a GNN-based interpretable model to reveal the atomistic contribution to the MPs of ILs using a data-driven procedure. According to our interpretable model, amino groups, S^+, N^+, and P^+ would increase the MPs of ILs, while the negatively charged halogen atoms, S^-, and N^- would decrease the MPs of ILs. The results of this study provide new insight into the rapid screening and synthesis of targeted ILs with appropriate MPs.

1. INTRODUCTION

Ionic liquids (ILs) are nonvolatile salts with melting points (MPs) below 100 °C, low vapor pressures, and strong thermal conductivity. ILs are widely used in greenhouse gas capture, catalysis, energy, separation, electrochemistry, pharmaceuticals, etc.[1−5] The MP of an IL is a crucial factor when using ILs for its application;[6] ILs with MPs < 100 °C are required by IL-based pharmaceuticals to avoid problems associated with polymorphism, agrochemicals, and the sorption of a precipitated artificial solid deposit.[7] Designing task-specific ILs with desired MPs is challenging because a wide range of factors, including hydrogen bonding, van der Waals interaction, and charge distribution, can affect the MPs.[4] Because trillion types of ILs can be synthesized in a lab, finding the appropriate ILs via experimental screening is costly and time-consuming.[8,9] Quantitative structure−property relationship (QSPR) studies[10−12] have been used to accurately forecast the MPs of ILs. The goal of QSPR techniques is to create mathematical representations of numerical properties based on the structural details of chemical substances.[13] However, owing to the intricacy of molecular interactions, conventional QSPR techniques, density functional theory, and molecular dynamics might be computationally difficult for large-scale ILs.[14,15]

Machine learning (ML) techniques show promise for accurately and effectively predicting the properties of chemical compounds.[9,16−20] In different fields, ML methods[8,21] are as accurate as traditional simulation techniques such as MD but require less computing power.[14,22] Molecular descriptor−based ML (DBML) models have been used to forecast the MPs of ILs.[23,24] Molecular descriptors are the numerical values that

Received: November 29, 2023
Revised: March 13, 2024
Accepted: March 15, 2024
Published: March 28, 2024



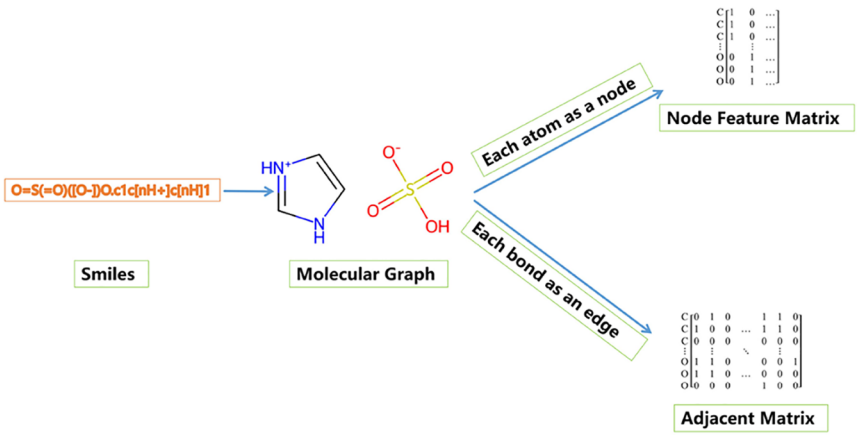


Figure 1. Illustration of constructing graphs from molecules with imidazolium hydrogen sulfate as an ionic liquid.

feature the molecular structures of the ILs. One of the earlier descriptors used group contribution,[25,26] which split molecules into fragments; the MP of an IL is calculated by summing the contribution of each fragment. However, this descriptor is largely dependent on human experience and could result in substructure information loss.[27,28] Another molecular descriptor, extended-connectivity fingerprint (ECFP),[29] creates a feature vector by iteratively compiling the neighboring data of each atom. Because ECFPs have access to molecular structural information, they can be more expressive in estimating the MPs of ILs. Some ML studies have been conducted to estimate the MPs of the ILs. To estimate the MPs of 2212 ILs, Venkatraman et al.[30,31] explored several ML techniques based on 113 molecular descriptors, which were calculated from computationally cheap semiempirical simulations. Their models could reasonably forecast MP trends by applying a tree-based ensemble technique. Makarov et al.[27,28] used eight ML approaches to predict the glass transition temperature, MP, and decomposition temperature of ILs over five sets of descriptors. Their transformer convolutional neural network (CNN) model exhibited a good correlation coefficient (R^2) of 0.67 and a root-mean-squared error (RMSE) of 44 °C when predicting the MPs. DBML models are capable of producing accurate estimations of MPs. However, they lack information regarding the molecular graph structure and are unable to completely explain the ML model results. Additionally, current ML research on the MPs of ILs often focuses on prediction performance evaluation; the model interpretability of the MPs of ILs has not been thoroughly examined.

Recently, the graph neural network (GNN) has been demonstrated as an effective tool for predicting the molecular properties of chemical compounds at the molecular level.[32,33] GNNs directly use the structure of molecules as input for modeling.[34] Considering a molecule as a graph, GNNs use an adjacent matrix to record the bond edge and connectivity properties and a node feature matrix to represent the atom and associated attributes.[14] GNNs have been used in molecular-level fields[35−38] such as drug discovery, quantum chemistry, and structural biology. In addition to achieving comparable high-prediction performance, GNNs explain the results of models at the atom and bond levels via graphical representation. Numerous recent approaches have been developed in the current field of research to interpret GNN predictions.[39−41] To more accurately predict and comprehend the MPs of ILs at the molecular level, GNNs with model interpretability need to be further extended on the modeling of MPs of ILs.[14]

In this study, the DBML and GNN models are introduced to predict the MPs of ILs. Fingerprints (FPs) are incorporated as molecular descriptors for several DBML models. Eight distinct GNN algorithms are included to estimate the MPs of ILs in GNN models. We initially assessed the performance of various models in predicting MPs using a diverse data set of 3080 ILs. Models that worked with molecules were interpreted by computing atomic contributions using the graph representation of GNNs. We examined the positive and negative impacts of atoms on the MPs of ILs to obtain significant insight into the atomistic level of an IL molecule. Additionally, we ranked the atomic contributions to the MPs of ILs across the entire data set. We believe that this is one of the first studies to apply GNNs and the interpretability model at the atomistic level for predicting the MPs of ILs.

2. MODELS

2.1. DBML Models. This work used ECFPs[29] as the input characteristics for modeling. ECFPs calculate a representation of a molecule in the form of a bag of words by decomposing it into local neighbors and hashing it into a bit vector. FPs with a bit vector of 2048 and a radius of 4 were produced using RDKit[42] for each molecule.

The MPs of ILs were predicted using three different ML techniques: support vector machine (SVM),[43] random forest (RF),[44,45] and multilayer perceptron (MLP)[46] techniques. The input variables for these ML algorithms were molecular descriptors and ECFPs.[29] The squared epsilon insensitive of SVM[43] was specified as the loss function, and the maximum number of iterations was set to 1000. The number of trees of RF[44,45] was 1000 and the squared error was used to measure the quality of a split. The LBFGS solver of MLP[46] was selected for weight optimization, and relu was chosen as the activation function in the hidden layer with a size of 100. These ML models were built using the Python code based on DeepChem[42] and SciKit-Learn.[47,48]

2.2. GNN Models. The GNN operates on a graph with nodes and edges[49,50] rather than descriptors. Each molecule represented a graph in this work, with each atom and bond acting as a node and edge, respectively. The construction of molecular graphs is shown in Figure 1. RDKit is used to transform the notation of the canonical simplified molecular-input line-entry system (SMILES) of each molecule into a

https://doi.org/10.1021/acsomega.3c09543
ACS Omega 2024, 9, 16016−16025

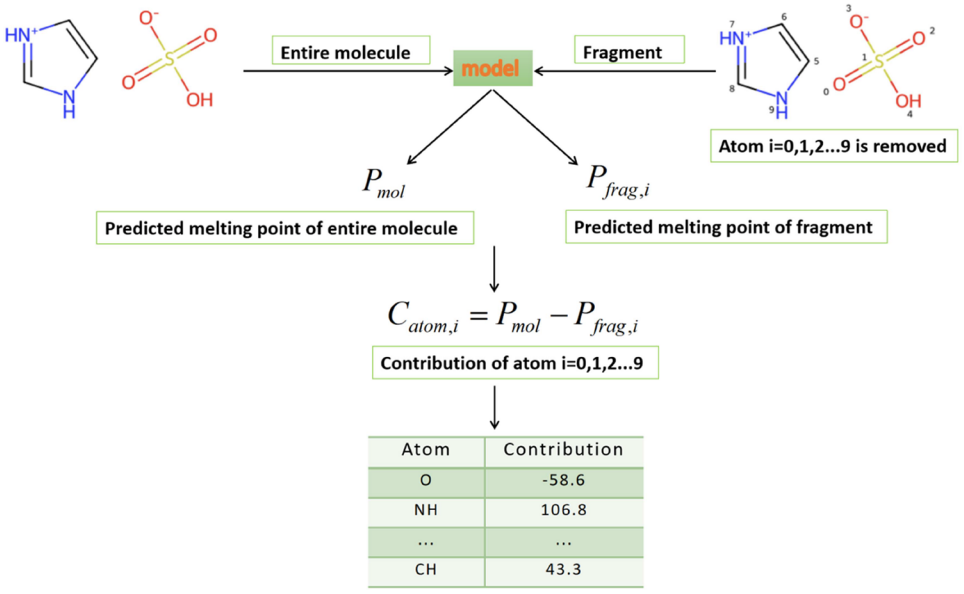


Figure 2. Calculation scheme for the atomic contribution.

molecular graph. Nodes and edges in the graph represent atoms and their interatomic interactions, respectively. A feature vector containing specific atom data is allocated to each node. One input for modeling is an N-by-C node feature matrix, where N and C denote the number of nodes and dimension of node features, respectively. The node feature vector includes details of the formal charge of atoms, degree of atom hybridization, aromaticity, and other properties. The edge feature vectors, containing details of the bond type, etc., are involved on the occasion. Furthermore, the bond connectivity of the molecular graph is represented by an N-by-N adjacent matrix.

GNN is a deep learning technique that requires the node feature matrix and adjacent matrix as inputs for processing data to be represented as graphs. Although DBML approaches might describe the structure information on molecules by adhering to certain criteria, they have difficulty encoding the topological and geometric information. GNNs, benefiting from molecular graph representation, can encode the graphical structure of molecules.

The MPs of ILs in this study were modeled by using eight common powerful GNN techniques: graph convolution network (GCN),[51] graph attention network (GAT),[52] Attentive FP,[53] path-augmented graph transformer network (PAGTN),[54] message passing neural network (MPNN),[55] directed acyclic graph (DAG),[56] weave,[50] and graph convolutional (GC)[49] techniques. DeepChem[42] was used to develop all GNN models.

The GCN model[51] updates the representation of nodes in graphs using a GCN variation. It calculates the representation of each graph by adding the weighted sum of its node representations, where weights are determined by computing a gating function on the node representation. The initial atom feature vectors have in this study the mentioned parameters: length = 30, the channel width of GC layers = (64,64), hidden representation size in the output MLP predictor = 128, and batch size = 128.

The GAT model[52] enhances the expressiveness of GNNs via the attention mechanism. Graph attention allocates a learnable weight for every edge while performing feature aggregation on nodes. In this study, each GAT layer has eight attention heads; the hidden representation size in the output MLP predictor is 128, 30 atom features, a batch size of 128, and the channel width per attention head for GAT layers is (8,8).

In the AttentiveFP model,[53] the initialization of node representations, entailing a round of message passing, mixes node features and edge information. It uses a gated recurrent unit to combine all node representations for each graph to determine the representation of each graph. Graph representations have a size of 200, the number of GNN layers is 2, the initial atom and bond feature vectors are 30 and 11, respectively, and the batch size is 128 in this study.

A GAT variation in the PAGTN model[54] is used to modify node representations in graphs by using a linear additive kind of attention. Concatenating the node and edge information on each bond yields attention weights. In this work, the input edge feature size is 42, the input and output node feature sizes are 94 and 256, respectively, two layers in the GNN, the dropout probability is 0.1, and the batch size is 128.

The node representations are updated in the MPNN model[55] by combining the most recent node representations with edge information, requiring numerous iterations of message passing. It computes the representation of each network by aggregating the representations of all of its nodes using a Set2Set layer. A graph data object with both node and edge characteristics must be produced by the feature generator when it is used with the MPNN model. In this study, the final node representation vectors and the hidden edge representation vectors are 64 and 128, respectively; there are six Set2Set steps and three Set2Set layers. Initial atom and bond feature vectors are 30 and 11, respectively, and the batch size is 128.

Molecules are regarded as directed graphs in the DAG model.[56] Although most chemical bonds lack inherent directions, a DAG on a molecule can be generated arbitrarily by identifying a central atom and specifying the directions of all bonds in certain orientations toward the atom. The batch size is 128 and each atom has 75 features in this study.

The weave model[50] uses the concept of adaptive learning to extract significant representations. The features of atoms are updated in weave models by incorporating data from all of the

https://doi.org/10.1021/acsomega.3c09543
ACS Omega 2024, 9, 16016−16025

other atoms and their associated pairing in the molecule. The main distinction lies in the size of the convolutions. Each dense layer in the network has a dimension of (2000, 100), the dropout for each fully connected layer is 0.25, the number of atom features is 75, and the batch size is 128 in this study. Each molecule has 128 output features.

The circular fingerprint decomposition concepts are expanded in the GC model.[49] The data vector of each graph node is the starting point. Convolutional and pooling layers mix and recombine information from connected nodes into descriptors to create a new data vector for each linked node. Contrary to the GCN model, this approach uses distinct labile weights for nodes with varying degrees. The learnable weight in the GCN model is shared by all of the nodes. Each atom has 75 features, the batch size is 128, and the model is adjusted for batch normalization in this study. The widths of channels of GC and atom-level dense layers after GraphPool are (64, 64) and 64, respectively.

2.3. Interpretability Model. GNN models can interpret our models and aid us in understanding the output by using molecular graphs. We examined the contribution of each atom of the IL molecule to the MP of the IL. The scheme is depicted in Figure 2. We removed one heavy atom at a time for each IL molecule. The remaining fragment was featured and predicted using the same trained model. Finally, the contribution of each atom to the MP is calculated by eq 1.

$$C_{\text{atom},i} = P_{\text{mol}} - P_{\text{frag},i} \tag{1}$$

The contribution of each atom ($C_{\text{atom},i}$) is determined via the prediction discrepancy of the MP between the entire molecule (P_{mol}) and left-over fragment after the removal of an atom ($P_{\text{frag},i}$), where i represents the index of each atom in the corresponding IL molecule. The influence of each atom on prediction and the contribution of atoms to the MPs of ILs were studied by examining changes in the prediction of the model.

We developed a weight function to measure the contribution of each atom relative to the entire data set. We calculated the weight (W_i) of each heavy atom (i) inside an IL molecule (j) using eq 2 for the entire data set.

$$W_i = \frac{\sum_{j=1}^{N_{\text{mol}}} (P_{\text{mol}} - P_{\text{frag},i})}{N_{\text{mol}}} \tag{2}$$

where N_{mol} denotes the total number of molecules of ILs that contain the specific atom i across the entire data set. The molecule would be counted for the corresponding times if i appears multiple times in one molecule. Positive and negative weights are calculated separately so that atoms having positive or negative contributions to the MPs of ILs can be determined by ranking the weights of all atoms in the entire data set.

3. DATA SETS AND METRICS

3.1. Data Sets. The data set used in this study includes MP values for 3080 different ILs, representing a wide range of IL families containing imidazolium, ammonium, pyrrolidinium, sulfonium, and other cations and tetrafluoroborate (BF_4), chloride (Cl), (trifluoromethylsulfonyl) amide (Tf_2N), and other anions. The data are based on studies by Makarov et al.[27,28] and Venkatraman et al.[30,31] reporting experimental MPs (between −96 and 319 °C) from published works. A canonical SMILES code is used to represent each IL in the data set. Figure 3 displays the MP distributions based on their quantity. The data sets are divided into training, validation, and test sets following the 80:10:10 ratio.

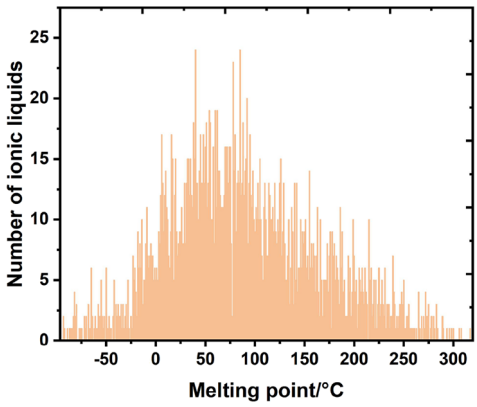


Figure 3. Distribution of the MPs of ILs in the data set.

3.2. Data Set Splitting. Random splitting is the most popular approach for dividing the data set into training and test sets. However, random splitting is not always the most effective technique for assessing models. Five distinct data set splitting techniques were used to evaluate the models in this study: random, random stratified,[42] scaffold,[57] fingerprint,[42] and molecular weight[42] splitting techniques. Training, validation, and test subsets are created randomly by dividing the samples in the random splitting method. The random stratified splitting[42] approach sorts data points in an ascending order of the label value and divides this sorted list into training, validation, and test sets with each set including the whole set of available labels. Scaffold splitting[57] divides the samples according to their two-dimensional structural frameworks and segregates structurally distinct molecules into several subsets. The fingerprint splitting[42] approach divides data sets into training, validation, and test sets based on the Tanimoto similarities of their ECFP4 fingerprints. This method aims to separate the data to make the molecules in each data set as dissimilar to each other as possible. Molecular weight splitting[42] uses the molecular weight determined using the SMILES string to divide internal compounds into training, validation, and test sets.

The Supporting Information (Table S1) displays the results of model evaluation using various data-splitting techniques. As the random stratified splitting[42] method produces a fairly accurate division with the best performance, we used the model with this data set splitting method for further interpretation analysis.

3.3. Metrics. RMSE, mean absolute error (MAE), and R^2 are presented to assess the accuracy of the models using eqs 3−5.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(ei - pi)^2} \tag{3}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}|e_i - p_i| \tag{4}$$

$$R^2 = \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2 - \sum_{i=1}^{n}(e_i - p_i)^2}{\sum_{i=1}^{n}(e_i - \bar{e})^2} \tag{5}$$

https://doi.org/10.1021/acsomega.3c09543
ACS Omega 2024, 9, 16016−16025

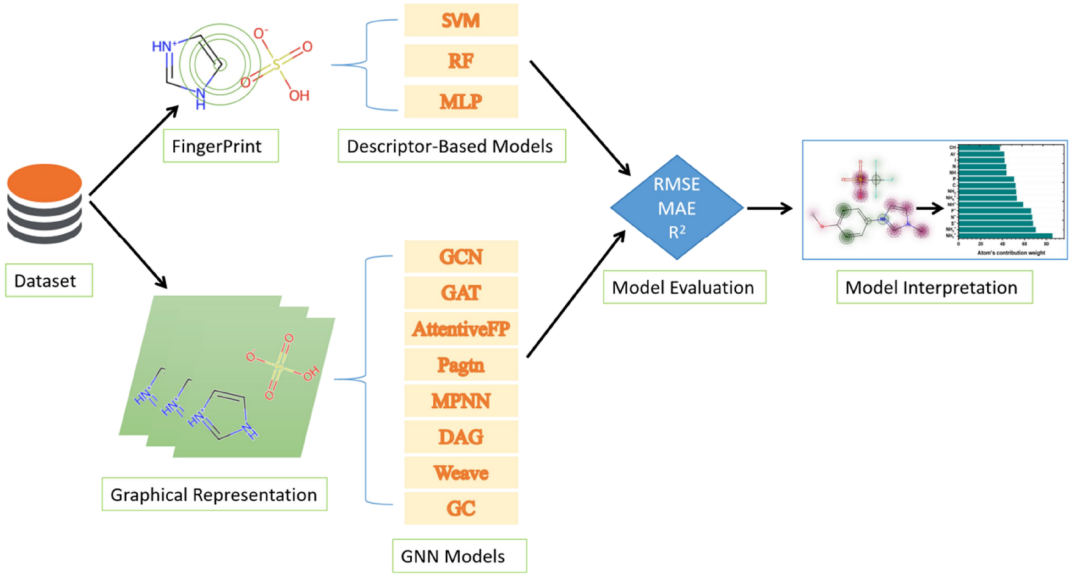


Figure 4. Overall scheme of modeling procedures in this work.

Table 1. Performances of Different Models on the MP Prediction of ILs

		training			test		
no.	model	RMSE	MAE	R^2	RMSE	MAE	R^2
1	SVM	37.56	25.05	0.74	44.36	33.12	0.68
2	RF	15.09	11.27	0.96	42.17	30.93	0.71
3	MLP	3.31	0.54	1.00	47.95	36.43	0.63
4	GCN	18.70	13.88	0.94	45.96	34.27	0.61
5	GAT	24.69	18.62	0.89	40.70	30.64	0.70
6	AttentiveFP	16.56	11.75	0.95	42.69	31.43	0.67
7	PAGTN	10.93	8.46	0.98	42.48	32.77	0.69
8	MPNN	14.74	10.52	0.96	43.17	31.60	0.61
9	DAG	6.21	4.25	0.99	45.93	35.54	0.65
10	weave	19.57	14.85	0.93	40.39	30.56	0.69
11	GC	14.00	10.07	0.96	37.06	28.79	0.76

where n is the number of data points, e_i is the laboratory MP values of ILs, p_i is the predicted MP values of ILs using models, and $\bar{e}$ is the average laboratory MP value of ILs.

3.4. Modeling Procedure. Figure 4 depicts the modeling process in this study. ECFPs and graph representations are used to depict IL molecules. Different ML models use ECFPs to perform the MP prediction. GNNs use graph representation to model the MPs of ILs; each model is trained for 500 epochs. Additionally, an explanation methodology is devised to analyze the significance of the atomic contribution to the MPs of ILs.

4. RESULT AND DISCUSSION

4.1. MP Prediction of ILs. The predicted MPs of ILs using different models are listed in Table 1. GNN models based on graph representation outperform ML models based on molecular fingerprints. GNN models can handle feature representation and extraction of graphs. They obtain excess molecular information and are superior to descriptor-based methods. The GC model having the lowest RMSE and MAE values of 37.06 and 28.79, respectively, and the highest R^2 value of 0.76 on the test set performs best among all of the models in this study. The GCN model is superior to RMSE and R^2 values of 45.96 and 0.61, respectively, because of the sharing of weights by all nodes and the simple update process of representations. Alternately, the GC model uses individual learnable weights for nodes of different degrees. As the GC model had the best performance, we utilized this model for further study.

Figure 5 shows the loss curve of the GC model. In this study, the MSE works as the loss score. The model converges after 500 epochs of training, indicating that the training is sufficient to form a well-trained model.

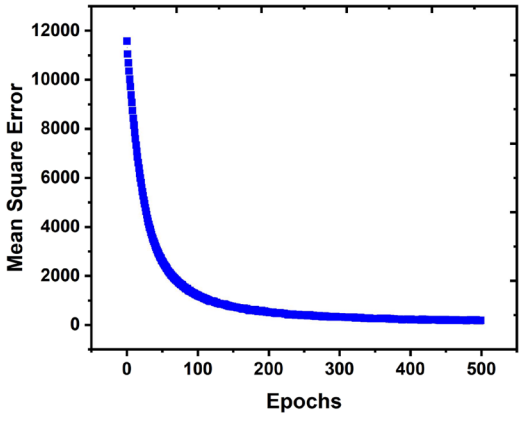


Figure 5. Training loss curve of the GC model. (B, C).

https://doi.org/10.1021/acsomega.3c09543
ACS Omega 2024, 9, 16016−16025

The parity plots of the predicted MPs of ILs vs experimental values using the GC model on test sets are illustrated in Figure 6. The visual comparison shows that the predicted values are consistent with the experimental data, indicating the high accuracy and reliability of the GC model.

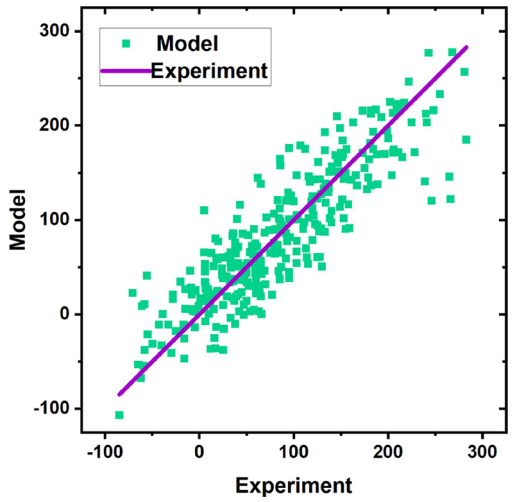


Figure 6. Parity plot of the predicted MPs with experimental data using the GC model on the test sets.

4.2. Comparison of Model Performances in Predicting MPs. The performance of different models is compared with the reported models, as shown in Table 2 with N representing the size of the whole data set. Table 2 shows that the GC model with the biggest size of the data set has a low RMSE, MAE, and high R^2 value, indicating that the performance of our model is the best among all the examined models.

Table 2. Performance Comparison of Different Models to Literature Values[a]

no.	model	N	RMSE	MAE	R^2
1	ANN[22]	799	33.33		0.54
2	QSPR[24]	808	26.85		0.72
3	RF[30,31]	2212	45.00	33.00	0.66
4	KRR[23]	2212	38.54	29.78	0.76
5	CNF[27,28]	3073	52.60	39.40	0.57
6	transformer CNF[27,28]	3073	46.60	35.00	0.64
7	transformer CNN[27,28]	3073	45.00	33.70	0.66
8	GC [this work]	3080	37.06	28.79	0.76

[a]ANN, artificial neural network; RF, random forest; KRR, kernel ridge regression; CNF, convolutional neural fingerprint; and CNN, convolutional neural networks.

4.3. Interpretation of the Model in Predicting MPs. The influence of IL molecular structure on MPs was studied by calculating the contribution of each atom in an IL using eq 1. The results of one IL (imidazolium hydrogen) are displayed in Table 3. C_{atom} refers to the contribution of atoms to the MP of the IL. A positive C_{atom} value indicates a positive impact, indicating that the atom would increase the MP of the corresponding IL. Alternately, a negative C_{atom} value indicates a negative impact, indicating that the atom would decrease the MP of the corresponding IL. The high absolute value of C_{atom}increases the impact. The NH^+ and NH of the imidazolium hydrogen sulfate−based IL have the highest C_{atom} values of 64.98 and 42.54, respectively. Hence, they have a positive impact on the MP, and their presence would significantly increase the MP of the IL. O^- and S have a negative impact with the lowest C_{atom} values of −55.51 and −53.82, respectively. Hence, O^- and S would effectively lower the MP of the corresponding IL.

The atomic contributions of each IL can be visualized using contribution maps.[58] The contribution maps of 3080 ILs (Figures S1−S3080) and the index number of each IL (Table S2) are provided in the Supporting Information. Atoms are colored based on their contributions to MPs in these maps. The influence of each atom on the MPs is represented by their colors. Green and red indicate positive and negative contributions, respectively, in our study. A darker shade of the color indicates more contribution than that of the other atoms.

We can intuitively observe the influence of each atom on the MPs of ILs from the 3080 nm visual maps. Cations are green for most ILs in this study, indicating that cations increase the MPs of ILs; anions are red, indicating that anions decrease the MPs of ILs. The map of IL-1815 is presented in Figure 7a. The cations and anions of some ILs have atoms that affect the MP favorably and unfavorably. Figure 7b shows that the presence of the anions O^- and S lowers the MP of IL 272, while F in anions increases the MP. A green IL indicates that it has excess atoms contributing positively, and its MP would be high. A red IL indicates that the IL has more atoms contributing negatively, and its MP would be low. The map of IL-1218 is presented in Figure 7c. All of the atoms in this IL show a positive impact. Hence, this IL has a high MP of 319 °C. Figure 7d shows the map of IL-384 where all atoms have a negative impact. Hence, this IL has a low MP of −81 °C, which is consistent with experimental results.[6,13]

4.4. Insight into the MP Based on Atom Weights. There are 46 unique heavy atoms or groups in the entire data set. The profiles of the top 15 atoms or groups with the highest frequency of occurrence are illustrated in Figure 8a. Alkyl series such as CH_2 occur most frequently. F, C, O, N, and S also frequently appeared in our study. The contribution values of the atoms are normally distributed. Figure 8b shows that most contribution values of atoms are between −100 and 100.

The weight of the contribution of each atom to the MPs of ILs on the entire data set is calculated using eq 2. Table 4 lists the top 10 atomic weights for both positive and negative contributions. NH_2^+, NH_3^+, and S^+ have the highest positive weights, with values of 85.45, 70.22, and 67.22, respectively. ILs have high MPs when their chemical structure contains these positive impact atoms or groups. S^-, N^-, and Br have the biggest negative contributions, with weights of −86.61, −78.76, and −76.11, respectively. The presence of these atoms or groups in the molecular structure of an IL lowers its

Table 3. Atomic Contribution Values to the MP of Imidazolium Hydrogen Sulfate

atom	O	NH	S	O	O^-	OH	CH	CH	CH	NH^+
C_{atom}	8.22	42.54	−53.82	8.22	−55.51	−3.99	−1.19	36.28	−15.43	64.98

https://doi.org/10.1021/acsomega.3c09543
ACS Omega 2024, 9, 16016−16025

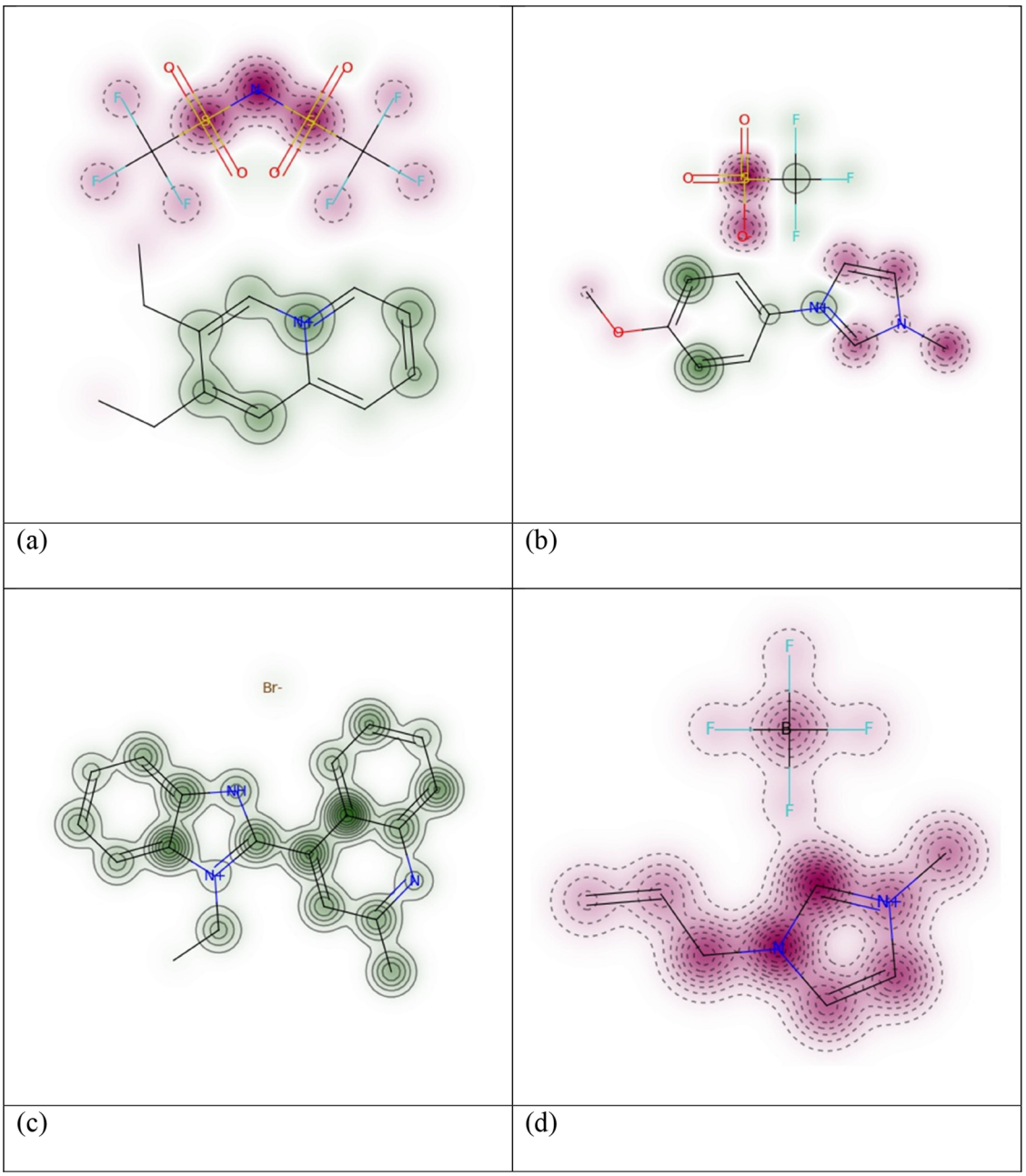
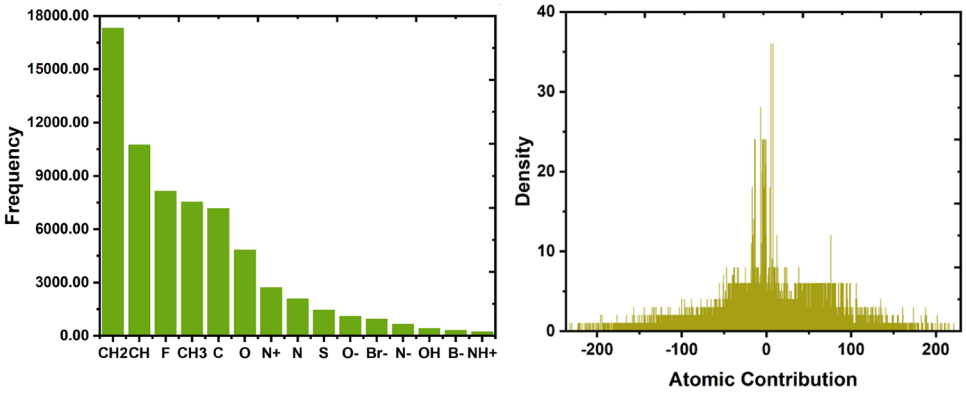


Figure 7. Atomic contribution maps of ILs: (a) IL-1815, (b) IL-272, (c) IL-1218, and (d) IL-384.

MP. Both positive (Table S3) and negative (Table S4) weights of all atoms involved in this work are provided in the Supporting Information.

The visualization results of atom or group contribution weights are displayed in Figure 9. Figure 9a shows the ranking of the atomic weights with positive contributions. Amino groups such as NH_2^+, and NH_3^+ have the top rankings. Thus, amino groups increase the MPs of ILs because they can form hydrogen bonds and polarization functions. This increases molecular interactions and increases the MPs of ILs. S^+, N^+, and P^+ are also favorable for increasing the MPs of ILs when these atoms are positively charged. Figure 9b shows the ranking of the negative contributions. S^-, and N^- are the top two atoms that can lower the MPs of ILs. S and N atoms have a strong impact on MPs. Their presence as positively charged cations plays an essential role in increasing the MPs of ILs. Alternately, their presence as negatively charged anions is the main driving force for reducing the MPs of the ILs. Halogen atoms such as Br are favorable for decreasing the MPs of ILs, which are attributed to the weak intermolecular forces of halogen atoms. The important discovery in this study is that the atoms or groups identified as favorable for increasing or decreasing the MPs of ILs are determined in a data-driven manner. This provides new insight into the melting of ILs from a molecular structure perspective via GNNs, which can aid in the synthesis of task-specific functional ILs. However, our model still has potential predictive limitations as it is based on only 3080 ILs among trillions of available ILs. The generalization ability of the model needs to be further improved. Although the model considers the molecular structure of ILs, it does not consider the interaction between molecules, which

https://doi.org/10.1021/acsomega.3c09543
ACS Omega 2024, 9, 16016−16025



(a) (b)

Figure 8. Profile of atom frequency and contribution values over the entire data set. (a) Distribution of top 15 atoms or groups with the highest frequency of occurrence. (b) Contribution value distribution.

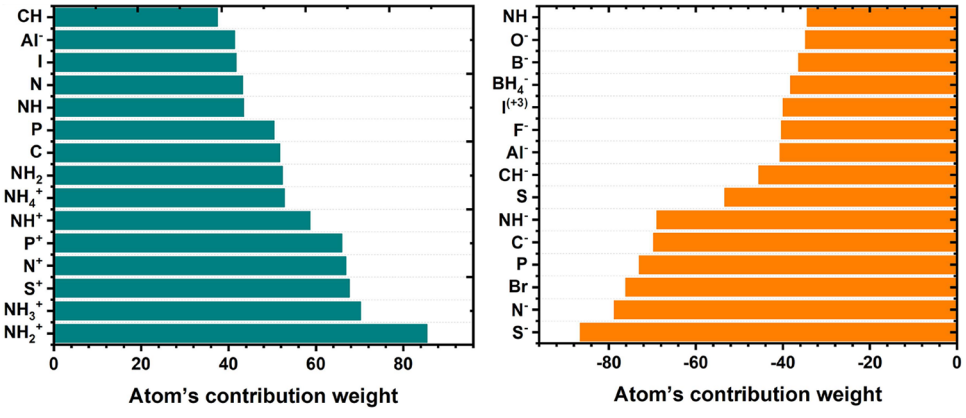
Table 4. Top 10 Atomic Weights for Both Positive and Negative Contributions

positive impact atom	positive weight	negative impact atom	negative weight
NH_2^+	85.45	S^-	−86.61
NH_3^+	70.22	N^-	−78.76
S^+	67.62	Br	−76.11
N^+	66.88	P	−73.00
P^+	65.95	C^-	−69.71
NH^+	58.69	NH^-	−68.98
NH_4^+	52.78	S	−53.36
NH_2	52.27	CH^-	−45.54
C	51.77	Al^-	−40.68
P	50.43	F^-	−40.32

also affects MPs. Thus, GNNs in the future should consider molecular interactions.

5. CONCLUSIONS

Three DBML models (SVM, RF, and MLP) and eight GNN models (GCN, GAT, AttentiveFP, PAGTN, MPNN, DAG, Weave, and GC) were used in this study to predict the MPs of ILs. ECFPs represented molecules as input features in the DBML models, while GNN models operated directly on the molecular graph. Furthermore, a GNN-based interpretability model was established to evaluate the contribution of each atom to the MP of the ILs. GNN models outperformed the DBML models. Among all the methods in this work, the GC model showed the best performance with the lowest RMSE and MAE values of 37.06 and 28.79, respectively, and the highest R^2 value of 0.76. Thus, the capability of GNNs to forecast the MPs of the ILs was established. The atomic contribution was calculated based on the graph representation of GNNs by the interpretability model. Atoms or groups positively contributing to an IL would increase the MP, while atoms or groups negatively contributing to an IL would lower the MP. A high absolute value of the contribution increased the impact. The weight of each atom on the MPs of ILs was calculated and ranked based on the entire data set. Amino groups (NH_2^+ and NH_3^+) occupied the top positive rankings and would significantly increase the MP of the amino-containing IL. The MPs of ILs were lowered by introducing halogen atoms (Br). S, N, and P atoms strongly impacted the MPs of the ILs. The positively charged atoms increased the MPs of ILs, while the negatively charged atoms decreased the MPs of ILs. Thus, ILs could be rapidly screened at room temperature with the aid of our accurate, interpretable GNN



(a) (b)

Figure 9. Distribution of top 15 atomic contribution weights: (a) positive contribution and (b) negative contribution.

https://doi.org/10.1021/acsomega.3c09543
ACS Omega 2024, 9, 16016−16025

models. The findings provide new insights into developing novel task-specific functional ILs.

■ ASSOCIATED CONTENT

Data Availability Statement

All data and codes used in this study are provided in the repository Zenodo: 10.5281/zenodo.10205501

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c09543.

Contribution maps of 3080 types of ionic liquids; evaluation of different data-splitting methods; ionic liquid index and SMILES structures; and positive and negative weights of atoms (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Haijun Feng − *School of Computer Sciences, Shenzhen Institute of Information Technology, Shenzhen, Guangdong 518172, China;* orcid.org/0000-0001-6221-3130; Email: fenghj@sziit.edu.cn

Authors

Lanlan Qin − *School of Chemistry and Chemical Engineering, South China University of Technology, Guangzhou, Guangdong 510640, China*

Bingxuan Zhang − *School of Computer Sciences, Shenzhen Institute of Information Technology, Shenzhen, Guangdong 518172, China*

Jian Zhou − *School of Chemistry and Chemical Engineering, South China University of Technology, Guangzhou, Guangdong 510640, China;* orcid.org/0000-0002-3033-7785

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c09543

Author Contributions

H.F.: Data curation, Project administration, Conceptualization, Methodology, Writing−Original draft preparation. L.Q.: Software, Validation. B.Z.: Investigation, Visualization. J.Z.: Writing−Reviewing and Editing, Supervision.

Funding

This study was funded by the education research and practice project from Shenzhen Institute of Information Technology (No. 2023djpjgyb01), and the Education Science "14th Five-Year Plan" of Shenzhen (No. xbjy23002)

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would also like to thank Dr. Dmitriy Makarov and Dr. Vishwesh Venkatraman for sharing and helping with the data.

■ REFERENCES

(1) Chatel, G.; Rogers, R. D. Oxidation of Lignin Using Ionic liquids−An Innovative Strategy to Produce Renewable Chemicals. *ACS Sustain. Chem. Eng.* **2014**, *2*, 322−339.

(2) Li, W.; Xiao, W.; Luo, Q.; Yan, J.; Zhang, G.; Chen, L.; Sun, J. Ionic Liquids Promoted Synthesis, Enhanced Functions, and Expanded Applications of Porous Organic Frameworks. *Coord. Chem. Rev.* **2023**, *493*, No. 215304.

(3) Watanabe, M.; Thomas, M. L.; Zhang, S.; Ueno, K.; Yasuda, T.; Dokko, K. Application of Ionic Liquids to Energy Storage and Conversion Materials and Devices. *Chem. Rev.* **2017**, *117* (10), 7190−7239.

(4) Plechkova, N. V.; Seddon, K. R. Applications of Ionic Liquids in the Chemical Industry. *Chem. Soc. Rev.* **2008**, *37* (1), 123−150.

(5) Li, Y. C.; Lee, S. Y.; Wang, H.; Jin, F. L.; Park, S. J. Enhanced Electrical Properties and Impact Strength of Phenolic Formaldehyde Resin Using Silanized Graphene and Ionic Liquid. *ACS Omega* **2024**, *9* (1), 294−303.

(6) Zhao, D.; Fei, Z.; Ohlin, C. A.; Laurenczy, G.; Dyson, P. J. Dual-Functionalised Ionic Liquids: Synthesis and Characterisation of Imidazolium Salts with a Nitrile-Functionalised Anion. *Chem. Commun. (Camb)* **2004**, *21*, 2500−2501.

(7) Hough-Troutman, W. L.; Smiglak, M.; Griffin, S.; Matthew Reichert, W. M.; Mirska, I.; Jodynis-Liebert, J.; Adamska, T.; Nawrot, J.; Stasiewicz, M.; Rogers, R. D.; Pernak, J. Ionic Liquids with Dual Biological Function: Sweet and Anti-Microbial, Hydrophobic Quaternary Ammonium-Based Salts. *New J. Chem.* **2009**, *33* (1), 26−33.

(8) Makarov, D. M.; Fadeeva, Y. A.; Safonova, E. A.; Shmukler, L. E. Predictive Modeling of Antibacterial Activity of Ionic Liquids by Machine Learning Methods. *Comput. Biol. Chem.* **2022**, *101*, No. 107775.

(9) Feng, H.; Zhang, P.; Qin, W.; Wang, W.; Wang, H. Estimation of Solubility of Acid Gases in Ionic Liquids Using Different Machine Learning Methods. *J. Mol. Liq.* **2022**, *349*, No. 118413.

(10) Matveieva, M.; Polishchuk, P. Benchmarks for Interpretation of QSAR Models. *J. Cheminform.* **2021**, *13* (1), 41.

(11) Matveieva, M.; Cronin, M. T. D.; Polishchuk, P. Interpretation of QSAR Models: Mining Structural Patterns Taking into Account Molecular Context. *Mol. Inform.* **2019**, *38* (3), No. e1800084.

(12) Polishchuk, P.; Tinkov, O.; Khristova, T.; Ognichenko, L.; Kosinskaya, A.; Varnek, A.; Kuz'min, V. Structural and Physicochemical Interpretation (SPCI) of QSAR Models and Its Comparison with Matched Molecular Pair Analysis. *J. Chem. Inf. Model.* **2016**, *56* (8), 1455−1469.

(13) Varnek, A.; Kireeva, N.; Tetko, I. V.; Baskin, I. I.; Solovev, V. P. Exhaustive QSPR Studies of a Large Diverse Set of Ionic Liquids: How Accurately Can We Predict Melting Points? *J. Chem. Inf. Model.* **2007**, *47* (3), 1111−1122.

(14) Jian, Y.; Wang, Y.; Farimani, A. B. Predicting CO_2 Absorption in Ionic Liquids with Molecular Descriptors and Explainable Graph Neural Networks. *ACS Sustain. Chem. Eng.* **2022**, *10* (50), 16681−16691.

(15) Feng, H.; Zhou, J.; Qian, Y. Atomistic Simulations of the Solid-Liquid Transition of 1-ethyl-3-methyl Imidazolium Bromide Ionic Liquid. *J. Chem. Phys.* **2011**, *135* (14), No. 144501.

(16) Baskin, I.; Epshtein, A.; Ein-Eli, Y. Benchmarking Machine Learning Methods for Modeling Physical Properties of Ionic Liquids. *J. Mol. Liq.* **2022**, *351*, No. 118616.

(17) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9* (2), 513−530.

(18) Makarov, D. M.; Fadeeva, Y. A.; Golubev, V. A.; Kolker, A. M. Designing Deep Eutectic Solvents for Efficient CO_2 Capture: A Data-Driven Screening Approach. *Sep. Purif. Technol.* **2023**, *325*, No. 124614.

(19) Jabed, M. A.; Kim, Y.; Yarbrough, C.; Harman-Ware, A. E.; Olstad, J.; Seiser, R.; Paeper, C.; Starace, A. K.; Kim, S. A Machine Learning Model for Predicting Composition of Catalytic Coprocessing Products from Molecular Beam Mass Spectra. *ACS Sustainable Chem. Eng.* **2023**, *11* (32), 11912−11923.

(20) Mishra, A. K.; Rajput, S.; Karamta, M.; Mukhopadhyay, I. Exploring the Possibility of Machine Learning for Predicting Ionic Conductivity of Solid-State Electrolytes. *ACS Omega* **2023**, *8* (18), 16419−16427.

(21) Feng, H.; Qin, W.; Hu, G.; Wang, H. Intelligent Prediction of Nitrous Oxide Capture in Designable Ionic Liquids. *Appl. Sci.* **2023**, *13* (12), 6900.
(22) Valderrama, J. O.; Faúndez, C. A.; Vicencio, V. J. Artificial Neural Networks and the Melting Temperature of Ionic Liquids. *Ind. Eng. Chem. Res.* **2014**, *53* (25), 10504−10511.
(23) Low, K.; Kobayashi, R.; Izgorodina, E. I. The Effect of Descriptor Choice in Machine Learning Models for Ionic Liquid Melting Point Prediction. *J. Chem. Phys.* **2020**, *153* (10), No. 104101.
(24) Farahani, N.; Gharagheizi, F.; Mirkhani, S. A.; Tumba, K. Ionic Liquids: Prediction of Melting Point by Molecular-Based Model. *Thermochim. Acta* **2012**, *549*, 17−34.
(25) Paduszyński, K.; Kłębowski, K.; Królikowska, M. Predicting Melting Point of Ionic Liquids Using QSPR Approach: Literature Review and New Models. *J. Mol. Liq.* **2021**, *344*, No. 117631.
(26) Mital, D. K.; Nancarrow, P.; Ibrahim, T. H.; Abdel Jabbar, N.; Khamis, M. I. Ionic Liquid Melting Points: Structure−Property Analysis and New Hybrid Group Contribution Model. *Ind. Eng. Chem. Res.* **2022**, *61* (13), 4683−4706.
(27) Makarov, D. M.; Fadeeva, Y. A.; Shmukler, L. E.; Tetko, I. V. Beware of Proper Validation of Models for Ionic Liquids! *J. Mol. Liq.* **2021**, *344*, No. 117722.
(28) Makarov, D. M.; Fadeeva, Y. A.; Shmukler, L. E.; Tetko, I. V. Machine Learning Models for Phase Transition and Decomposition Temperature of Ionic Liquids. *J. Mol. Liq.* **2022**, *366*, No. 120247.
(29) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.
(30) Venkatraman, V.; Evjen, S.; Knuutila, H. K.; Fiksdahl, A.; Alsberg, B. K. Predicting Ionic Liquid Melting Points Using Machine Learning. *J. Mol. Liq.* **2018**, *264*, 318−326.
(31) Venkatraman, V.; Evjen, S.; Lethesh, K. C.; Raj, J. J.; Knuutila, H. K.; Fiksdahl, A. Rapid, Comprehensive Screening of Ionic Liquids Towards Sustainable Applications. *Sustainable Energy Fuels* **2019**, *3* (10), 2798−2808.
(32) Ahmad, W.; Tayara, H.; Chong, K. T. Attention-Based Graph Neural Network for Molecular Solubility Prediction. *ACS Omega* **2023**, *8* (3), 3236−3244.
(33) Rittig, J. G.; Ben Hicham, K. B.; Schweidtmann, A. M.; Dahmen, M.; Mitsos, A. Graph Neural Networks for Temperature-Dependent Activity Coefficient Prediction of Solutes in Ionic Liquids. *Comput. Chem. Eng.* **2023**, *171*, No. 108153.
(34) Martinez-Hernandez, E.; Valencia, D.; Arvizu, C.; Romero Alatorre, D. F.; Aburto, J. Molecular Graph Modularity as a Descriptor for Property Estimation—Application to the Viscosity of Biomass-Derived Molecules. *ACS Sustainable Chem. Eng.* **2021**, *9* (20), 7044−7052.
(35) Atz, K.; Grisoni, F.; Schneider, G. Geometric Deep Learning on Molecular Representations. *Nat. Mach. Intell.* **2021**, *3* (12), 1023−1032.
(36) Wang, Y.; Magar, R.; Liang, C.; Barati Farimani, A. Improving Molecular Contrastive Learning via Faulty Negative Mitigation and Decomposed Fragment Contrast. *J. Chem. Inf. Model.* **2022**, *62* (11), 2713−2725.
(37) Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. Molecular Contrastive Learning of Representations via Graph Neural Networks. *Nat. Mach. Intell.* **2022**, *4* (3), 279−287.
(38) Karamad, M.; Magar, R.; Shi, Y.; Siahrostami, S.; Gates, I. D.; Barati Farimani, A. B. Orbital Graph Convolutional Neural Network for Material Property Prediction. *Phys. Rev. Mater.* **2020**, *4* (9), 93801.
(39) Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug Discovery with Explainable Artificial Intelligence. *Nat. Mach. Intell.* **2020**, *2* (10), 573−584.
(40) Yuan, H.; Yu, H.; Gui, S.; Ji, S. Explainability in Graph Neural Networks: A Taxonomic Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45* (12), 5782−5799.
(41) Wellawatte, G. P.; Gandhi, H. A.; Seshadri, A.; White, A. D. A Perspective on Explanations of Molecular Prediction Models. *J. Chem. Theory Comput.* **2023**, *19* (4), 2149−2160.
(42) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V. *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*; O'Reilly Media, Inc., 2019.
(43) Fan, R.; Chang, K.; Hsieh, C.; Wang, X.; Lin, C. LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.* **2008**, *9* (9), 1871−1874.
(44) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5−32.
(45) Geurts, P.; Ernst, D.; Wehenkel, L. Extremely Randomized Trees. *Mach. Learn.* **2006**, *63* (1), 3−42.
(46) Hinton, G. E. Connectionist Learning Procedures. *Mach. learn.* **1990**, 555−610.
(47) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(48) Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J. API Design for Machine Learning Software: Experiences from the Scikit-learn Project. *Preprint arXiv:1309.0238* **2013**, DOI: 10.48550/arXiv.1309.0238.
(49) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Adv. Neural Inf. Process* **2015**, *2*, 2224−2232.
(50) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30* (6), 595−608.
(51) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *Preprint arXiv:1609.02907* **2016**, DOI: 10.48550/arXiv.1609.02907.
(52) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph Attention Networks. *Preprint arXiv:1710.10903* **2017**, DOI: 10.48550/arXiv.1710.10903.
(53) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* **2019**, *63* (20), 8749−8760.
(54) Chen, B.; Barzilay, R.; Jaakkola, T.Path-Augmented Graph Transformer Network. *arXiv Preprint arXiv:1905.12712* **2019**
(55) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *Proc. Mach. Learn. Res.* **2017**, *70*, 1263−1272.
(56) Lusci, A.; Pollastri, G.; Baldi, P. Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-like Molecules. *J. Chem. Inf. Model.* **2013**, *53* (6), 1563−1575.
(57) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887−2893.
(58) Riniker, S.; Landrum, G. A. Similarity Maps-A Visualization Strategy for Molecular Fingerprints and Machine-Learning Methods. *J. Cheminformatics* **2013**, *5* (1), 1−7.

https://doi.org/10.1021/acsomega.3c09543
ACS Omega 2024, 9, 16016−16025