基于图卷积的离子液体 CO_2 溶解度可解释性预测

张茜茜[1,2],陈　平[1,2]
(1. 中北大学 信息与通信工程学院,山西 太原 030051;
2. 信息探测与处理山西省重点实验室(中北大学),山西 太原 030051)

摘　要:为构建离子液体的 CO_2 溶解度的准确预测模型,考虑到传统模型存在的描述符计算复杂、成本高、关联结构与性质困难、结构特征提取不充分等问题,提出一种融合了加入注意力机制的图卷积神经网络和 XGBoost 的预测模型(APGCN-XGBoost)。对 9 897 组离子液体的 CO_2 溶解度数据的分析结果显示,所提出的 APGCN-XGBoost 模型在预测性能上优于传统的分子指纹模型和图卷积神经网络模型。此外,通过注意力池化层与 SHAP 方法对模型进行解释,APGCN-XGBoost 模型学习到了离子液体中各个原子和结构的特征信息与分子非局部信息,这些特征信息不仅可以用于性质预测,还可以用于探索化学结构与性质之间的联系,即通过模型的解释,筛选出对于溶解度预测重要的离子液体结构信息,从而实现 CO_2 捕获过程中理想离子液体的计算机辅助设计和筛选。

关键词:图卷积神经网络;离子液体;性质预测;溶解度;可解释性
中图分类号:TP181　　**文献标识码**:A　　**文章编号**:1673-629X(2024)02-0134-08
doi:10.3969/j.issn.1673-629X.2024.02.020

Interpretable Prediction of CO_2 Solubility of Ionic Liquids Based on Graph Convolution Neural Network

ZHANG Qian-qian[1,2], CHEN Ping[1,2]
(1. School of Information and Communication Engineering, North University of China, Taiyuan 030051, China;
2. Shanxi Province Key Laboratory of Information Detection and Processing, North University of China, Taiyuan 030051, China)

Abstract: In order to build an accurate prediction model of CO_2 solubility of ionic liquids, considering the problems existing in traditional models, such as complex descriptor calculation, high cost, difficult related structure and properties, and insufficient structural feature extraction, a prediction model APGCN-XGBoost combining graph convolution neural network and XGBoost with attention mechanism was proposed. The analysis results of CO_2 solubility data of 9 897 groups of ionic liquids show that the prediction performance of the proposed APGCN-XGBoost model is better than that of the traditional molecular fingerprint model and graph convolutional neural network model. In addition, the APGCN-XGBoost model learned the characteristic information and molecular non-local information of each atom and structure in ionic liquids, which can be used not only to predict the properties, but also to explore the relationship between the chemical structures and properties, that is, to screen out the structural information of ionic liquids that is important for solubility prediction through the interpretation of the model, thus realizing the computer-aided design and screening of ideal ionic liquids in the process of CO_2 capture.

Key words: graph convolutional neural networks; ionic liquids; property prediction; solubility; interpretability

0　引　言

离子液体是一种理想的绿色 CO_2 吸收剂[1-2],具有化学稳定性好、腐蚀性低、从中分离 CO_2 需要更少的能量、可供选择的种类比较多等优点[3-5]。CO_2 在离子液体中的溶解度值是重要信息,可以为筛选、设计满足工业需求的离子液体提供重要指导[6]。目前,获取离子液体中 CO_2 溶解度的主要方法包括实验测量和建模。而通过实验测量大量离子液体的 CO_2 溶解度并不现实,实验过程不仅耗时费力且成本高,故借助模型实现准确的预测具有重要意义。基于热力学或动力学方

收稿日期:2023-04-03　　**修回日期**:2023-08-04
基金项目:国家自然科学基金(62122070);山西省研究生创新项目(2022Y623)
作者简介:张茜茜(1998-),女,硕士研究生,CCF 会员(P3570G),研究方向为图神经网络;通信作者:陈　平(1983-),男,博士,教授,博导,CCF 会员(26644M),研究方向为人工智能、图像处理与重建。

法构建的模型难以处理大量的性质数据,同时预测准确度不够高、通用性差。因此,根据现有的实验数据开发机器学习模型实现准确、快速、高效的离子液体的 CO_2 溶解度预测是十分必要的。

构建预测模型过程中离子液体的表征方式选取尤为重要,当前的模型构建选择的分子表征方式主要包括经验描述符、简单热物理参数、基团信息以及包含多种物理化学性质值、拓扑性质的分子描述符。

经验描述符需要化学背景以及相关先验知识。简单热物理参数多借助现有的数据集,数据集中包含离子液体的偏心因子、临界性质(临界温度与临界压力)以及分子量等。Moosanezhad – Kermani H 等采用 GMDH 分组数据处理方法,提出了一种数学关联形式的白盒模型,模型输入参数为传统的热力学参数[7]。基团信息是指将离子液体手动分解为若干基团的组合,Song Z 等使用基团信息表征离子液体,使用人工神经网络(Artificial Neural Network,ANN)和支持向量机(Support Vector Machines,SVM)构建模型预测离子液体中 CO_2 溶解度[8]。这两种表征方式都受限于数据集,对新的离子液体应用困难。Aghaie M 等引入了几何、拓扑、信息指数、基于 3D 矩阵的描述符和量子化学特征等作为输入参数,用决策树(Decision Tree,DT)与随机森林(Random Forest,RF),并与热力学参数模型进行对比,结果证明相比于热力学参数,结构性质与离子液体的 CO_2 溶解度有更强的相关性[9]。然而大量的分子描述符需要借助专业的软件获取,计算成本高。分子指纹作为容易获取的表征方式,常见于药物分子的虚拟筛选模型中,分子指纹是由结构信息编码而成的高维、稀疏向量,这一过程会造成分子的结构信息表征能力受限,且高维的向量在模型训练中易过拟合。

而图卷积神经网络(Graph Convolutional Neural networks,GCN)进行离子液体的结构特征学习可以解决上述分子表征方式存在的问题。GCN 可以从分子图中捕获结构特征信息,已经被应用于生物化学领域的活性分析、性能预测等任务中[10],在这些领域的成功应用证明 GCN 能够有效学习分子图的结构数据,挖掘化合物分子图中的结构信息实现与分子性质的关联。这些任务与离子液体中 CO_2 溶解度预测任务的共性是要实现结构性质与性质的关联,说明通过 GCN 实现离子液体的预测任务同样是可行的。

该文以离子液体中的 CO_2 溶解度预测为研究目的,提出了基于注意力图卷积神经网络的全流程建模方法,主要贡献概括如下:

(1)在图嵌入模块引入注意力机制,使模型关注原子的非局部特征信息。

(2)将图卷积神经网络与 XGBoost 有效组合为预测模型,解决了预测模型容易过拟合的问题。

(3)借助注意力池化层与 SHAP 对预测模型进行解释,能够筛选出对于 CO_2 溶解度预测重要的离子液体结构片段。

1　CO_2 溶解度预测模型

1.1　总体框架流程

整体流程见图 1。首先将离子液体的分子式转化为 GCN 所需要的图数据,原子的特征向量经过图卷积后生成新的原子特征向量,然后经过池化转换为整个分子图的特征向量,也就是图嵌入网络生成的离子液体的图嵌入向量。

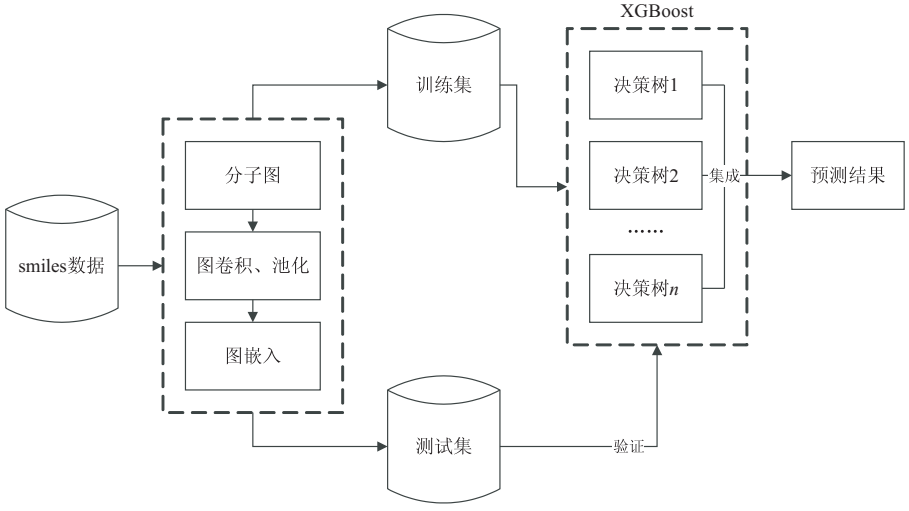


图 1　二氧化碳溶解度预测模型构建及应用流程

1.2　图数据获取

要对分子进行表征,通过软件 ChemDraw[11](大多数化学结构软件都支持)获取阳离子、阴离子的简化分子线性输入规范(Simplified Molecular Input Line

Entry System,SMILES)。SMILES 字符串是一种类似语言的描述,可以通过利用化学信息学工具方便地表达分子二维和三维结构特征,从而用于表征分子特征。然后使用 Python 中的化学信息学工具包 RDKit 将 SMILES 字符串转换为所需要的图数据与 Morgan 分子指纹。RDKit 获取的离子液体的 SMILES 表达式所包含的图数据包含节点特征(如原子的元素类型、价态等)与结构特征(如键的种类、键是否在环中等)。

1.3 图卷积

训练过程旨在学习这些特征来获得分子的整体特征信息,最终表示为一个固定长度的向量[12]。假设离子液体中包含 n 个原子,包含所有原子的特征向量组成的 $n \times d$ 维特征矩阵 $\boldsymbol{X}$,以及原子之间的 $n \times n$ 维的邻接矩阵 $\boldsymbol{A}$,$\boldsymbol{X}$ 与 $\boldsymbol{A}$ 也就是 GCN 模型的输入,$\boldsymbol{H}^{(l)}$ 表示所有原子在第 l 层的特征向量矩阵,$\boldsymbol{H}^{(l+1)}$ 表示所有原子经过一次卷积操作之后的特征向量矩阵,GCN 层与层之间的传播方式如式 1 所示:

$$\boldsymbol{H}^{(l+1)} = \sigma(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}) \tag{1}$$

其中,$\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$,$\boldsymbol{I}$ 为单位矩阵,$\tilde{D} = \sum \tilde{A}_{ij}$,$\boldsymbol{H}^{(0)} = \boldsymbol{X}$,$\sigma$ 为非线性激活函数,$\boldsymbol{W}^{(l)}$ 表示当前层卷积变换的可训练的参数矩阵[13]。

在每一层中,将原子的特征与邻域原子的特征信息聚合,生成新的原子表示,对原子的特征向量进行更新,作为下一层的输入。随着网络层数的增加,逐渐获取所有的原子特征信息与结构信息生成新的原子特征向量。

1.4 注意力池化层

为了实现性质预测任务,希望从离子液体的分子图生成图级表示,大多数先前的研究用简单的最大值池化、平均池化等[14-16]来聚合分子图中所有的原子特征来生成整个分子图的表示,但可能会导致不必要的合并、不完整的解释等问题,因此使用注意力池化层作为传统池化方式的替代。目标是将分子图编码成固定大小的嵌入矩阵,同时最大化原子特征向量所隐含的信息,学习原子的全局信息。

注意力机制允许方法关注神经网络的任务相关部分。对具有序列结构数据的任务应用注意力机制已成为一种惯例,以使模型能够集中注意输入的最相关部分并实现更好的预测[17]。

通过注意力池化层生成注意力系数矩阵,使用注意力机制,将从图卷积模块学习的原子特征向量作为输入来输出权重向量。注意力系数也就是输出的权重,代表不同特征的重要性,权重越大,特征越重要,根据注意力系数大小筛选出对于溶解度预测相对重要的特征,这里的特征是分子图中的化学结构片段,经过图卷积模块的卷积、合并,每个原子的特征向量不仅包含原子本身的信息,还包含原子的邻域信息。

经过图卷积模块,离子液体的结构特征被编码为每个原子的特征向量组合,此时的图表示为 $\boldsymbol{G}$,$\boldsymbol{\alpha}_{vn}$ 为原子的特征向量,$\boldsymbol{G}$ 的每一行表示一个原子的特征向量,如下式:

$$\boldsymbol{G} = (\boldsymbol{\alpha}_{v1}, \boldsymbol{\alpha}_{v2}, \cdots, \boldsymbol{\alpha}_{vn}) \tag{2}$$

通过注意力池化:

$$\boldsymbol{A} = \mathrm{softmax}(\boldsymbol{\beta}_2 \tanh(\boldsymbol{\beta}_1 \boldsymbol{G}^T)) \tag{3}$$

其中,$\boldsymbol{\beta}_1$,$\boldsymbol{\beta}_2$ 表示权重矩阵,则最终输出的图的表达形式变为:

$$\boldsymbol{G}' = A\boldsymbol{G} \tag{4}$$

加入注意力池化层的图卷积神经网络(Attention Pool Artificial Neural Network,APGCN)所构成的图嵌入生成模块流程如图 2 所示,通过图卷积神经网络获取分子图中每个原子的特征向量。

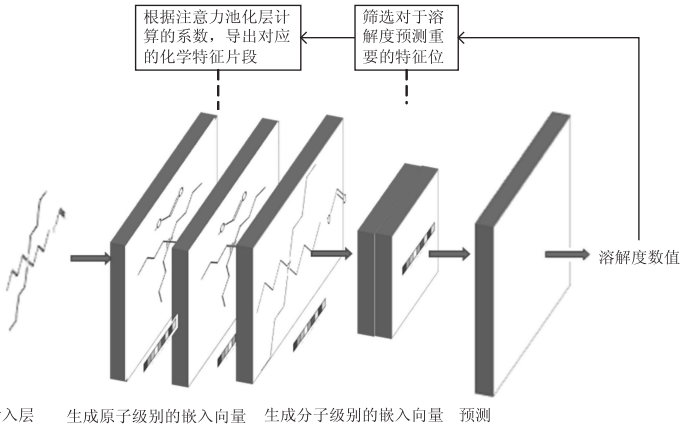


图 2 APGCN 模块示意图

1.5 预测模块算法选择

目前,离子液体的 CO_2 溶解度预测模型构建中 ANN,SVM,RT 模型等被广泛应用。在输入数据为分子描述符、高维向量时机器学习算法综合表现更好,训练效率高、不易过拟合[18]。通过 6 折交叉验证选择最合适的机器学习算法进行建模,在数据集上对机器学

习算法 XGBoost(e-Xtreme Gradient Boosting)[19]、梯度提升决策树(Gradient Boosting Decision Tree, GBDT)[20]、LightGBM(Light Gradient Boosting Machine)[21]进行 6 折交叉验证实验对比,评估指标选择决定系数(R^2)、平均绝对误差(MAE)、均方误差(MSE)。表 1 中结果表明,经过验证,在选定的数据集上使用 XGBoost 构建的模型表现更好,更能避免过拟合,且相比于其他机器学习算法有较高的准确性,同时计算成本低、效率高,故选择 XGBoost 算法来构建 CO_2 的预测模型。

表 1　图嵌入和 Morgan 指纹交叉验证对比

算法	图嵌入			Morgan		
	R^2	MAE	MSE	R^2	MAE	MSE
XGBoost	0.996 2	0.009 7	0.000 2	0.994 6	0.012 0	0.000 3
LightGBM	0.990 6	0.015 3	0.000 5	0.989 9	0.015 8	0.000 6
GBDT	0.976 6	0.025 0	0.001 3	0.974 8	0.026 7	0.001 4

此预测任务的输入不仅有图数据,还有对于离子液体的 CO_2 溶解度不可忽略的重要热力学参数 T,P,因此模型要实现基于图数据与数值型混合数据的预测。综上所述,充分考虑到离子液体的 CO_2 预测模型需要输入数据的特点,经过图嵌入生成网络将图数据转化为一个低维向量,将其与对应的 T,P 条件结合作为 XGBoost 预测模型的输入,输出最终的 CO_2 溶解度预测值。

XGBoost 是基于决策树的集成模型,每次训练时对数据集进行采样,随机抽取一定数量的样本作为每棵树的根节点。每个样本所包含的分子指纹所表示的离子液体结构特征、温度 T、压力 P 即为待分裂的特征。每一棵树生成过程如图 3 所示,温度 T、压力 P、结构特征都是影响离子液体 CO_2 溶解度的重要因素,从根节点开始分裂,图 3 中的树根节点先对压力 P 进行判断,中间节点有温度 T、分子的结构特征向量,对于 APGCN-XGBoost 模型来说,分子的结构特征向量为 APGCN 生成的图嵌入向量,对于 APGCN-XGBoost 模型来说,分子的结构特征向量为 Morgan 分子指纹,中间节点不断进行分裂生长,直到叶子节点结束。

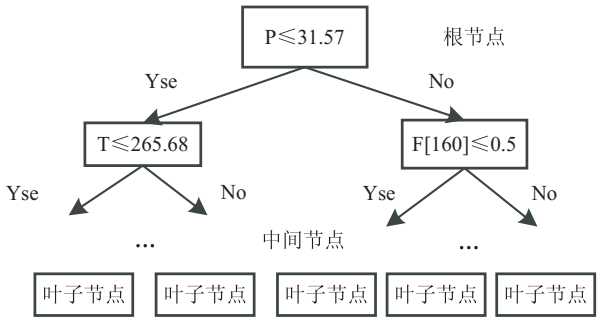


图 3　决策树生成

XGBoost 将增益作为特征选择和分裂点选择的准则,增益值越大,说明分裂后能使目标函数减小越多,就越好,用增益值来寻找每次分裂的最优特征和最优分裂点。通过不断地进行特征分裂来生成一棵树,每次添加一棵树,其实是学习一个新函数,去拟合上次预测的残差。在目标函数中加入了正则项控制模型复杂度,以防止过拟合,并对目标函数进行二阶泰勒展开,保留了更多有关目标函数的信息。最终得到 XGBoost 的最优目标函数如式 5 所示,

$$\mathrm{Obj} = -0.5\sum_{j=1}^{T} G_j^2/(H_j+\lambda) + \gamma T \tag{5}$$

其中,G_j,H_j 分别为损失函数对前 $t-1$ 轮迭代的一阶导数和二阶导数和,γ 和 λ 为加权因子,ω_j 为叶子节点权重,T 为叶子的个数。

当训练完成后得到了多棵树,对一个样本进行预测时,每棵树的节点对相应的离子液体的分子指纹和温度 T、压力 P 进行判断,最终每棵树都会落到对应的一个叶子节点,模型最终预测的 CO_2 溶解度为每棵树对应的叶子节点之和。

2　实验及结果分析

2.1　实验环境及参数设置

设计 APGCN 为 4 层、最终生成的图嵌入维度为 50 的向量构建 APGCN-XGBoost 预测模型进行实验。选择半径为 8、阳离子和阴离子指纹长度均为 512 的 Morgan 分子指纹[22-23]来构建 Morgan-XGBoost 预测模型进行对比实验,RDKit 将 SMILES 转化 Morgan 分子指纹,也就是指定长度的二进制比特串。分子结构、T、P 对于离子液体的 CO_2 溶解度均为重要的因素,故最终不管是哪种分子表示方式,都需要结合相应的 T、P 作为预测模型的输入。

超参数优化使用了优化工具 Optuna,对基于树的超参数搜索进行了优化,它使用被称为 TPE("Tree-structured Parzen Estimator")Sampler 的方法,这种方法依靠贝叶斯概率来确定哪些超参数选择是最有希望的并迭代调整搜索[24]。反复调用目标函数,找到使目标函数最小化的值。最终优化后模型的超参数见表 2。

表 2 模型的主要超参数

超参数	GCN/APGCN-XGBoost	Morgan-XGBoost
max_depth	10	15
learning_rate	0.1	0.1
n_estimators	200	200
min_child_weight	5	1

2.2 实验环境及参数设置

使用的数据来源于 Lei 等早期的工作[25],其中包含在不同温度和压力下多种离子液体的 CO_2 溶解度数据。共有 118 种离子液体 9 897 组 CO_2 溶解度数(0.000 064 8-0.951 6 mol · mol^{-1}),温度 243.2-453.15 K,压力 0.798 bar-49 990 kPa。阳离子有[BMIM],[EMIM],[MMIM],[HMIM][OMIM],[BMPYR],[N1,4,4,4],[P1,4,4,4]等,阴离子有[BF_4][Cl],[DCA],[NO_3],[PF_6],[SCN],[C(CN)$_3$],[Tf_2N],[$MeSO_4$]等。将数据集划分出外部测试集,包含 8 种离子液体的 1 040 组溶解度数据,这 8 种离子液体未被包含在训练集中。将训练集中的实验数据随机划分出 20% 作为测试集,测试集与外部测试集中的数据均不参与模型的训练。

2.3 评价指标

对于预测任务,模型的性能可以用预测的准确度来评估。选择平均相对误差(Average Absolute Relative Deviation,AARD)与确定系数 R^2 作为模型评估指标,$y(i)$ 与 $\hat{y}(i)$ 分别为溶解度的实验值和模型的预测值。AARD 与 R^2 越大说明预测值与实验值越接近。

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\hat{y}(i) - y(i))^2}{\sum_{i=1}^{N} (\hat{y}(i) - \frac{1}{N}\sum_{i=1}^{N} y(i))^2} \tag{6}$$

$$\mathrm{AARD} = \frac{100}{N}\sum_{i=1}^{N} \frac{|\hat{y}(i) - y(i)|}{y(i)} \tag{7}$$

2.4 对比分析

不同模型的相关参数与训练集和测试集的预测性能如表 3 所示。其中 Daryayehsalameh B 仅仅对[BMIM][BF_4]中 CO_2 溶解度预测建模[26],而 APGCN-XGBoost 模型可以对多种离子液体实现更准确的预测。

表 3 模型的预测结果

Model	Inputs	Datasets	AARD/%	R^2	Ref
DT	Pc,Tc,Mw,ω,T,P	Train	2.85	0.965	[9]
		Test	21.24	0.94	
RF		Train	4.72	0.988	
		Test	12.05	0.96	
DT	Chi_G/D 3D, Disps,SpMax_RG,Homo-Lumo Fraction, T,P	Train	0.53	0.999	
		Test	20.26	0.958	
RF		Train	5.89	0.997	
		Test	12.36	0.983	
CFNN	T,P	Train	6.85	0.989 18	[25]
		Test	7.01	0.984 08	
XGBoost	Morgan 指纹, T,P	Train	7.445 3	0.997 3	
		Test	10.071 2	0.986 7	
XGBoost	GCN 图嵌入, T,P	Train	5.127 3	0.998 6	
		Test	9.502 6	0.987 6	
XGBoost	APGCN 图嵌入, T,P	Train	4.442 8	0.999 0	
		Test	9.840 5	0.988 3	

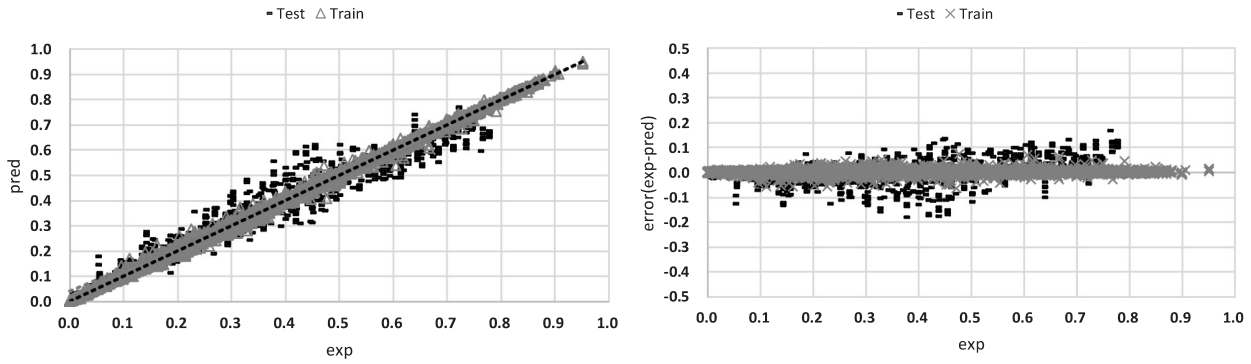
由预测结果的对比可以看出,APGCN 获取的图嵌入构建的模型测试集的预测结果与实验值的 AARD 最小,R^2 最大,能够实现离子液体中 CO_2 溶解度更为准确的预测。APGCN-XGBoost 与 Morgan-XGBoost,GCN-XGBoost 预测模块均采用 XGBoost 模型,三个模型外部测试集决定系数 R^2 分别为 0.943 3,0.729 9,0.924 1,外部测试集的结果说明 APGCN-XGBoost 模型对于未知的离子液体也能实现更为准确的预测。

APGCN-XGBoost 模型不仅在预测准确度方面表现优秀,同时相比于量子化学、分子轨道、拓扑信息、基团信息等描述符的计算,APGCN 获取的分子图嵌入无需复杂的计算、专业的软件或者物理化学先验知识,特征工程成本低,易于推广到新的离子液体。综合表明 APGCN 实现了更好的分子表征,注意力池化层的

加入提高了离子液体结构信息的提取效率,提高了预测模型的准确性、泛化性。

训练集与测试集的实验值与 APGCN-XGBoost 模型预测值进行比较,图 4(a)实验值与预测值的拟合效果表明模型能够实现较为准确的预测;图 4(b)可以看出测试集大多数误差落在[-0.1,0.1]之间,表明模型的预测值与实验值具有较高的对应性。

实验结果验证了模型的有效性,证明模型可以很好地提取离子液体中的特征信息用于 CO_2 溶解度预测,不仅在准确度上具有明显的优势,且通过图卷积获取的图嵌入维度更低,这能够降低下游的计算量,为预测 CO_2 在离子液体中的溶解度提供了一种便捷、准确的方法。

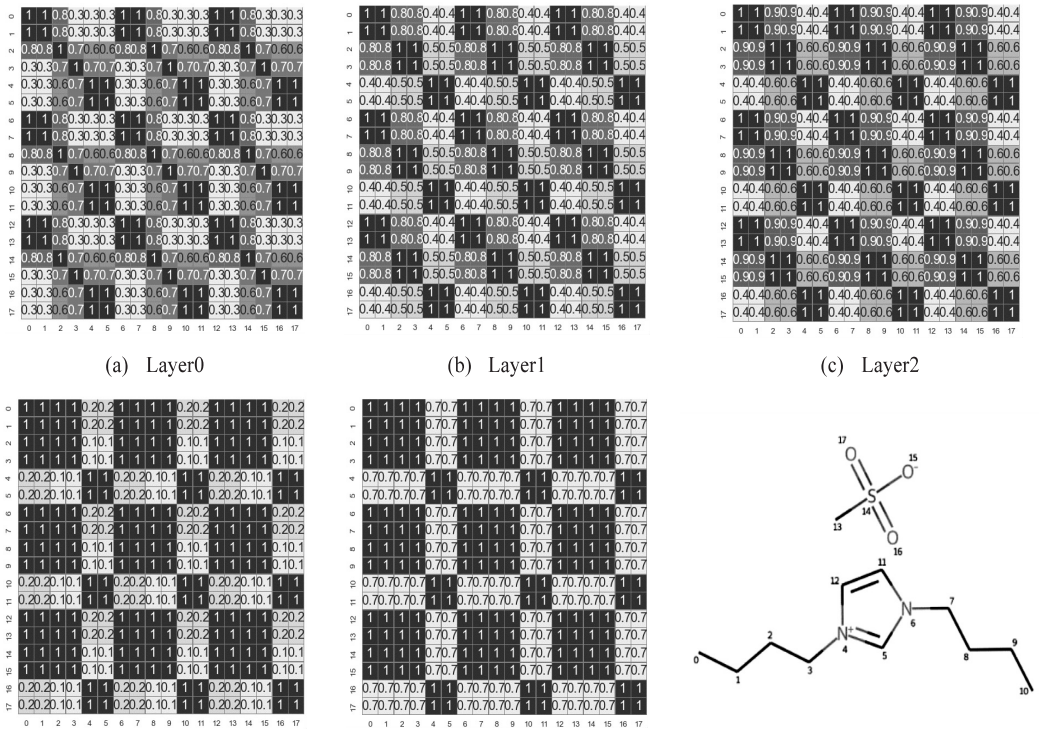


(a) Train, Test 的预测值-实验值　　(b) Train, Test 的预测误差-实验值

图 4　实验值和 APGCN-XGBoost 模型预测的 CO_2 溶解度值比较

2.5　模型学习效果

原子的特征向量在学习过程中不断变化,可通过计算不同层中不同原子的特征向量之间的 Pearson 相关系数(取值为-1 到 1 之间,接近-1 或 1 被称为具有强负相关或正相关),并绘制原子相似性矩阵的热图,以观察学习过程中的变化。以[BBIM][MeSO3]为例,原子的标号为 0-17,第 0-4 层的原子相关性热图见图 5,具体标号对应位置如图 5(f)所示。在训练之前,如图 5(a),不同原子之间相似性混乱,并没有明显的区分,而在训练过程中,具有相似化学特性的原子的特征向量越来越相似,最终整个离子液体按照原子的化学环境分为几个区域,每个区域内的原子的化学特性相似,如图 5(e)所示,原子 0-3 在同一支链上,具有相似的节点特征且随着 APGCN 层数的增加,几个原子特征向量之间的相关系数增大,最终为 1,也就是说原子特征向量之间相似性也随之提高。



(a) Layer0　(b) Layer1　(c) Layer2

(d) Layer3　(e) Layer4　(f) [BBIM][MeSO3]分子图

图 5　[BBIM][MeSO3]的热图

具有相似的节点特征并且在拓扑上彼此接近的原子在最终训练后的嵌入空间中具有更高的相似性。说明 APGCN 通过学习成功提取了原子化学特性相关的信息,并且能够捕获到原子的非局部效应。

2.6 模型的可解释性

模型不仅能够进行准确的预测,还能对模型的预测机理进行解释,实现结构-性质之间的关联,借助注意力池化层计算的注意力系数与 SHAP(SHapley Additive exPlanation)[27] 实现模型的解释。首先通过分析温度、压力对模型溶解度预测的影响,验证模型中特征重要性的可信度。SHAP 的部分依赖图(partial dependence plot,pdp)可以显示特征与预测值之间的关系。如图 6 所示,pdp 图显示了温度、压力的不同取值与模型预测的溶解度数值之间的关系。可以看出,离子液体中的 CO_2 溶解度随着温度的升高而降低,随着压力的增加而增加,结果与溶解度随着温度和压力变化的理论是一致的,说明模型的预测和解释是可信的。

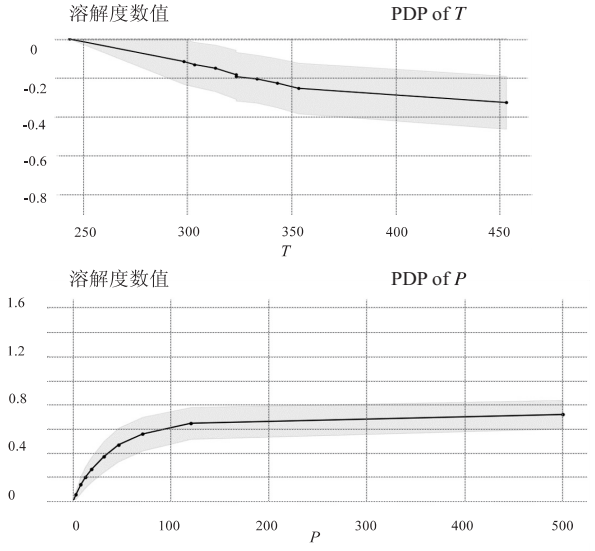


图 6 温度和压力的 pdp

用注意力池化层得到的注意力系数来衡量离子液体中结构片段对于溶解度预测的重要性,最终导出对溶解度预测影响较大的结构信息,并将其与已知的结论进行对比。相应的结构信息见图 7,分子图中的浅色部分为标注的重要结构片段。从结果可以看出:标注结构均为阴离子,S=O 基团与含氟阴离子对溶解度影响较大。已知:离子液体 CO_2 溶解度中阴离子起关键作用[28];S=O 基团能够提高离子液体 CO_2 溶解度[29];阴阳离子氟化都能使溶解度升高,阴离子氟化作用更为明显[30]。通过对离子液体 CO_2 溶解度重要结构的可视化,寻找溶解度预测中重要的子结构。将化学常识与模型自动筛选得到的重要特征对比,有利于进一步研究结构与性质的关系,对碳捕集过程中的理想离子液体的筛选与设计具有重要意义。

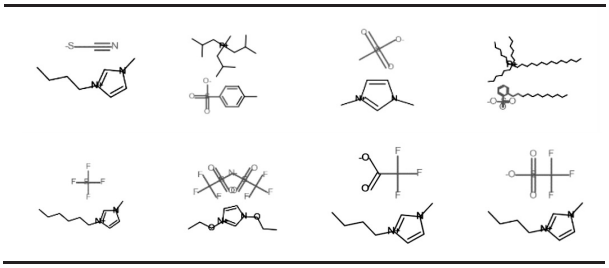


图 7 对模型预测贡献较大的结构片段

从模型的结果可以看出,APGCN 能够学习到离子液体分子图中更全局的特征,基于 APGCN 的离子液体表征方式足够“学习”化学结构与性质之间的关系,并且 APGCN-XGBoost 模型的开发和使用不依赖于化学反应、化学性质和描述符计算的任何高级先验知识,准确预测的同时还可以指出哪些化学结构与化学特性最相关,可以将这些结构筛选出来,此项工作旨在通过深度学习来解释结构与性质的关联,作为相关离子液体的设计和筛选的辅助工具。这种方法还可以很容易地扩展到许多其他领域,用于多种化学分子的性质预测以及筛选、结构优化工作。

3 结束语

为实现离子液体中 CO_2 溶解度的高效预测,通过图卷积神经网络从离子液体的分子图中捕获有效的特征信息,将 APGCN 与 XGBoost 组成高效的预测模型,图数据经过 APGCN 转化为包含结构特征的低维向量,结合其他参数数据作为输入特征,APGCN-XGBoost 模型输出溶解度的预测值。最终结果表明,对于新的离子液体,APGCN 能够实现比其他分子表征方式更为准确的预测,且对比于传统的热力学、量子化学、拓扑信息等表征方式特征工程成本低,更容易推广到新的离子液体,说明 APGCN 实现了更好的分子表征。一方面,APGCN-XGBoost 模型能够实现在较宽的温度、压力范围内离子液体 CO_2 溶解度准确的预测,解决了实验测量耗时、成本高、传统模型结构复杂、通用性差、计算成本高等问题,为预测 CO_2 在离子液体中的溶解度提供了一种高效的方法。另一方面,APGCN 成功学习到了各个原子和结构的特征信息,这些特征信息不仅可以用于预测离子液体的性质,还可以用于探索不同的原子和结构是如何影响最终的性质预测的,有利于进一步研究结构与性质之间的关系。

参考文献:

[1] MAO J,IOCOZZIA J,HUANG J,et al. Graphene aerogels for efficient energy storage and conversion[J]. Energy & Environmental Science,2018,11(4):772-799.

[2] ZENG S,ZHANG X,BAI L,et al. Ionic-liquid-based CO2

capture systems:structure,interaction and process[J]. Chemical Reviews,2017,117(14):9625-9673.

[3] HALLETT J P,WELTON T. Room-temperature ionic liquids:solvents for synthesis and catalysis. 2[J]. Chemical reviews,2011,111(5):3508-3576.

[4] CHONG F K,ANDIAPPAN V,NG D K,et al. Design of ionic liquid as carbon capture solvent for a bioenergy system:integration of bioenergy and carbon capture systems[J]. ACS Sustainable Chemistry & Engineering,2017,5(6):5241-5252.

[5] WANG J,SONG Z,CHENG H,et al. Computer-aided design of ionic liquids as absorbent for gas separation exemplified by CO2 capture cases[J]. ACS Sustainable Chemistry & Engineering,2018,6(9):12025-12035.

[6] 王　欢,吴云雁,赵燕飞,等. 离子液体介导 CO2 化学转化研究进展[J]. 物理化学学报,2021,37(5):157-168.

[7] MOOSANEZHAD-KERMANI H,REZAEI F,HEMMATI-SARAPARDEH A,et al. Modeling of carbon dioxide solubility in ionic liquids based on group method of data handling[J]. Engineering Applications of Computational Fluid Mechanics,2021,15(1):23-42.

[8] SONG Z,SHI H,ZHANG X. Prediction of CO2 solubility in ionic liquids using machine learning methods[J]. Chemical Engineering Science,2020,223:115752.

[9] AGHAIE M,ZENDEHBOUDI S. Estimation of CO2 solubility in ionic liquids using connectionist tools based on thermodynamic and structural characteristics[J]. Fuel,2020,279:117984.

[10] 江钰哲,成　全. 图嵌入式双层图卷积网络药物推荐模型[J/OL]. 计算机工程与应用. https://kns.cnki.net/kcms/detail//11.2127.TP.20230228.1650.038.html.

[11] COUSINS K R. ChemDraw ultra 9.0[J]. Journal of the American Chemical Society,2005,127(11):4115-4116.

[12] DUVENAUD D,MACLAURIN D,AGUILERA-IPARRA J,et al. Convolutional networks on graphs for learning molecular fingerprints[C]//Proceedings of the 28th international conference on NeurIPS. Montreal:MIT Press,2015:2224-2232.

[13] KIPF T N,WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv:1609.02907,2016.

[14] HENAFF M,BRUNA J,LECUN Y. Deep convolutional networks on graph-structured data[J]. arXiv:1506.05163,2015.

[15] ZHANG M,CUI Z,NEUMANN M,et al. An end-to-end deep learning architecture for graph classification[C]//Proceedings of the AAAI conference on artificial intelligence. New Orleans:AAAI,2018:4438-4445.

[16] DEFFERRARD M,BRESSON X,VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[C]//Proceedings of the 30th international conference on NeurIPS. Barcelona:Curran Associates Inc.,2017:3844-3852.

[17] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]//Proceedings of the 31st international conference on NeurIPS. Long Beach:Curran Associates Inc.,2017:6000-6010.

[18] 聂长森,白　勇,柳贤德. 基于机器学习的 COX 抑制剂预测模型研究[J]. 计算机技术与发展,2017,27(10):74-77.

[19] CHEN T,GUESTRIN C. XGBoost:a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco:ACM,2016:785-794.

[20] FRIEDMAN J H. Greedy function approximation:a gradient boosting machine[J]. Annals of Statistics,2001,29(5):1189-1232.

[21] KE G,MENG Q,FINLEY T,et al. Lightgbm:a highly efficient gradient boosting decision tree[C]//Proceedings of the 31st international conference on NeurIPS. Long Beach:Curran Associates Inc.,2017:3149-3157.

[22] MORGAN H L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service[J]. Journal of Chemical Documentation,1965,5(2):107-113.

[23] ROGERS D,HAHN M. Extended-connectivity fingerprints[J]. Journal of Chemical Information and Modeling,2010,50(5):742-754.

[24] AKIBA T,SANO S,YANASE T,et al. Optuna:a next-generation hyperparameter optimization framework[C]//Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. Anchorage:Association for Computing Machinery,2019:2623-2631.

[25] LEI Z,DAI C,CHEN B. Gas solubility in ionic liquids[J]. Chemical Reviews,2014,114(2):1289-1326.

[26] DARYAYEHSALAMEH B,NABAVI M,VAFERI B. Modeling of CO2 capture ability of [Bmim][BF4] ionic liquid using connectionist smart paradigms[J]. Environmental Technology & Innovation,2021,22:101484.

[27] LUNDBERG S M,LEE S I. A unified approach to interpreting model predictions[C]//Proceedings of the 31st international conference on NeurIPS. Long Beach:Curran Associates Inc,2017:4768-4777.

[28] 周晨杰,于光认. 氨基功能化离子液体-胺-水复合相变吸收体系在 CO2 捕集中的应用研究[D]. 北京:北京化工大学,2021.

[29] PRINGLE J M,GOLDING J J,BARANYAI K,et al. The effect of anion fluorination in ionic liquids—physical properties of a range of bis(methanesulfonyl)amide salts[J]. New Journal of Chemistry,2003,27:1504-1510.

[30] 陈　阳,胡　辉. 多位点功能化离子液体吸收剂的研制及其捕集 CO_2 规律研究[D]. 武汉:华中科技大学,2017.