

研究论文

DOI: 10.11949/0438-1157.

面向未知离子结构外推的离子液体 CO₂溶解度物理引导级联预测模型

包赢政¹, 周怀荣¹, 杨勇¹, 王东亮¹, 杨庆春²(¹ 兰州理工大学石油化工学院, 甘肃 兰州 730050; ² 合肥工业大学化学与化工学院, 安徽 合肥 230009)

摘要: 针对统一回归模型难以同时表征低溶解度尾部、低压近 Henry 区 and 高压非理想区的问题, 构建融合 COSMO-RS Henry 先验、Apparent Henry 校准、分类门控和压力路由的物理引导级联模型。基于 932124 个 COSMO-RS Henry 数据和 297 种离子液体的 22294 个实验数据进行建模。分类门控后, 低压和高压分支测试集决定系数 (coefficient of determination, R²) 分别为 0.905 和 0.865; 完整模型在离子液体配对 (ionic-liquid pair, IL-pair)、Cation 和 Anion 划分下的 R² 分别为 0.823、0.591 和 0.780, 较最佳基线提高 0.189、0.048 和 0.473。该模型可用于高 CO₂ 溶解度离子液体的计算初筛, 但仍需独立实验及多物性评价。

关键词: 离子液体; CO₂ 溶解度; COSMO-RS; Henry 常数; 物理引导机器学习; 级联模型

中图分类号: O645.12; TQ028.2; TP181

文献标志码: AA

文章编号: 0438-1157 (XXXX) XX-0001-14

Physics-guided cascade prediction of CO₂ solubility in ionic liquids for extrapolation to unseen ionic structures

BAO Yingzheng¹, ZHOU Huairong¹, YANG Yong¹, WANG Dongliang¹, YANG Qingchun²

(¹ College of Petrochemical Engineering, Lanzhou University of Technology, Lanzhou 730050, Gansu, China; ² School of Chemistry and Chemical Engineering, Hefei University of Technology, Hefei 230009, Anhui, China)

Abstract: A physics-guided cascade model integrating a COSMO-RS Henry prior, apparent-Henry calibration, classification gating, and pressure-based routing was developed to address the difficulty of simultaneously describing the low-solubility tail, the low-pressure near-Henry region, and the high-pressure non-ideal region using a unified regression model. The framework was constructed using 932,124 COSMO-RS Henry-constant data points and 22,294 experimental CO₂-solubility data points for 297 ionic liquids. After classification gating, the low- and high-pressure branches achieved test R² values of 0.905 and 0.865, respectively. The complete cascade model achieved test R² values of 0.823, 0.591, and 0.780 under the IL-pair, Cation, and Anion splits, representing improvements

Keywords: ionic liquid; CO₂ solubility; COSMO-RS; Henry constant; physics-guided machine learning; cascade model

of 0.189, 0.048, and 0.473 over the corresponding computational prescreening and candidate-ranking best baselines. The model can be used as a tool for high-CO₂-solubility ionic liquids, while

收稿日期: 2026-05-18 修回日期: 2026-06-20

通信作者: 周怀荣(1988—), 男, 博士, 副教授, zhouhr@lut.edu.cn

第一作者: 包赢政(2002—), 男, 硕士研究生, 2043983041@qq.com

基金项目: 国家自然科学基金项目(22568027)

引用本文: 包赢政, 周怀荣, 杨勇, 王东亮, 杨庆春. 面向未知离子结构外推的离子液体 CO₂溶解度物理引导级联预测模型[J]. 化工学报, XXXX, XX(XX): 1–14

Citation: BAO Yingzheng, ZHOU Huairong, YANG Yong, WANG Dongliang, YANG Qingchun. Physics-guided cascade prediction of CO₂ solubility in ionic liquids for extrapolation to unseen ionic structures[J]. CIESC Journal, XXXX, XX(XX): 1–14

independent experiments and additional physicochemical-property evaluations remain necessary.

引 言

CO₂ 捕集与资源化利用是碳中和背景下过程工业绿色低碳发展的重要方向^[1]。离子液体因其低挥发性、结构可设计性和 CO₂ 亲和能力可调等特点,被认为是 CO₂ 吸收与富集过程中的潜在绿色溶剂^[2,3]。然而,离子液体由阳离子和阴离子组合构成,候选空间巨大,依靠实验逐一测定 CO₂ 溶解度成本高、周

期长,难以满足高通量筛选需求。机器学习方法可从分子结构和操作条件中学习结构—性质关系,为离子液体性质预测提供了有效手段^[4,5]。已有模型多采用统一回归策略直接预测所有样本的 CO₂ 溶解度,但该策略容易忽略低压和高压条件下溶解机理的差异,也难以区分低溶解度无效候选物与具有进一步评价价值的高溶解度候选物。其中,随机划分可能使相同阳离子、阴离子或离子配对在不同温度和压力下的数据同时进入训练集和测试集,因此随机划分下的较高精度主要反映同分布插值能力,不能直接说明模型能够外推至实验训练集中未出现的离子结构^[6]。

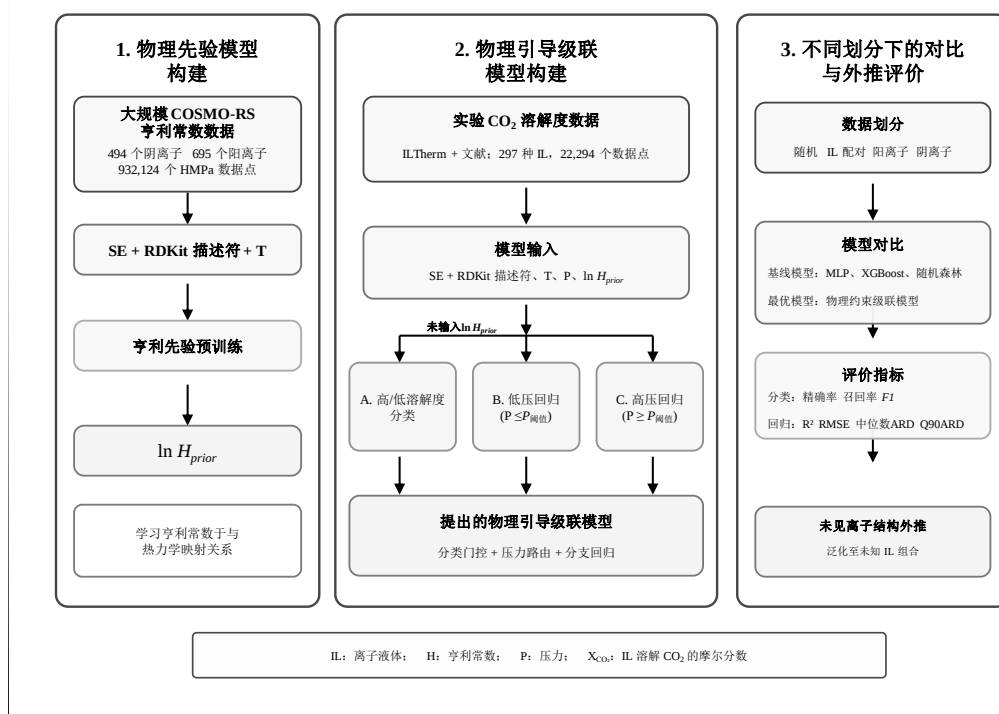


图 1 基于 COSMO-RS Henry 先验的 CO₂ 溶解度级联预测框架

Fig.1 Framework of CO₂ solubility cascade prediction based on COSMO-RS Henry prior

此外,CO₂在离子液体中的溶解度不仅是评价吸收容量的基础热力学性质,也直接影响气液传质推动力、吸收剂循环负荷和候选溶剂的工程筛选顺序。对于离子液体这类结构可设计溶剂而言,离子液体阴、阳离子结构及功能位点的变化会改变其极性环境、CO₂相互作用强度和捕集性能,因此不同离子组合可能表现出显著不同的溶解与吸收行为^[7]。因此,在实验数据有限的条件下,建立能够面向未知离子结构进行外推的 CO₂ 溶解度预测模型,对高通量绿色溶剂初筛具有实际意义^[8,9]。

从热力学角度看,CO₂在离子液体中的低压溶解度与 Henry 常数密切相关^[10,11]。低压条件下,CO₂溶解度近似满足 Henry 定律;而在高压条件下,气相非理想性、液相活度偏差和分子间相互作用增强会导致溶解度偏离简单线性关系^[12-16]。因此,若将 Henry 常数作为连接离子结构与宏观 CO₂ 溶解度的热力学桥梁,并结合实验数据学习其校准关系或偏离修正,有望提升模型的物理解释性和严格结构外推能力。基于

此,本文提出一种基于 COSMO-RS Henry 先验

的CO₂溶解度级联预测方法。整体研究框架如图1所示。

为将离子液体结构输入机器学习模型,本文采用标准化后的阳离子和阴离子简化分子线性输入规范(simplified molecular-input line-entry system, SMILES)分别构建结构编码(structural encoding, SE)和RDKit开源化学信息学工具包(RDKit open-source cheminformatics toolkit)。SE用于表示离子结构中的离散结构片段和局部连接模式;RDKit descriptors用于补充分子组成、拓扑结构、极性相关描述符及其他二维分子结构信息。对于每一种离子液体,阳离子和阴离子分别独立编码,形成阳离子结构特征向量 Z_{cat} 和阴离子结构特征向量 Z_{ani} ^[17]。在Henry prior预训练阶段,模型输入为 Z_{cat} 、 Z_{ani} 和温度 T ,输出为 $\ln H_{prior}$;在实验校准和级联预测阶段,分类门控模型使用阳、阴离子结构特征、温度和压力;低压和高压回归分支进一步引入预训练模型输出的 $\ln H_{prior}$ 。

与仅依赖结构描述符和操作条件直接回归CO₂溶解度的模型相比,本文将COSMO-RS Henry先验引入实验溶解度预测,并通过Apparent Henry校准建立离子结构、无限稀释热力学行为与实际压力条件下CO₂溶解度之间的联系。在此基础上,进一步设置高/低溶解度分类门控和压力分区回归,以分别处理低溶解度尾部、低压近Henry行为和高压非理想偏离。最后,采用random、IL-pair、Cation和Anion划分,区分评价模型的同分布插值能力和实验监督数据中未见离子结构的外推能力。

1 数据与方法

1.1 数据来源与预处理

本文使用的数据包括COSMO-RS计算数据和实验CO₂溶解度数据。COSMO-RS数据来自DATABASE-COSMO-2021中的阳离子和阴离子结构组合。基于695种阳离子和494种阴离子进行组合,共形成343330种候选离子液体配对;经批量计算后获得932124个Henry常数数据点,记为 H_{MPa} ,用于Henry prior模型预训练。

实验数据来自ILTherm开源离子液体数据库和Chen等^[18, 19]文献中的CO₂-离子液体溶解度数据。本文预测对象为CO₂在离子液体相中的摩尔分数 X_{CO_2} 。经相态错误剔除、离子配对异常检查、关键字

段缺失值删除和SMILES标准化处理后,最终获得297种离子液体和22294个实验数据点^[18, 19]。实验字段包括温度 T 、压力 P 、CO₂摩尔分数 X_{CO_2} 、阳离子SMILES、阴离子SMILES及标准化离子结构信息。数据集组成见表1。

本文采用结构编码(structural encoding, SE)和RDKit二维分子描述符表征阳离子和阴离子结构。SE包括半径为2的1024位Morgan指纹和167位分子访问系统结构键(Molecular ACCess System structural keys, MACCS keys)结构键;RDKit描述符主要覆盖分子组成、拓扑指数、环与芳香性、氢键供受体、拓扑极性表面积、疏水性及片段计数等二维结构信息。需要说明的是, σ -profile仅用于COSMO-RS计算Henry常数的底层热力学计算,未作为分段描述符直接输入本文机器学习模型。经无效值和常数描述符剔除后,每个离子最终保留217项RDKit描述符,总输入维数为1408。

从数据分布看,实验数据在压力维度上并不均衡。低压子集包含18022个数据点,而高压子集仅包含4272个数据点;低溶解度样本为1478个,明显少于主分支样本20816个。该分布特征说明,若直接采用统一回归模型,模型更容易被主导样本区间影响,并可能削弱对低溶解度边界样本和高压非理想区域的刻画能力。因此,本文后续引入高/低溶解度分类门控和压力分区回归,并非单纯增加模型复杂度,而是针对实验数据分布不均和不同压力区间溶解行为差异进行建模。

1.2 Henry prior 预训练模型

COSMO-RS计算得到的Henry常数包含离子结构、CO₂-离子液体相互作用和无限稀释溶解行为等热力学信息^[2, 10, 11]。本文首先利用COSMO-RS计算数据训练Henry prior预训练模型,以学习离子液体结构与Henry常数之间的映射关系。该模型采用基于PyTorch的前馈神经网络回归结构,以阳离子和阴离子的结构特征向量以及温度作为输入,以COSMO-RS计算得到的Henry常数对数值作为监督学习目标^[20]。

设由SE和RDKit descriptors得到的阳离子结构特征为 Z_{cat} ,阴离子结构特征为 Z_{ani} ,温度为 T ,则Henry prior模型可表示为:

$$\ln H_{prior} = f_{\theta}^H(Z_{cat}, Z_{ani}, T) \quad (1)$$

式中, f_{θ}^H 表示Henry prior神经网络模型, θ 为模型参数, H_{prior} 为模型预测得到的COSMO-RS Henry

表 1 数据集来源与组成

Table 1 Sources and composition of datasets

数据模块	数据来源	划分依据	温度范围/K	压力范围/MPa	离子液体数量	数据点数
COSMO-RS Henry 数据	DATABASE-COSMO-2021	Henry prior 预训练	373.15—423.15	0.100—1	343330	932124
实验总数据	ILTherm、Chen 等文献 ^[18, 19]	清洗后全数据	243.20—453.15	1.3×10^{-9} —34.55	297	22294
低压实验子集	实验数据	$P \leq 1$ MPa	243.20—453.15	1.3×10^{-9} —1	184	18022
高压实验子集	实验数据	$P > 1$ MPa	243.20—403.15	1—34.55	166	4272
低溶解度样本	实验数据	$X_{\text{CO}_2} < 0.01$	281.90—413.15	1.3×10^{-9} — 5.93×10^{-1}	127	1478
主分支样本	实验数据	$X_{\text{CO}_2} \geq 0.01$	243.20—453.15	3×10^{-6} —34.55	272	20816

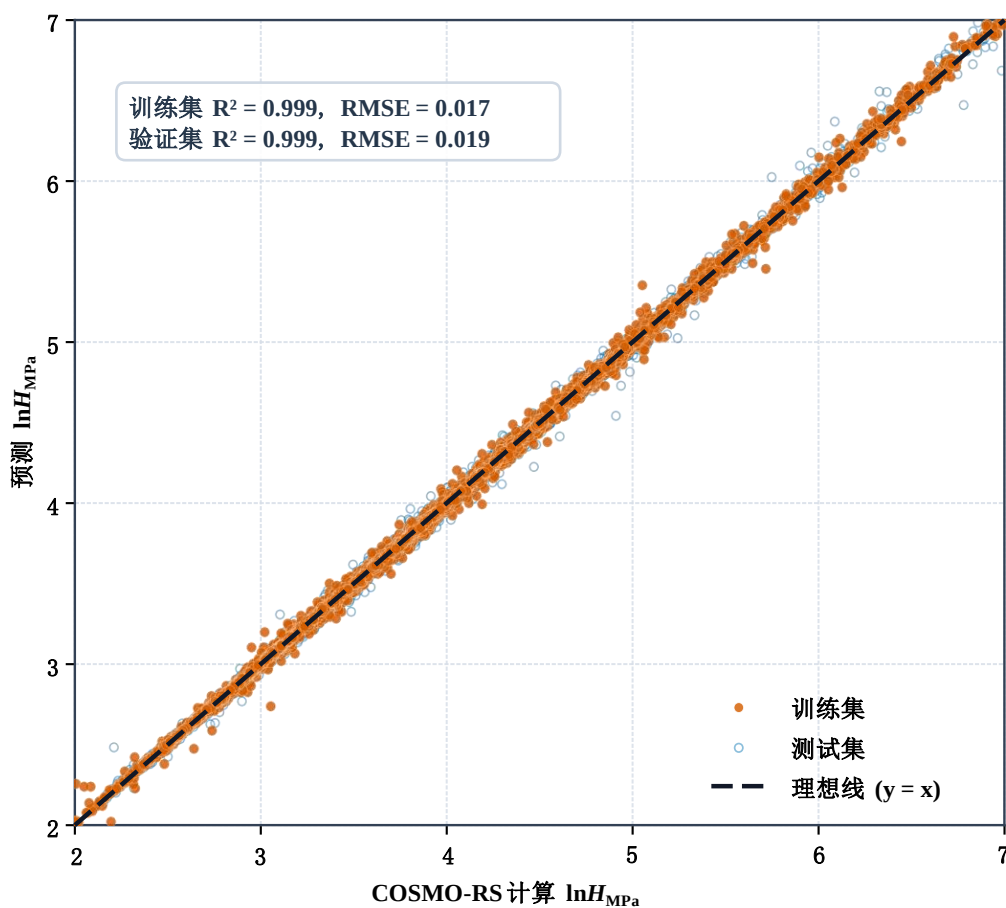


图 2 Henry prior 模型在 COSMO-RS Henry 常数数据上的预测性能

Fig.2 Prediction performance of the Henry prior model on COSMO-RS Henry-constant data

常数先验。模型训练目标为最小化预测值 $\ln H_{\text{prior}}$ 与 COSMO-RS 计算值 $\ln H_{\text{MPa}}$ 之间的误差。采用对数 Henry 常数作为学习目标,可以降低 Henry 常数跨数量级变化对模型训练稳定性的影响。

在特征输入方面,阳离子和阴离子分别由标准化 SMILES 构建 SE 特征和 RDKit 分子描述符,并形

成对应的特征矩阵。后续实验校准、分类门控和级联建模均不重新计算 COSMO-RS,而是调用预训练模型及其保存的特征矩阵、特征字典和归一化参数,为实验数据预测提供 $\ln H_{\text{prior}}$ 热力学先验特征。

Henry prior 模型的训练和验证结果如图 2 所示。训练集 $R^2=0.999$,均方根误差(root mean square

error, RMSE)=0.017; 验证集 $R^2=0.999$, RMSE=0.019, 表明该模型能够较好复现 COSMO-RS 计算体系内部的结构—Henry 常数映射关系。需要说明的是, 该高精度主要反映模型对 COSMO-RS 计算数据分布的学习能力, 并不意味着 Henry prior 可直接替代实验 CO_2 溶解度预测。本文仅将其作为热力学先验提取器, 用于后续 Apparent Henry 校准、分类门控和压力分区回归。其结构外推有效性则通过 IL-pair、Cation 和 Anion 严格划分下的实验溶解度预测结果进一步验证^[11, 21]。

1.3 Henry 定律引导与偏离修正策略

为分析 Henry prior 在不同压力条件下的迁移作用, 本文在低压和高压区域分别构建并比较三种 Henry-prior 物理迁移策略。需要说明的是, 实验压力原始单位为 kPa, 而 COSMO-RS 计算的 Henry prior 的单位为 MPa, 因此在所有 Henry 相关公式中均先进行单位统一。实验压力在建模前统一由 kPa 换算为 MPa, COSMO-RS Henry prior 和实验表观 Henry 常数均采用 MPa。

第一种为 Raw Henry prior 策略。该策略不引入实验校准参数, 而是直接利用 COSMO-RS Henry prior 和压力计算 CO_2 溶解度:

$$\ln x_{\text{CO}_2}^{\text{raw}} = \ln P_{\text{MPa}} - \ln H_{\text{prior}} \quad (2)$$

该策略可用于评价未经实验校准的 COSMO-RS Henry prior 是否能够直接迁移至实验溶解度预测任务。若该策略表现较差, 说明 COSMO-RS 计算值与实验溶解度之间存在系统偏差, 不能简单将 Henry prior 直接作为实验溶解度预测结果^[21]。

第二种为 Apparent Henry 策略。根据 Henry 定律, 可由实验溶解度反推出表观 Henry 常数^[11, 13]:

$$H_{\text{app}} = \frac{P_{\text{MPa}}}{x_{\text{CO}_2}} \quad (3)$$

其对数形式为:

$$\ln H_{\text{app}} = \ln P_{\text{MPa}} - \ln x_{\text{CO}_2} \quad (4)$$

模型学习表观 Henry 常数与 Henry prior 之间的校正关系:

$$\ln H_{\text{app,pred}} = a \ln H_{\text{prior}} + b + \Omega_{\theta} \quad (5)$$

最终由预测得到的表观 Henry 常数反算 CO_2 溶解度:

$$\ln x_{\text{CO}_2}^{\text{pred}} = \ln P_{\text{MPa}} - \ln H_{\text{app,pred}} \quad (6)$$

该策略的优势在于将溶解度预测转化为“Henry prior 校准—表观 Henry 常数预测—溶解度反算”的过程。对于低压区域, H_{app} 更接近真实 Henry

常数, 模型主要学习 COSMO-RS Henry prior 与实验 Henry 行为之间的校准关系^[11, 13]; 对于高压区域, H_{app} 不再是严格意义上的 Henry 常数, 而应理解为吸收压力非线性、气相非理想性、液相活度偏差及 CO_2 —离子液体相互作用增强等因素的有效表观参数^[12-15]。因此, Apparent Henry 策略既保留了 Henry prior 的热力学参照作用, 又允许模型在高压区域通过有效表观参数吸收 beyond-Henry 偏离^[22]。

第三种为 Log-residual 策略。该策略首先对 Henry prior 进行线性校准, 并构建 Henry-law 物理基线:

$$\ln x_{\text{CO}_2}^{\text{base}} = \ln P_{\text{MPa}} - (a \ln H_{\text{prior}} + b) \quad (7)$$

然后由模型学习实验溶解度相对于该物理基线的对数残差:

$$\ln x_{\text{CO}_2}^{\text{low}} = \ln P_{\text{MPa}} - (a_{\text{low}} \ln H_{\text{prior}} + b_{\text{low}}) + \Delta_{\theta}^{\text{low}}(Z_{\text{cat}}, Z_{\text{ani}}, T, P) \quad (8)$$

$$\ln x_{\text{CO}_2}^{\text{high}} = \ln P_{\text{MPa}} - (a_{\text{high}} \ln H_{\text{prior}} + b_{\text{high}}) + \Delta_{\theta}^{\text{high}}(Z_{\text{cat}}, Z_{\text{ani}}, T, P) \quad (9)$$

其中, Δ_{θ} 为实验数据驱动学习得到的偏离修正项。与 Raw Henry prior 策略相比, Log-residual 策略能够校正 COSMO-RS Henry prior 与实验值之间的系统偏差; 与 Apparent Henry 策略相比, 该策略直接在溶解度残差空间学习偏离修正, 而不是显式预测表观 Henry 常数。

低压和高压分支虽然均采用 Apparent Henry 表达形式, 但其物理含义不同。低压区域中, CO_2 溶解行为更接近 Henry 定律控制, $H_{\text{app}}^{\text{low}}$ 可近似反映实验 Henry 常数^[23], 模型主要学习 COSMO-RS Henry prior 与实验值之间的校准关系。高压区域中, 简单 Henry 线性关系逐渐失效, $H_{\text{app}}^{\text{high}}$ 则可视作等效表观参数, 用于表征压力非线性、气相非理想性和液相相互作用增强等 beyond-Henry 偏离。因此, 压力分区并非简单的数据拆分, 而是针对不同溶解行为区域进行物理分流预测。

1.4 分类门控与压力分区回归

为提高模型对高 CO_2 溶解度候选离子液体的筛选针对性, 本文在回归预测前引入高/低溶解度分类门控。该门控的作用不是给出严格的热力学相界, 而是从筛选任务角度优先识别低溶解度样本, 减少其对后续主回归模型的干扰。本文将 $X_{\text{thr}}=0.01$ 设为面向候选筛选任务的低溶解度识别阈值, 用于分离数据集的低端尾部, 并不代表热力学相界或普适临界溶解度。按此阈值划分, 低溶解度样本为 1478 个, 占全部实验数据的 6.63%, 当 $X_{\text{CO}_2} < 0.01$ 时定义为

低溶解度样本,当 $X_{\text{CO}_2} \geq 0.01$ 时定义为进入主回归分支的候选样本。分类标签定义为:

$$c = \begin{cases} 1, & x_{\text{CO}_2} < x_{\text{thr}} \\ 0, & x_{\text{CO}_2} \geq x_{\text{thr}} \end{cases} \quad (10)$$

式中, $c=1$ 表示低溶解度样本, $c=0$ 表示主分支候选样本。

分类门控模型采用极端梯度提升(extreme gradient boosting, XGBoost)二分类器^[24],并通过贝叶斯优化确定主要超参数^[25]。分类器输入包括阳离子结构特征、阴离子结构特征、阳/阴离子特征绝对差、温度和压力特征,可表示为:

$$X_{\text{cls}} = [Z_{\text{cat}}, Z_{\text{ani}}, |Z_{\text{cat}} - Z_{\text{ani}}|, T, \ln P_{\text{MPa}}, P_{\text{MPa}}] \quad (11)$$

其中, Z_{cat} 和 Z_{ani} 分别为阳离子和阴离子经SE与RDKit descriptors编码得到的结构特征向量, P_{MPa} 为以MPa为单位的压力。分类器输出样本属于低溶解度类别的概率:

$$p_{\text{low}} = f_{\theta}^{\text{cls}}(X_{\text{cls}}) \quad (12)$$

式中, f_{θ}^{cls} 表示分类门控模型, p_{low} 为低溶解度预测概率。预测阶段采用硬阈值路由规则:

$$r = \begin{cases} \text{reject}, & p_{\text{low}} \geq T_{\text{low}} \\ \text{main}, & p_{\text{low}} < T_{\text{low}} \end{cases} \quad (13)$$

其中, $T_{\text{low}}=0.03$ 为低溶解度路由阈值,仅依据验证集确定,并在测试集评价前固定。其选择综合考虑低溶解度召回率、low-to-main leakage和真实主分支样本保留率,而不是单纯追求总体分类准确率最大。

进入main分支的样本进一步以 $P_{\text{thr}}=1$ MPa作为低压和高压回归分支的操作分区边界,分别进入低压或高压回归分支。综合考虑了低压近Henry行为、实验压力覆盖和两个分支的样本量。该值属于本文建模中的物理—数据折中边界,不代表所有 CO_2 —离子液体体系的普适临界压力:

$$S(P) = \begin{cases} \text{low pressure}, & P_{\text{MPa}} \leq P_{\text{thr}} \\ \text{high pressure}, & P_{\text{MPa}} > P_{\text{thr}} \end{cases}, P_{\text{thr}} = 1 \text{ MPa} \quad (14)$$

根据1.3节的物理迁移策略比较,本文在低压和高压分支中均采用Apparent Henry策略。低压分支的表现Henry常数预测可表示为:

$$\ln H_{\text{app}}^{\text{low}} = a_{\text{low}} \ln H_{\text{prior}} + b_{\text{low}} + \Omega_{\theta}^{\text{low}}(Z_{\text{cat}}, Z_{\text{ani}}, T, P) \quad (15)$$

并由Henry关系反算 CO_2 摩尔分数:

$$\ln x_{\text{CO}_2}^{\text{low}} = \ln P_{\text{MPa}} - \ln H_{\text{app}}^{\text{low}} \quad (16)$$

高压分支可表示为:

$$\ln H_{\text{app}}^{\text{high}} = a_{\text{high}} \ln H_{\text{prior}} + b_{\text{high}} + \Omega_{\theta}^{\text{high}}(Z_{\text{cat}}, Z_{\text{ani}}, T, P) \quad (17)$$

$$\ln x_{\text{CO}_2}^{\text{high}} = \ln P_{\text{MPa}} - \ln H_{\text{app}}^{\text{high}} \quad (18)$$

式中, a_{low} 、 b_{low} 、 a_{high} 和 b_{high} 为Henry prior校准参数; $\Omega_{\theta}^{\text{low}}$ 和 $\Omega_{\theta}^{\text{high}}$ 分别为低压和高压分支中的数据驱动校正项。低压分支中的 $\Omega_{\theta}^{\text{low}}$ 主要用于修正COSMO-RS Henry prior与实验表现Henry常数之间的偏差;高压分支中的 $\Omega_{\theta}^{\text{high}}$ 则进一步表征压力非线性、气相非理想性和液相相互作用增强引起的beyond-Henry偏离。

综上,最终级联预测流程可写为:

$$x_{\text{CO}_2} = \begin{cases} 0, & P_{\text{low}} \geq T_{\text{low}} \\ f_{\theta}^{\text{low}}(x), & P_{\text{low}} < T_{\text{low}}, P \leq P_{\text{thr}} \\ f_{\theta}^{\text{high}}(x), & P_{\text{low}} < T_{\text{low}}, P > P_{\text{thr}} \end{cases} \quad (19)$$

该级联结构将低溶解度样本识别、低压Henry主导溶解行为和高压非理想溶解行为分开处理,避免单一统一回归模型同时拟合不同物理区域所造成的样本异质性问题^[21, 26]。

1.5 模型评价方式

为评价模型在不同泛化场景下的性能,本文采用random、IL-pair、Cation和Anion四种数据划分方式。其中本文所称的“未知”或“未见”离子结构,是指在相应实验监督训练划分中未出现的离子配对、阳离子或阴离子。部分离子结构可能已经包含在COSMO-RS Henry预训练数据中,因此IL-pair、Cation和Anion划分不等同于整个预训练—实验建模流程中的完全零样本预测,也不等同于独立前瞻性实验验证。random划分用于评价同分布插值能力;IL-pair划分用于评价未知离子液体组合的泛化能力;Cation和Anion划分分别用于评价未知阳离子和未知阴离子的结构外推能力^[2-4]。在每种划分方式下,数据均进一步划分为训练集、验证集和测试集,比例为70%/10%/20%。训练集用于模型参数学习,验证集用于超参数选择、早停和分类路由阈值确定,测试集用于最终性能评价。为保证不同模型之间的公平比较,XGBoost、Random Forest、多层感知机(multilayer perceptron, MLP)基线模型和本文级联模型均采用相同的训练/验证/测试划分。回归模型采用 R^2 、RMSE、平均绝对误差(mean absolute error, MAE)、绝对相对偏差中位数(median absolute relative deviation, MedianARD)和绝对相对偏差第90百分位数(90th percentile of absolute relative deviation, Q90ARD)作为评价指标,其中相对误差定义为:

$$ARD_i = \left| \frac{x_{\text{pred}} - x_i}{x_i} \right| \times 100\% \quad (20)$$

分类模型除 accuracy、precision、recall和F1分

数(F1-score)外,重点考察低溶解度样本误进入主分支比例和潜在高溶解度样本误剔除比例。

1.6 特征贡献分析

为量化不同输入信息对分类门控和回归分支预测结果的影响,本文在冻结已训练分类器、Henry prior 模型及低压和高压回归模型参数的条件下,开展分组置换重要性和局部分组扰动分析。考虑到模型输入同时包含高维 Morgan 指纹、MACCS 结构键、RDKit 二维描述符、操作条件及 Henry prior,且单个指纹位难以与唯一化学片段建立稳定对应关系,本文不对单个指纹位进行排序,而是按照特征来源将输入划分为阳离子 Morgan、阳离子 MACCS、阳离子 RDKit、阴离子 Morgan、阴离子 MACCS、阴离子 RDKit、阳阴离子差异特征、温度、压力和 Henry prior 等特征组。

对于分类门控模型,在 IL-pair 独立测试集上分别整体置换各特征组,并重新计算分类 Balanced Accuracy 及样本路由结果。设共有(G)个特征组,记 $g \in 1, \dots, G$ 为特征组索引,则第(g)个特征组的分类重要性定义为:

$$I_g^{BA} = BA_{base} - \frac{1}{N} \sum_{r=1}^N BA_{g,r}^{perm} \quad (21)$$

式中, BA_{base} 为未扰动测试集的 Balanced Accuracy, $BA_{g,r}^{perm}$ 为第(r)次置换第(g)组特征后的 Balanced Accuracy, (N=30)为重复置换次数。

同时,本文统计特征扰动前后样本门控结果发生变化的比例,并将其定义为路由变化率:

$$C_g = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(y_{i,g}^{perm} \neq y_i^{base}) \quad (22)$$

为减少分类门控错误对回归重要性评价的干扰,回归部分采用 forced-main 评价方式,即对真实 $x_{CO_2} \geq 0.01$ 的主分支样本不执行低溶解度拒绝,而是根据压力分别进入低压或高压回归模型。第(g)个特征组的回归重要性定义为:

$$I_g^{R^2} = R_{base}^2 - \frac{1}{N} \sum_{r=1}^N R_{g,r,perm}^2 \quad (23)$$

其中, R_{base}^2 为原始测试集上的回归决定系数, $R_{g,r,perm}^2$ 为第(r)次特征组置换后的决定系数。由于置换后模型预测可能显著劣于使用测试集均值的预测结果,置换后的 R^2 可以为负,因此 R^2 下降量可能大于1。

在全局置换分析基础上,进一步对完整级联模型开展局部分组扰动分析。对于测试样本(i)和特征组(g),局部扰动效应定义为:

$$E_{i,g} = f(x_i) - \frac{1}{B} \sum_{b=1}^B f(x_i^{g \leftarrow b}) \quad (24)$$

式中, $f(x_i)$ 为原始完整级联模型输出, $x_i^{g \leftarrow b}$ 表示使用第(b)个条件匹配背景样本替换第(g)组输入后的样本, (B=30)为背景扰动次数。正值表示该特征组相对于背景值提高模型输出,负值表示其降低模型输出。

特征组在全部测试样本中的平均绝对局部效应按下面公式计算:

$$\bar{E}_g^{abs} = \frac{1}{n} \sum_{i=1}^n |E_{i,g}| \quad (25)$$

该指标用于量化不同特征组在全部测试样本中的平均局部影响幅度,而不区分其作用方向。

为避免产生明显的分布外输入组合,压力扰动限定在原压力分支内部,即低压样本仅使用 $P \leq 1$ MPa 的压力背景,高压样本仅使用 $P > 1$ MPa 的压力背景;Henry prior 增量效应则在保持离子结构、温度和压力不变的条件下,使用相同路由分支和相近温度范围内的背景值进行替换。阳离子、阴离子或温度被扰动后,相关结构描述符、阳阴离子差异特征及 Henry prior 均同步重新计算。

需要说明的是,局部分组扰动效应反映当前模型对不同输入信息的局部统计敏感性,不满足 SHAP 解释方法 (SHapley additive explanations, SHAP) 值的严格加和性质,也不能直接解释为相应变量对 CO_2 溶解行为的物理因果效应。

2 结果与讨论

2.1 不同物理引导模型策略下的低压与高压回归性能

为验证 1.3 节中三种 Henry-prior 物理迁移策略的有效性,本文分别在低压区域 ($P \leq 1$ MPa) 和高压区域 ($P > 1$ MPa) 比较 Raw Henry prior、Apparent Henry 和 Log-residual 三种策略的预测性能。图 3 给出了三种策略对应的 R^2 和 MedianARD 对比结果。

结果表明,未经实验校准的 Raw Henry prior 策略在两个压力区域中均表现较差。低压区域中,该策略的 $R^2 = -0.754$, MedianARD 为 99.9%; 高压区域中,其 $R^2 = -0.203$, MedianARD 为 58.5%。这说明 COSMO-RS Henry prior 虽然能够提供结构相关的热力学基线,但若直接代入 Henry 定律预测实验溶解度,仍会受到系统偏差影响,难以准确再现实验数据^[27]。

相比 Raw Henry prior, Apparent Henry 和 Log-residual 策略均显著改善了低压和高压区域的预测性能,说明 COSMO-RS Henry prior 不能直接替代实验溶解度,而需要通过实验数据进行校准或偏离修正后才能有效迁移至 CO_2 -离子液体实验溶解度预测任务。Apparent Henry 和 Log-residual 策略均显著优于 Raw Henry prior,但二者在不同指标上的表现存在差异。低压区域中,Log-residual 的 R^2 和 MedianARD 略优于 Apparent Henry;高压区域中,Apparent Henry 的 R^2 略高,而 Log-residual 的 MedianARD 更低。因此,本文选择 Apparent Henry 并非因为其在所有指标上均最优,而是综合考虑其

表观参数含义、低压和高压统一的目标表达形式、高压 R^2 表现以及与分类门控和压力分区级联框架的适配性。

从物理角度看,Apparent Henry 策略并未直接将 COSMO-RS Henry prior 作为实验溶解度预测结果,而是通过实验数据学习 H_{prior} 与 H_{app} 之间的校准关系^[27]。在低压区域, H_{app} 更接近实验 Henry 常数,主要反映 Henry prior 的校正;在高压区域, H_{app} 则视为有效表观参数,用于表征非理想溶解行为造成的 beyond-Henry 偏离。由此,Henry prior 的作用是提供结构相关的热力学参照,而不是直接替代实验溶解度。

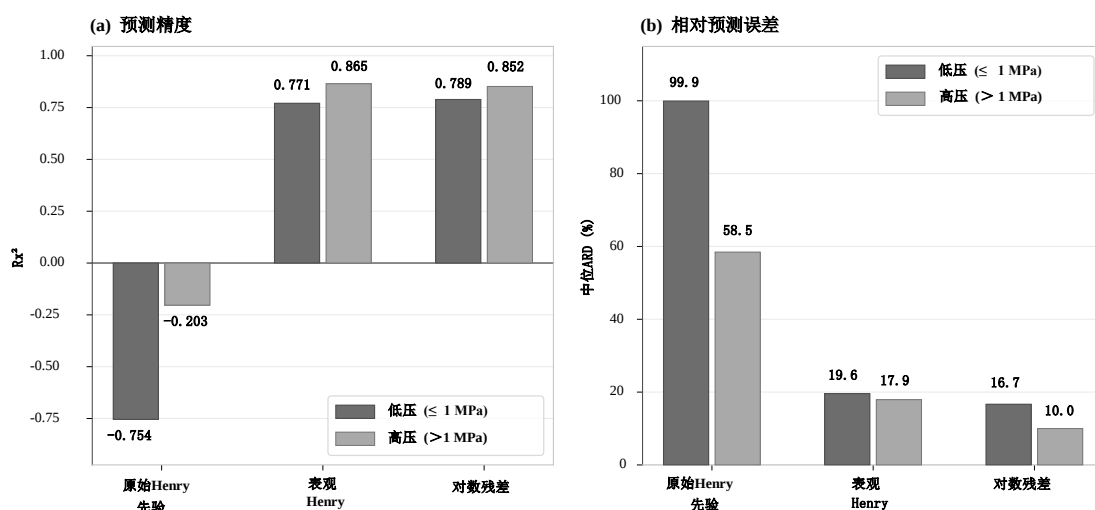


图3 不同物理引导模型策略在低压和高压区域的预测性能比较

Fig.3 Performance comparison of different physics-guided strategies in low- and high-pressure regions

2.2 分类门控模型性能

分类门控模型用于识别低溶解度体系并控制其进入主回归分支的比例。与普通二分类任务不同,本文更关注 low-to-main leakage 和 false rejection 两类筛选相关错误。优化后的分类模型采用高低筛分的硬路由规则,即 $P_{\text{low}} \geq T_{\text{low}}$ 时判定为低溶解度并剔除,否则进入主分支。分类结果如图4所示。

测试集结果显示,分类门控捕获了 97.6% 的低溶解度样本,仅有 2.4% 的真实低溶解度样本误进入 main 分支,说明该模型能够有效拦截低价值样本。与此同时,85.8% 的真实高溶解度样本被保留进入主回归分支,表明所采用的路由阈值并未过度牺牲潜在高溶解度候选物,分类门控在降低低溶解度样本干扰和

保留后续回归评价对象之间取得了较合理的折中。值得强调的是, $T_{\text{low}}=0.03$ 由验证集确定并在

测试集评价前固定,测试结果仅用于评价已固定路由规则的泛化性能。该阈值针对当前数据分布和筛选任务确定,不应视为其他数据集上的通用阈值。

2.3 分类门控后的低压与高压分支回归性能

根据 2.1 节的策略比较,本文最终在低压和高压分支中均采用 Apparent Henry 策略。图5给出了加入分类门控后,IL-pair 严格划分下最终低压和高压分支模型的预测结果及残差分布。该图用于验证所选 Apparent Henry 策略在级联筛选框架中的实际预测性能。

在低压区域 ($P \leq 1 \text{ MPa}$),模型测试集 $R^2=0.905$, RMSE 为 0.067, MAE 为 0.038。与未加入分类门控时的策略比较结果相比,分类门控后低压分支性能进一步提升,说明前置剔除低溶解度样本和边界样本能够提高主回归任务的一致性,使低压分支更集

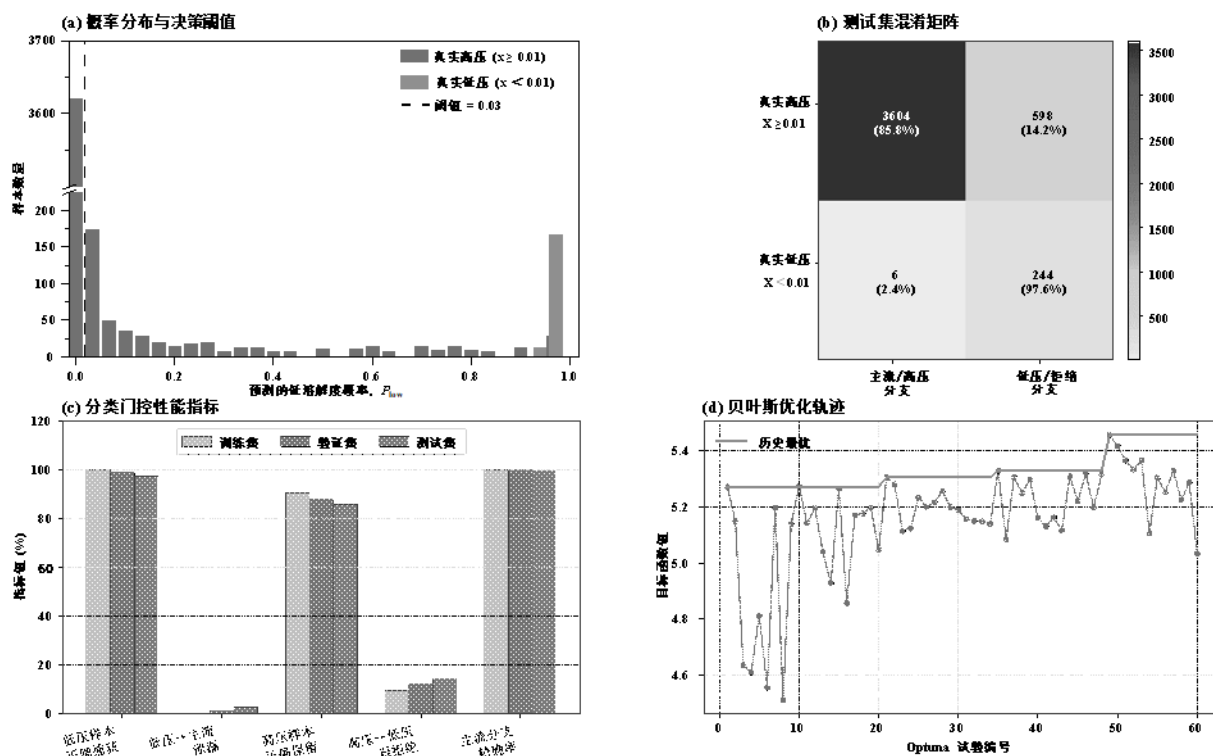


图4 高/低溶解度分类门控模型性能

Fig.4 Performance of the high/low-solubility classification gate

中地学习高溶解度候选体系中的 Henry 常数校准 关系。

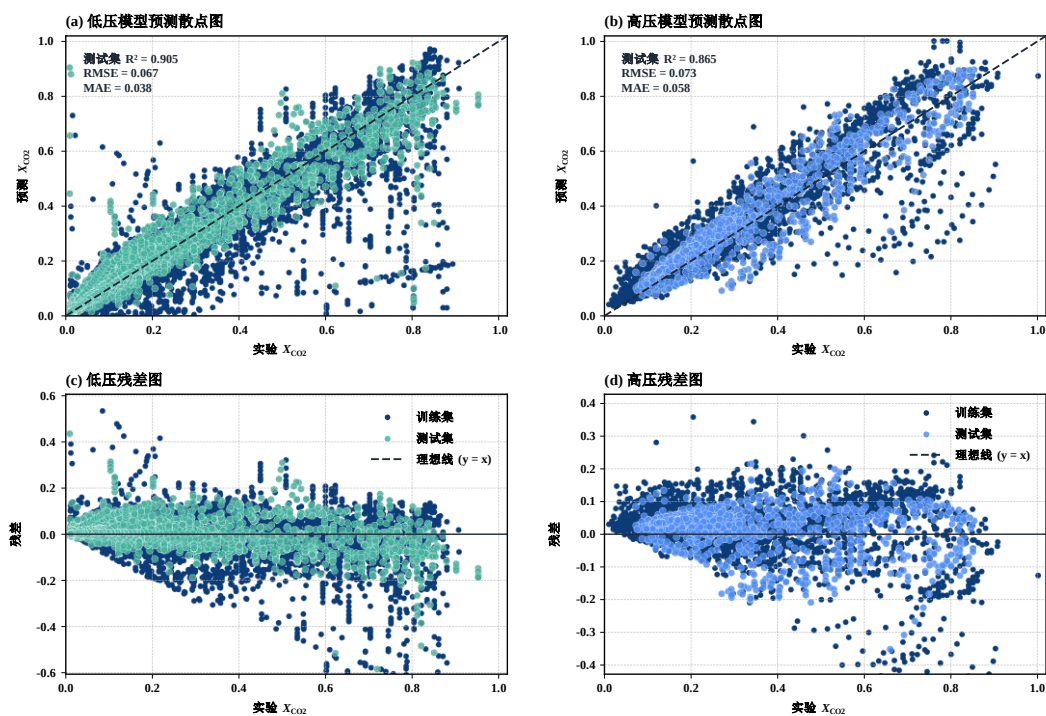


图5 IL-pair 严格划分下分类门控后的低压与高压回归模型预测性能

Fig.5 Prediction performance of low-pressure and high-pressure regression models after classification gating under IL-pair strict split

在高压区域($P > 1$ MPa), 模型测试集 $R^2 = 0.865$, RMSE 为 0.073, MAE 为 0.058。该结果说明, Apparent Henry 策略中的有效表观 Henry 参数能够吸收高压条件下的 beyond-Henry 偏离, 从而保持较好的预测稳定性。上述结果表明, Apparent Henry 物理先验迁移策略与分类门控机制具有协同作用, 共同提升了最终级联模型的预测稳定性。

高压分支 R^2 略低于低压分支, 主要与高压溶解行为复杂性和数据覆盖不足有关。一方面, 高压区域中 CO_2 溶解度不再主要受 Henry 线性关系控制, 压力非线性、气相非理想性和液相相互作用增强会共同影响表观溶解行为^[3, 12-15]; 另一方面, 表1显示高压实验子集仅有 4272 个数据点, 明显少于低压子集的 18022 个数据点。因此, 高压模型精度略低并不说明 Apparent Henry 策略失效, 而反映了高压 beyond-Henry 行为和样本覆盖不足带来的预测难度^[28]。

图5中仍存在少量残差较大的样本, 表明模型对近阈值溶解度样本和高压非理想体系的预测仍存在一定不确定性。该现象与表1中的数据分布基本一致: 高压实验子集仅包含 4272 个数据点, 明显少于低压子集的 18022 个数据点, 样本覆盖度相对不足; 同时, 数据集中仍有 1478 个低溶解度样本, 低摩尔分数或近阈值样本更容易造成相对误差放大。

2.4 级联模型在不同划分方式下的泛化性能分析

为进一步评价所提出级联模型在不同泛化场景下的预测能力, 本文在 random、IL-pair、Cation 和 Anion 四种划分方式下, 将最终级联模型与 XGBoost、Random Forest 和 MLP 基线模型进行比较^[29]。需要说明的是, 为避免评价对象不一致, 图6中所有模型均在相同的 retained true-main 样本上进行回归性能比较, 即低溶解度样本的识别与拦截由分类门控在 2.2 节中单独评价, 回归性能仅针对被保留进入主回归分支且真实 $X_{\text{CO}_2} \geq 0.01$ 的样本计算。因此, 图6反映的是在相同主分支评价样本空间内, 统一回归基线模型与分类门控—压力路由—Apparent Henry 分支级联模型之间的性能差异, 而不是由测试样本范围不同造成的偏差。

由图6(a)可知, 在 random 划分下, Random Forest 基线模型取得最高 $R^2 = 0.911$, 高于最终级联模型的 $R^2 = 0.784$, 最终级联模型相对最佳基线的 R^2 差值为 -0.127。这说明在训练集和测试集分布高度相近的随机插值场景中, 传统集成模型仍具有较强的数据拟合能力。该结果并不削弱级联模型的意义, 因为本文关注的核心问题不是同分布插值, 而是未知离子液体组合、未知阳离子和未知阴离子条件下的结构外推能力。

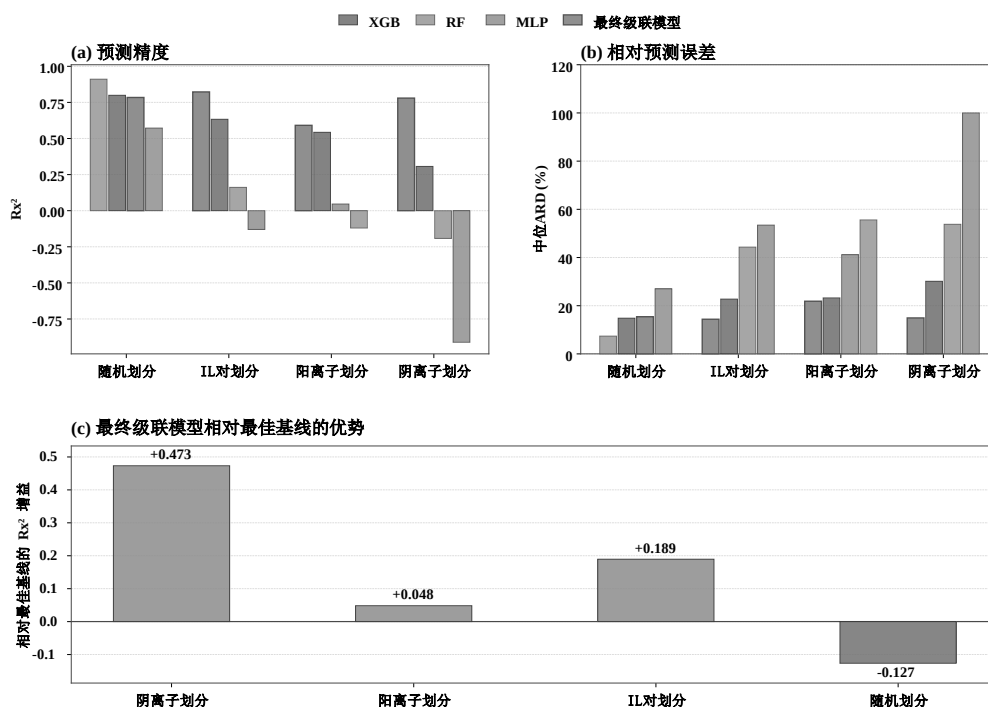


图6 不同数据划分方式下模型预测性能比较

Fig.6 Prediction performance comparison under different splitting strategies

在更接近实际溶剂筛选任务的严格结构外推场景中,最终级联模型表现出明显优势。在 IL-pair、Cation 和 Anion 划分下,最终级联模型的测试集 R^2 分别为 0.823、0.591 和 0.780,相对对应最佳基线模型分别提升 0.189、0.048 和 0.473。其中,IL-pair 划分用于评价未知离子液体组合的泛化能力,最终级联模型明显优于统一基线模型,说明 Henry prior 与 Apparent Henry 校准能够有效缓解未知离子配对条件下的预测不稳定性。Anion 划分下提升最为显著,表明引入 COSMO-RS Henry 先验后,模型对未知阴离子结构具有更强的外推适应能力^[30]。Cation 划分下最终级联模型仍优于最佳基线,但提升幅度相对有限,说明未知阳离子外推仍是模型后续需要进一步改进的方向^[31]。

从图 6(b)的 MedianARD 结果看,最终级联模型在严格划分下同样表现出更好的相对误差稳定性。在 IL-pair 和 Anion 划分下,最终级联模型的 MedianARD 分别为 14.37% 和 14.91%,明显低于对应最佳基线的 22.74% 和 30.13%;在 Cation 划分下,最终级联模型的 MedianARD 为 21.85%,较最佳基线的 23.24% 也有一定降低。上述结果表明,最终级联模型的优势不仅体现在 R^2 的提升上,也体现在严格结构外推场景中相对误差分布更加稳定。

综合来看,最终级联模型的主要优势并不体现在 random 随机插值场景,而体现在 IL-pair、Cation 和 Anion 等更具挑战性的结构外推场景^[2-4]。其原因在于,COSMO-RS Henry prior 为未知离子结构提供了热力学先验引导,Apparent Henry 策略将 CO_2 溶解度预测转化为表观 Henry 常数校准与反算问题,分类门控减少了低溶解度样本对主回归任务的干扰,而压力分区回归进一步区分了低压 Henry 常数校准行为和高压 effective apparent-Henry 偏离行为。因此,所提出的级联模型更适合用于未知离子液体结构空间中的高 CO_2 溶解度候选物筛选。上述结果属于现有实验数据库内部的严格结构留出评价,反映的是模型对实验训练集中未见离子配对、阳离子或阴离子的泛化能力,不等同于对模型筛选得到的全新候选离子液体开展独立实验验证。

2.5 完整级联模型特征贡献分析

为进一步分析级联模型预测性能的来源,本文对分类门控和回归分支分别开展分组置换重要性分析,并利用局部分组扰动热力图展示不同输入信息在测试样本中的贡献差异,结果如图 7 所示。

分类门控结果表明,压力是影响高、低溶解度分类及样本路由的首要变量。置换压力后,分类模型的 Balanced Accuracy 下降 0.362 ± 0.013 ,并导致 53.3% 的测试样本门控结果发生变化,明显高于其他特征组。温度和阳离子 RDKit 描述符的重要性接近, Balanced Accuracy 分别下降 0.087 ± 0.010 和 0.086 ± 0.026 ,对应路由变化率分别为 11.3% 和 12.8%。阴离子 RDKit 描述符和阳阴离子 RDKit 差异特征分别导致 Balanced Accuracy 下降 0.039 ± 0.022 和 0.038 ± 0.011 ,说明连续二维分子描述符能够为低溶解度尾部样本识别提供一定结构信息。阴离子 MACCS 结构键的影响相对较小, Balanced Accuracy 下降为 0.011 ± 0.006 , Henry prior 未作为分类门控模型的输入,因此不参与分类门控分组置换评价,其增量贡献仅在 forced-main 回归评价和完整级联局部扰动分析中考察。

上述结果与分类器的功能一致。分类门控的主要任务是依据当前温度、压力及离子结构判断样本是否落入 $x_{\text{CO}_2} < 0.01$ 的低溶解度区域。压力直接决定气体溶解驱动力,因此其对门控结果的影响最为显著;温度及 RDKit 描述符所包含的分子组成、拓扑结构、极性和片段计数信息则进一步修正不同离子体系在相同压力条件下的溶解度差异。与高维离散指纹相比,连续 RDKit 描述符在当前分类任务中表现出更稳定的全局贡献。

在 forced-main 回归评价中,压力和 Henry prior 表现出远高于其他特征组的重要性。置换压力后,回归模型的 R^2 下降 1.873 ± 0.039 ;置换 Henry prior 后, R^2 下降 1.033 ± 0.022 。该结果表明,压力不仅参与低压和高压分区,还直接进入 Apparent Henry 反算过程,是决定 CO_2 溶解度预测水平的主要操作变量。Henry prior 虽然不能未经实验校准直接代替实验溶解度,但能够为回归模型提供离子结构相关的无限稀释热力学参照,其贡献主要体现在低压和高压回归分支,而不是分类门控阶段。

温度置换后,回归 R^2 下降 0.254 ± 0.013 ,说明温度通过影响 Henry prior 和实际气体溶解行为,对最终预测产生不可忽略的作用。结构特征的总体贡献低于压力、Henry prior 和温度,其中阴离子 Morgan、阳离子 Morgan、阴离子 RDKit 和阳离子 RDKit 置换后, R^2 分别下降 0.065 ± 0.049 、 0.056 ± 0.029 、 0.046 ± 0.027 和 0.028 ± 0.024 。阴离子结构特征的平均贡献略高于对应阳离子特征,这与阴离子

在CO₂—离子液体相互作用、局部极性环境和气体亲和能力调节中的重要作用相一致^[30]。但不同结构特征之间的差异总体较小,且标准差相对较大,说明高维结构指纹的作用具有明显的样本依赖性。

图7进一步展示了500个测试样本中各输入特征组的局部分组扰动效应。样本按照低溶解度门控、低压分支和高压分支进行排列,并在各分支内部按照完整级联模型输出排序。热力图表明,压力在低压和高压分支中均产生最显著的局部扰动,其中低压分支后段以正向效应为主,而高压分支中同时存在正、负效应,说明压力对模型输出的影响取决于具体工况及样本所处的溶解度区间。

Henry prior在低压和高压分支中均表现出较强的局部效应,但其贡献方向随离子体系和操作条件发生变化。在部分样本中, Henry prior 相对于条件背景提高预测溶解度,而在另一些样本中则降低预测结果。这表明实验校准模型并未简单复制COSMO-RS Henry数值,而是根据离子结构、温度和压力学习 Henry prior 与实验表观 Henry 行为之间的体系相关校正关系。

温度、阳离子 RDKit 和阴离子 RDKit 描述符的

局部效应低于压力和 Henry prior,但在低溶解度门控区和部分高压样本中仍表现出明显贡献。Morgan 和 MACCS 指纹的平均绝对局部效应相对较小,但在少数离子体系中出现较强的正向或负向响应,说明特定局部结构片段能够影响个别样本的预测结果,但其作用并不具有跨全部离子体系的一致性。

综合全局置换和局部扰动结果,完整级联模型的预测主要由压力、Henry prior 和温度共同驱动, RDKit 连续分子描述符为分类门控及结构差异识别提供补充信息,高维 Morgan 和 MACCS 指纹则主要表征特定离子体系的局部结构效应。该结果说明,级联模型并非仅依赖操作条件进行经验拟合,而是在压力和温度信息基础上进一步利用 Henry 热力学先验及阳、阴离子结构信息完成预测。

需要强调的是,图7中的局部分组扰动效应和右侧平均绝对局部效应反映的是模型输出对输入信息变化的统计响应,不代表相应变量对CO₂溶解度的直接因果贡献。尤其是Morgan和MACCS属于分布式结构编码,单个指纹位不具有唯一、稳定的化学含义,因此本文仅在特征组层面讨论其作用,不对单个指纹位进行机理归因。

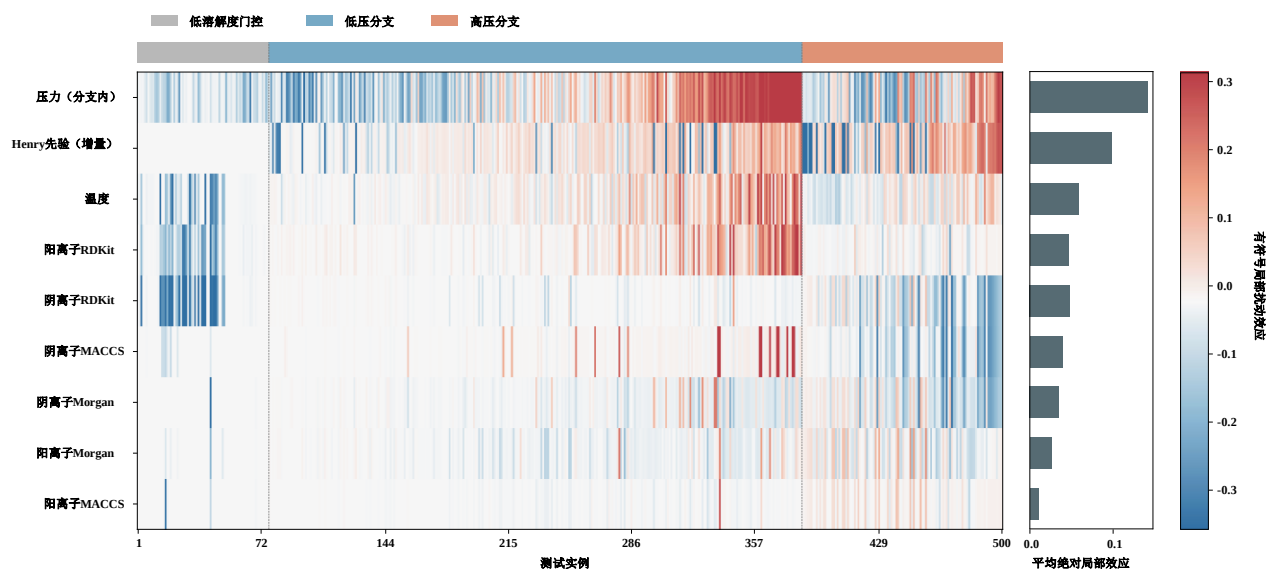


图7 输入信息对完整级联模型输出的局部分组扰动效应

Fig.7 Local grouped perturbation effects of input information on the complete cascade-model output

注:测试样本首先按照低溶解度门控、低压分支和高压分支进行划分,并在各分支内依据级联模型预测结果排序。热图中红色和蓝色分别表示正向和负向局部扰动效应,右侧柱状图表示对应特征组的平均绝对局部效应。该图用于表征模型对不同特征组的局部敏感性及其相对贡献,不构成SHAP值分解结果,也不直接代表物理因果关系。

2.6 面向绿色溶剂筛选的适用性与局限性

CO₂在离子液体中的溶解度是评价其作为CO₂吸收、富集和转化辅助介质的重要基础物性。本文提出的级联模型可用于大规模离子液体结构空间的初步筛选:首先通过分类门控排除低CO₂溶解度候选物,再对保留样本进行压力相关溶解度预测,从而为后续实验验证提供优先级排序。与直接统一回归相比,该流程的优势在于能够区分低溶解度样本、低压Henry主导区域和高压非理想区域,避免将不同物理行为强行交由单一模型处理。

需要指出的是,本文尚未对模型筛选得到的全新离子液体候选物开展前瞻性实验盲测。IL-pair、Cation和Anion划分属于现有实验数据库内部的严格结构留出评价,不能替代独立实验验证。因此,本文模型应定位为候选物计算初筛和优先级排序工具,不能直接据此认定候选物具备实际工程适用性^[32]。

此外,CO₂溶解度不是绿色溶剂筛选的唯一判据。实际应用中还需综合考虑离子液体黏度、热稳定性、腐蚀性、价格与合成可得性、再生能耗、H₂或其他气体共溶行为以及与催化剂或工艺介质的相容性。因此,本文模型更适合作为高通量初筛工具,用于识别值得进一步实验验证的高CO₂溶解度候选体系,而不能直接替代实验测定或完整过程评价^[15]。

3 结论

本文基于类导体真实溶剂屏蔽模型(conductor-like screening model for real solvents, COSMO-RS)计算数据构建了CO₂-离子液体Henry先验数据集,并训练Henry prior模型学习离子结构与Henry常数之间的映射关系。结果表明,该模型在COSMO-RS验证集上取得R²=0.999、RMSE=0.019,说明其能够较好复现COSMO-RS计算体系内部的结构-Henry常数关系。需要指出的是,该模型并不直接作为实验溶解度预测器,而是作为热力学先验提取器,为后续Apparent Henry校准、分类门控和压力分区回归提供 $\ln H_{\text{prior}}$ 特征。

(1)围绕Henry定律引导,本文比较了Raw Henry prior、Apparent Henry和Log-residual三种物理先验迁移策略。结果表明,未经实验校准的Raw Henry prior难以直接用于实验CO₂溶解度预测,而

Apparent Henry和Log-residual策略均能有效改善预测性能。结合策略比较结果与分类门控后的分支模型表现,本文最终采用Apparent Henry作为低压和高压分支的回归策略。该策略将溶解度预测转化为表观Henry常数校准与反算问题,其中低压分支主要对应Henry常数校准,高压分支则用于表征beyond-Henry偏离。Apparent Henry并非在所有单项指标上均优于Log-residual,其最终选用主要基于表观参数含义、统一目标表达和完整级联框架中的综合表现。

(2)本文建立的高/低溶解度分类门控与压力分区回归框架提高了主分支回归任务的样本一致性。分类门控在测试集中实现了97.6%的低溶解度捕获率,并将low-to-main leakage控制在2.4%。在IL-pair严格划分下,加入分类门控后的低压和高压分支模型分别取得测试集R²=0.905和测试集R²=0.865,表明该框架能够较好地地区分低溶解度样本、低压Henry主导区域和高压非理想区域。

(3)与XGBoost、Random Forest和MLP基线模型相比,最终级联模型的优势主要体现在严格结构外推场景中。在IL-pair、Cation和Anion划分下,最终级联模型相对最佳基线模型的R²分别提升0.189、0.048和0.473。其中,IL-pair和Anion划分下的提升较为明显,说明COSMO-RS Henry先验与Apparent Henry校准有助于增强未知离子液体组合和未知阴离子结构下的预测稳定性;Cation划分下提升幅度相对有限,表明未知阳离子外推仍是后续模型改进的重点。

(4)本文提出的级联框架可作为离子液体绿色溶剂筛选中的CO₂溶解度初筛工具,用于在未知离子结构空间中识别具有较高CO₂溶解潜力的候选体系。但CO₂溶解度并不能单独决定溶剂的工程适用性,后续仍需结合黏度、热稳定性、再生能耗、气体共溶行为和实验验证结果进行综合评价。需要指出的是,上述结构外推结果属于现有实验数据库内部的严格留出评价,模型筛选得到的全新候选体系仍需开展独立实验验证。

参考文献

- [1] 崔国凯,吕书贞,王键吉.功能化离子液体在二氧化碳吸收分离中的应用[J].化工学报,2020,71(1):16-25.
Cui G K, Lyu S Z, Wang J J. Functional ionic liquids for carbon dioxide capture and separation[J]. CIESC Journal, 2020, 71(1): 16-25.

- [2] Kuroki N, Suzuki Y, Kodama D, et al. Machine learning–boosted design of ionic liquids for CO₂ absorption and experimental verification[J]. The Journal of Physical Chemistry B, 2023, **127** (9): 2022–2027.
- [3] Ali M, Sarwar T, Mubarak N M, et al. Prediction of CO₂ solubility in Ionic liquids for CO₂ capture using deep learning models[J]. Scientific Reports, 2024, **14**: 14730.
- [4] Laakso J P, Gorji A E, Uusi-Kyyny P, et al. Machine learning modeling of the CO₂ solubility in ionic liquids by using σ -profile descriptors[J]. Chemical Engineering Science, 2025, **307**: 121226.
- [5] Kong X Y, Qiu Y X, Wang K, et al. Deep learning driven ionic liquid screening for toluene and styrene capture[J]. Chemical Engineering Science, 2026, **320**: 122724.
- [6] 张瑞娜, 田媛, 葛春亮, 等. 人工智能机器学习助力离子液体 CO₂捕集性能分析[J]. 科学通报, 2025, **70**(26): 4462–4472.
Zhang R N, Tian Y, Ge C L, et al. Artificial intelligence and machine learning for analysis of CO₂ capture by ionic liquids[J]. Chinese Science Bulletin, 2025, **70**(26): 4462–4472.
- [7] 张威, 张瑞娜, 葛春亮, 等. 功能化离子液体活性位点强化 CO₂捕集[J]. 低碳化学与化工, 2024, **49**(7): 23–33, 41.
Zhang W, Zhang R N, Ge C L, et al. Intensification of CO₂ capture by active sites in functional ionic liquids[J]. Low-Carbon Chemistry and Chemical Engineering, 2024, **49**(7): 23–33, 41.
- [8] Lei Z G, Dai C N, Chen B H. Gas solubility in ionic liquids[J]. Chemical Reviews, 2014, **114**(2): 1289–1326.
- [9] Eckert F, Klamt A. Fast solvent screening *via* quantum chemistry: COSMO–RS approach[J]. AIChE Journal, 2002, **48**(2): 369–385.
- [10] Zhang X C, Liu Z P, Wang W C. Screening of ionic liquids to capture CO₂ by COSMO–RS and experiments[J]. AIChE Journal, 2008, **54**(10): 2717–2728.
- [11] Sistla Y S, Khanna A. Validation and prediction of the temperature–dependent henry’s constant for CO₂ – ionic liquid systems using the conductor–like screening model for realistic solvation (COSMO–RS) [J]. Journal of Chemical & Engineering Data, 2011, **56**(11): 4045–4060.
- [12] Moya C, Gonzalez–Miquel M, Rodriguez F, et al. Non–ideal behavior of ionic liquid mixtures to enhance CO₂ capture[J]. Fluid Phase Equilibria, 2017, **450**: 175–183.
- [13] Kapoor U, Shah J K. Molecular origins of the apparent ideal CO₂ solubilities in binary ionic liquid mixtures[J]. The Journal of Physical Chemistry B, 2018, **122**(42): 9763–9774.
- [14] Karmakar A, Mukundan R. Modeling solubility of CO₂ gas in room temperature ionic liquids using the COSMO–RS model: a first principles study[J]. Physical Chemistry Chemical Physics, 2019, **21**(35): 19667–19685.
- [15] Liu X Y, O’Harra K E, Bara J E, et al. Solubility behavior of CO₂ in ionic liquids based on ionic polarity index analyses[J]. The Journal of Physical Chemistry B, 2021, **125**(14): 3665–3676.
- [16] 王伟彬, 银建中, 孙丽华, 等. CO₂/离子液体体系热力学性质的分子动力学模拟[J]. 物理化学学报, 2009, **25**(11): 2291–2295.
Wang W B, Yin J Z, Sun L H, et al. Molecular dynamics simulation of thermodynamic properties for CO₂/ionic liquid systems[J]. Acta Physico–Chemica Sinica, 2009, **25**(11): 2291–2295.
- [17] Rogers D, Hahn M. Extended–connectivity fingerprints[J]. Journal of Chemical Information and Modeling, 2010, **50**(5): 742–754.
- [18] Dong Q, Muzny C D, Kazakov A, et al. ILThermo: a free–access web database for thermodynamic properties of ionic liquids[J]. Journal of Chemical & Engineering Data, 2007, **52**(4): 1151–1159.
- [19] Chen X X, Chen G Z, Xie K C, et al. Exploring the chemical space of ionic liquids for CO₂ dissolution through generative machine learning models[J]. Green Chemical Engineering, 2025, **6**(3): 335–343.
- [20] Klamt A, Eckert F. COSMO–RS: a novel and efficient method for the *a priori* prediction of thermophysical data of liquids[J]. Fluid Phase Equilibria, 2000, **172**(1): 43–72.
- [21] Fan J, Dai Z X, Cao J, et al. Hybrid data–driven and physics–based modeling for viscosity prediction of ionic liquids[J]. Green Energy & Environment, 2024, **9**(12): 1878–1890.
- [22] Blanchard L A, Gu Z Y, Brennecke J F. High–pressure phase behavior of ionic liquid/CO₂ systems[J]. The Journal of Physical Chemistry B, 2001, **105**(12): 2437–2444.
- [23] Baltus R E, Culbertson B H, Dai S, et al. Low–pressure solubility of carbon dioxide in room–temperature ionic liquids measured with a quartz crystal microbalance[J]. The Journal of Physical Chemistry B, 2004, **108**(2): 721–727.
- [24] Chen T Q, 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 California USA. ACM, 2016: 785–794.
- [25] 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 AK USA. ACM, 2019: 2623–2631.
- [26] Specht T, Nagda M, Fellenz S, et al. HANNA: hard–constraint neural network for consistent activity coefficient prediction[J]. Chemical Science, 2024, **15**(47): 19777–19786.
- [27] Qin H L, Wang K, Ma X F, et al. Predicting the solubility of CO₂ and N₂ in ionic liquids based on COSMO–RS and machine learning[J]. Frontiers in Chemistry, 2024, **12**: 1480468.
- [28] Shiflett M B, Yokozeki A. Solubilities and diffusivities of carbon dioxide in ionic liquids: [bmim][PF₆] and [bmim][BF₄] [J]. Industrial & Engineering Chemistry Research, 2005, **44**(12): 4453–4464.
- [29] Breiman L. Random forests[J]. Machine Learning, 2001, **45**(1): 5–32.
- [30] Cadena C, Anthony J L, Shah J K, et al. Why is CO₂ so soluble in imidazolium–based ionic liquids [J]. Journal of the American Chemical Society, 2004, **126**(16): 5300–5308.
- [31] Yang K, Swanson K, Jin W G, et al. Analyzing learned molecular representations for property prediction[J]. Journal of Chemical Information and Modeling, 2019, **59**(8): 3370–3388.
- [32] 吕超, 张习文, 金理健, 等. 新型两相吸收剂–离子液体系统高效捕获 CO₂ [J]. 化工进展, 2023, **42**(6): 3226–3232.
Lü C, Zhang X W, Jin L J, et al. Efficient capture of CO₂ by a new biphasic solvent–ionic liquid system[J]. Chemical Industry and Engineering Progress, 2023, **42**(6): 3226–3232.