文章编号：1006-3080(2025)06-0793-11 DOI：10.14135/j.cnki.1006-3080.20250121001

基于机器学习的酸烃界面预测与离子液体设计

田一凡[1]， 高维群[2,3]， 卢静宜[1]， 郑伟中[2,3]， 孙伟振[2,3]
（华东理工大学 1. 信息科学与工程学院； 2. 化学工程与低碳技术全国重点实验室；
3. 化工学院，上海 200237）

摘要：为了精准预测离子液体对酸烃界面性质的影响并设计出新颖的离子液体组合，本文采用随机森林等机器学习方法，分析离子液体与酸烃界面性质之间的内在联系，并基于不同的描述符构建了酸烃界面性质的预测模型。为了合理设计新型的离子液体组合，将离子液体的SMILES编码转换为连续且数据驱动的分子描述符(Continuous and Data-Driven Molecular Descriptors，CDDD)，并利用基于成功历史的自适应差分进化(Success-History Based Adaptive Differential Evolution，SHADE)算法在潜在空间中进行精准搜索和解码，同时，结合子结构约束以确保生成结构的合理性。以烷基化反应中的界面厚度(δ_w)和界面张力(γ)为例，构建的预测模型在测试集上对界面厚度的决定系数为0.952，界面张力的决定系数为0.901，显示出较高的预测精度。此外，通过SHADE算法优化设计，成功生成了328对满足界面厚度和界面张力要求的离子液体组合，显著扩展了符合标准的离子液体组合的可行域。本文研究不仅提高了对酸烃界面性质的预测能力，还为离子液体的合理设计提供了新的方法和思路。

关键词：离子液体；机器学习；酸烃界面性质；合理设计；进化算法

中图分类号：TP181 **文献标志码：**A

C4烷基化反应是异丁烷与C3~C5烯烃在浓硫酸为催化剂条件下生成C8烷基化油的反应[1]。研究表明[2]离子液体(Ionic Liquids, ILs)作为C4烷基化反应的潜在添加剂，能够显著增加浓硫酸与C4烃之间的界面厚度，有效降低界面张力，提高界面区域反应物浓度，进而显著提升烷基化油产品质量。由于阳离子和阴离子的多种组合，离子液体表现出多变的行为[3]，这使得在C4烷基化反应中选择表现优异的离子液体组合变得困难。因此，研究不同离子液体组合及其对应的酸烃界面性质对于筛选高性能的离子液体组合至关重要。通过设计离子液体组合实现反应的精细调控，可以推动反应向更高效和环保的方向发展。

以硫酸为催化剂的C4烷基化反应中，界面厚度(δ_w)、界面张力(γ)等界面性质的评估始终是研究的重点之一。如Dokoohaki等[4]利用密度泛函理论(Density Functional Theory, DFT)[5]、分子动力学模拟(Molecular Dynamics, MD)[6]等基于理论计算的方法获取界面性质数值，然而这些方法往往需要消耗大量的计算资源。近年来，越来越多的研究者采用基团贡献法[7]和真实溶剂似导体屏蔽模型(Conductor-like Screening Model for Real Solvents, COSMO-RS)[8]等经验模型方法来预测酸烃界面性质。这些经验模型方法在计算效率上具有显著优势，但在处理非平衡状态的性质以及应用范围方面仍存在一定的局限性。

随着人工智能技术的不断发展，机器学习(ML)方法被广泛应用于离子液体性质预测与高性能离子液体筛选。Zheng等[9]基于DFT计算离子液体的特征，构建的机器学习模型高效地预测了11 000组离子液体对酸烃界面性质的影响。除了界面性质，机

收稿日期：2025-01-21
基金项目：科技部科技创新2030人工智能重大专项(2023ZD0121001)；中央高校基础科研业务费支持
作者简介：田一凡(2000—)，男，江西人，硕士生，主要从事新材料逆向设计研究。E-mail: 1179460951@qq.com
通信联系人：卢静宜，E-mail: jylu_cise@ecust.edu.cn
引用本文：田一凡，高维群，卢静宜，等. 基于机器学习的酸烃界面预测与离子液体设计 [J]. 华东理工大学学报(自然科学版)，2025，51(6)：793-803.
Citation: TIAN Yifan, GAO Weiqun, LU Jingyi, *et al*. Machine Learning-Based Acid-Hydrocarbon Interface Prediction and Ionic Liquids Design[J]. Journal of East China University of Science and Technology, 2025, 51(6): 793-803.

器学习也被用于预测离子液体的其他性质，如生物毒性和熔点。Cao 等[10] 与 Low 等[11] 利用量子力学计算离子液体的结构特征，并结合机器学习构建预测模型，以预测其生物毒性和熔点。Kuroki 等[12] 则通过量子化学和机器学习筛选出具有高 CO_2 溶解度的离子液体。这些研究通过结合传统化学计算方法和机器学习算法，实现高效率、高精度的离子液体性质预测。然而，在利用传统化学计算方法计算分子结构特征时，研究者仍需消耗大量计算资源。此外，这种将传统化学计算方法与机器学习算法相结合的方法主要依赖已知结构来预测性质，适用于高通量筛选，但在逆向设计中存在局限性。逆向设计旨在从期望性质出发生成相应结构，而目前计算特征与结构之间缺乏直接映射，导致逆向设计面临挑战。因此，尽管机器学习在预测离子液体性质方面取得了进展，但在逆向设计方面仍需进一步研究。

针对逆向设计问题，传统的计算模拟和高通量筛选方法[13-14] 虽然能够有效指导化学空间的探索，但这些方法通常需要高昂的计算成本，并且依赖于专家的知识和经验。一种新兴的策略是计算机辅助分子设计，特别是结合启发式方法从头设计具有理想性质的新分子，避免了数据库的大规模访问。Reutlinger 等[15] 的研究展示了计算机辅助药物设计中的成功应用，通过构建模块选择和属性优化，成功合成了新型生物活性化合物。Dey 等[16] 和 Yuan 等[17] 的工作也展示了在蛋白质抑制剂设计和配体优化中的创新实践，尽管这些研究主要集中在生物分子领域，但其原理为离子液体的逆向设计提供了宝贵启示。Winter 等[18] 将分子特性的预测与粒子群优化(Particle Swarm Optimization, PSO)算法相结合，在连续潜在空间中进行搜索，实现了高效的分子优化。在此基础上，为了进一步提高搜索效率，许多学者也开始广泛采用基于成功历史的自适应进化(Success-History Based Adaptive Differential Evolution, SHADE)算法，在高维空间中进行搜索[19-20]。然而，在离子液体的反向设计中，相较于药物设计，数据的稀缺性和质量问题尤为显著。此外，离子液体组合的多样性和结构不稳定性导致在高维空间中容易搜索到不合理的离子结构，这使得传统方法在生成合理的化学结构时面临诸多挑战。为应对这些挑战，结合物理化学知识和先验信息对搜索空间进行约束成为一种可行策略，但其效果受限于专家经验的广度和深度[21-22]。另一种途径是利用大规模数据和先进的机器学习技术进行无监督学习，以捕捉分子结构的内在规律。然而，生成的分子中仍可能出现不合理的化学结构[23-25]，但这种方法同样面临数据获取和计算成本的瓶颈。

本文首先探讨了不同离子液体组合与酸烃界面性质之间的构效关系；接着，筛选出多种描述符，并基于这些描述符构建了用于预测酸烃界面性质的模型；通过对各模型预测性能进行分析，选择能精确表征离子液体结构的描述符构建预测模型。为了反向设计具有强化特定界面性质的新型离子液体组合，将离子液体的 SMILES 编码转换为连续且数据驱动的分子描述符(Continuous and Data-driven Molecular Descriptors, CDDD)，并基于多种机器学习模型构建预测模型，通过对比分析筛选出最优模型。该模型用于评估在 CDDD 映射的潜在空间中候选解的界面性质。基于预测模型，采用 SHADE 优化算法在该潜在空间中进行搜索，同时利用开源化学信息学工具包 RDKit 制定子结构约束，以避免生成不合理结构，确保设计过程集中在具有可靠结构的离子液体组合上。通过上述方法，实现对离子液体强化的酸烃界面性质更精确的预测，并成功地生成了一系列新颖的离子液体候选组合。

1 酸烃界面性质预测及离子液体合理设计

首先介绍不同的分子描述符，比较不同描述符构建预测模型的性能。在此基础上，设计一种基于优化算法的离子液体合理性设计的方法，以推导出具有强化界面性质的离子液体组合。整体框架如图 1 所示。其中 RF 为随机森林(Random Forest)，SVR 为支持向量回归(Support Vector Regression)，GBM 为梯度提升机(Gradient Boosting Machine)。

将阴离子和阳离子 SMILES 编码维度为 n 的连续潜在空间均表示为 L，并以 M^- 和 M^+ 分别表示阴离子和阳离子。具体来说，用编码器将阴离子 M^- 映射为 $\boldsymbol{x}^-$，阳离子 M^+ 映射为 $\boldsymbol{x}^+$，即：

$$\boldsymbol{x}^- = \mathrm{Encoder}(M^-) \tag{1}$$

$$\boldsymbol{x}^+ = \mathrm{Encoder}(M^+) \tag{2}$$

定义 $\boldsymbol{x} = [(\boldsymbol{x}^-)^{\mathrm{T}}, (\boldsymbol{x}^+)^{\mathrm{T}}]^{\mathrm{T}}$，$\boldsymbol{x}$ 就是预测和优化任务的决策变量。

1.1 离子液体的描述符

1.1.1 分子指纹　在离子液体强化的酸烃界面性质预测任务中，将离子液体的 SMILES 编码转换为分子指纹(Molecular Fingerprint, FP)[26]。图 2 所示方法利用 RDKit 中的 Chem.MolFromSmiles函数，将离子液体中的阴离子和阳离子分别解析成一个 Mol 对象。

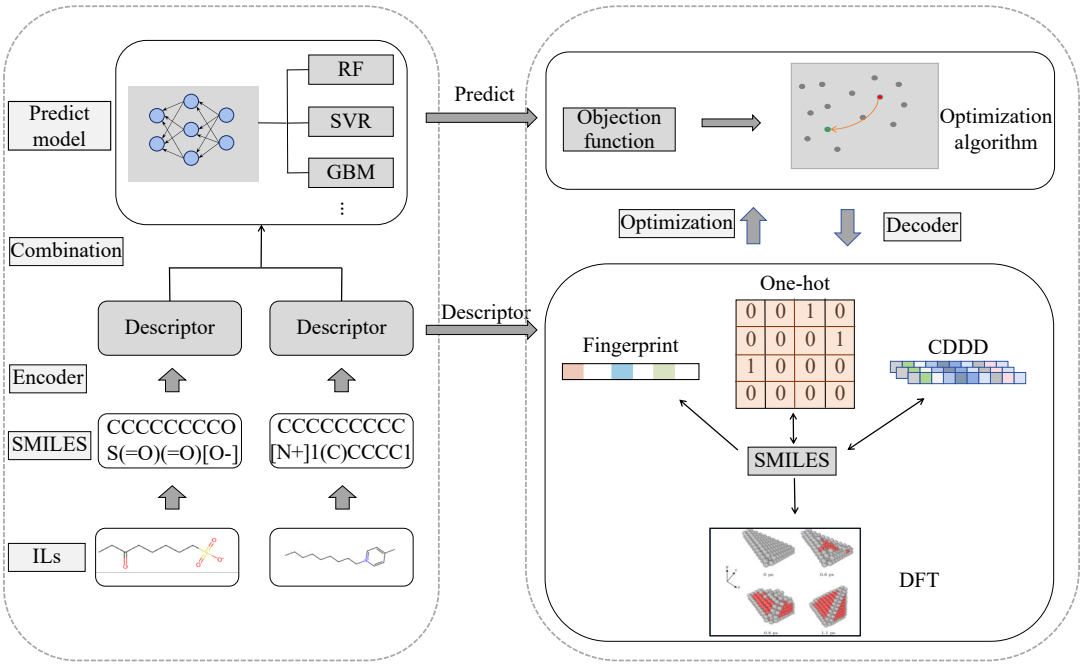


图 1　离子液体强化酸烃界面性质预测 (a) 和合理性设计 (b) 的框架

Fig. 1　Framework for ionic liquids enhanced acid-hydrocarbon interfacial property prediction(a) and reasonable design (b)

然后, 利用 AllChem 中的函数计算 Mol 对象的 Morgan 指纹, 将计算得到的阴、阳离子的指纹对象分别进一步转换为二进制字符串形式 $\boldsymbol{x}^-$ 和 $\boldsymbol{x}^+$, 并拼接为 $\boldsymbol{x}=[(\boldsymbol{x}^-)^{\mathrm{T}},(\boldsymbol{x}^+)^{\mathrm{T}}]^{\mathrm{T}}$。分子指纹具有较高的计算效率, 适合处理大规模数据集, 且能够部分保留分子的结构信息, 从而在一定程度上还原分子结构。然而, 它通常只能提供分子的局部结构信息, 对于复杂的分子结构描述不够精确, 且在处理结构相似的分子时, 其泛化能力不足。

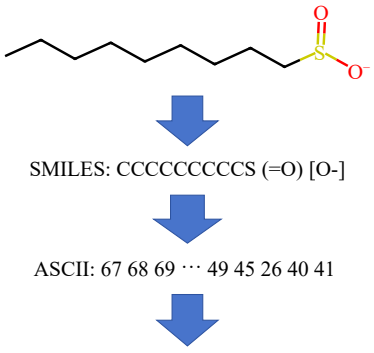


图 2　计算分子指纹

Fig. 2　Calculate molecular fingerprints

1.1.2　独热编码　为了将数据集中的 SMILES 进行独热(One-hot)编码[27], 创建了字符到索引的映射字典, 其中包含所有出现在数据集中离子液体的 SMILES 字符串中的字符, 及其对应的唯一索引。将 SMILES 字符串分解为单个字符, 并利用映射字典分别将阴、阳离子的 SMILES 中每个字符转换为与其对应的 One-hot 编码向量 $\boldsymbol{x}^-$ 和 $\boldsymbol{x}^+$, 拼接为 $\boldsymbol{x}=[(\boldsymbol{x}^-)^{\mathrm{T}},(\boldsymbol{x}^+)^{\mathrm{T}}]^{\mathrm{T}}$, 并且可以根据映射字典找到对应的离子液体。然而, 机器学习模型通常要求输入的数据具有统一的长度, 因此, 需要对不同长度的 SMILES字符串进行处理。将原始 SMILES 字符串中的每个字符转换为 one-hot 编码向量, 对于长度小于设定最大长度的向量, 在其末尾填充零, 直至达到最大长度。这一操作确保了所有输入向量的维度一致, 从而满足了机器学习模型对输入数据长度统一的要求。One-hot 编码可以直接从 SMILES 字符串生成, 并且能够完全还原分子结构。然而, One-hot 编码也存在明显的缺点, 对于长的 SMILES 字符串, One-hot 编码会导致输入特征维度大幅增加, 从而增加计算复杂度。

1.1.3　CDDD　以 Winter 等[28] 的编码器-解码器结构为基础, 将离子液体组合转换为潜在向量 $\boldsymbol{x}$。具体而言, 编码器利用离子液体的 SMILES 描述符, 将其映射到潜在空间, 通过简化数据结构并提取关键特征, 生成潜在向量。这些潜在向量可以通过解码器映射回对应的离子液体组合。编码器负责将分子表示转换为潜在向量, 而解码器则负责将这些潜在向量重构为分子表示。本文将编码器 $E(M)$ 定义为一个将分子 SMIELS 描述符 M 映射到潜在空间 $L\in\mathbb{R}^n$ 的函数, 其输出为潜在向量 $\boldsymbol{x}$。$\boldsymbol{x}$ 可通过解码器 $D(M)$ 解码回相应的 M。因此, 将阴离子 M^- 和阳离子 M^+ 分别编码为 $\boldsymbol{x}^-$ 和 $\boldsymbol{x}^+$, 并拼接为 $\boldsymbol{x}=[(\boldsymbol{x}^-)^{\mathrm{T}},(\boldsymbol{x}^+)^{\mathrm{T}}]^{\mathrm{T}}$。

CDDD通过数据驱动的方法生成描述符,具有较高的计算效率。它能够从潜在空间映射回具体的分子结构,这使得CDDD在逆向设计中具有显著优势。此外,CDDD能够捕捉分子结构的内在规律,从而具有较强的泛化能力,能够更好地处理具有复杂结构的分子。

1.2 离子液体结构-酸烃界面性质建模

使用来自Zheng等[9]的离子液体数据集构建离子液体-酸烃界面性质预测模型,数据集包含1 200组基于DFT的离子液体结构特征和对应的酸烃界面厚度和界面张力数据。首先将数据集中的界面厚度和界面张力进行归一化处理,具体是通过助剂体系的界面厚度和界面张力与它们的纯硫酸体系的界面厚度(δ_0)和界面张力(γ_0)的比值来实现,得到归一化后的界面厚度(δ_w/δ_0)和界面张力(γ/γ_0)。将离子液体数据集定义为 $S=(M_i^-,M_i^+,y_i^{\delta_w},y_i^{\gamma})$,$i=1\cdots N$,其中 N 为数据量;$y_i^{\delta_w}$ 和 y_i^{γ} 分别为归一化处理后的界面厚度和界面张力。将数据集中的离子液体SMILES编码通过1.1节的方法得到不同描述符的模型训练数据集 $D=(\boldsymbol{x}_i^-,\boldsymbol{x}_i^+,y_i^{\delta_w},y_i^{\gamma})$,其中 $\boldsymbol{x}_i=[(\boldsymbol{x}_i^-)^{\mathrm{T}},(\boldsymbol{x}_i^+)^{\mathrm{T}}]^{\mathrm{T}}$ 作为模型的输入特征向量,$y_i^{\delta_w}$ 和 y_i^{γ} 分别作为模型的目标值。用 $F^{\delta_w}(\boldsymbol{x})$ 和 $F^{\gamma}(\boldsymbol{x})$ 分别表示界面厚度和界面张力性质预测模型。由于数据集规模较小,使用随机森林[29]构建预测模型。随机森林回归模型的训练可以描述为基于离子液体的训练集 $D=(\boldsymbol{x}_i,\boldsymbol{y}_i)$,其中 $\boldsymbol{y}_i=(y_i^{\delta_w},y_i^{\gamma})$。随机森林回归模型由 T 个决策树组成,每个决策树 t 都是一个基本的回归树。对于给定的输入特征向量 $\boldsymbol{x}_i$ 和目标向量 $\boldsymbol{y}_i$,随机森林训练过程可以表示为:

$$F_t(\boldsymbol{x})=\sum_{j=1}^{j_t}c_{j_t}\cdot I(x\in R_{j_t}) \tag{3}$$

$$F(\boldsymbol{x})=\frac{1}{T}\sum_{t=1}^{T}F_t(\boldsymbol{x}) \tag{4}$$

其中:j_t 是决策树 t 的叶子节点数,R_{j_t} 是第 j 个叶子节点的区域,c_{j_t} 是该叶子节点的预测值,I 是指示函数。$F_t(\boldsymbol{x})$ 表示第 t 个基本回归树对输入特征向量 $\boldsymbol{x}$ 的预测值,随机森林的预测值 $F(\boldsymbol{x})$ 由所有基本回归树的预测值的平均值得到。模型在训练过程中,主要通过不断调整决策树数量等超参数,最小化均方误差(MSE)损失函数,从而使得模型的预测值 $F(\boldsymbol{x})$ 逐渐接近真实值,提高预测准确性。

为了全面评估构建的机器学习模型的预测性能,通常采用一些关键指标对模型进行评估,如相关系数(R^2)、平均绝对误差(MAE)和均方根误差(RMSE),其计算式分别如下:

$$R^2=1-\frac{\sum_{i=1}^{N}(y_i-\widehat{y_i})^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2} \tag{5}$$

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}|\widehat{y_i}-y_i| \tag{6}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{y_i}-y_i)^2} \tag{7}$$

其中,y_i 是真实值,$\widehat{y_i}$ 是预测值,$\bar{y}$ 是真实值的平均值,N 是数据量。

1.3 基于SHADE优化算法的离子液体合理性设计

基于SHADE算法实现在潜在空间中搜索满足目标性质且结构合理的离子液体组合。为了确保生成的离子液体结构合理,优化过程中使用RDKit对候选解的阴、阳离子结构进行分析,并对不合理结构施加惩罚。惩罚分为3类:当RDKit分析候选解结构为非分子结构时,输出惩罚值为2($p^-=p^+=2$);当候选解为合理结构(或目标结构)时,输出惩罚值为0($p^-=p^+=0$);其他情况时输出惩罚值为1($p^-=p^+=1$)。其中:p^- 和 p^+ 分别为RDKit对阴离子结构和阳离子结构的分类输出。为了加速在庞大的潜在空间中的搜索,选择几组性能较优的数据作为基点,并用其来引导搜索方向。通过不断缩小候选种群与优秀解之间的欧氏距离,使候选种群逐渐逼近优秀解的结构和性质。在优化目标函数设计方面,将多个目标函数通过权重转化为一个单目标任务,从而使离子液体组合能够在不同目标间取得平衡。定义该非线性优化问题候选解为 $\boldsymbol{x}\in\mathbb{R}^n$($n$ 表示向量的维度),该多目标的非线性优化问题可以描述为在可行域 L 中,寻找令目标函数最小的决策向量 $\boldsymbol{x}$,即:

$$\min_{x\in\mathbb{R}^{2n}} w_{\mathrm{p}}f_{\mathrm{p}}(\boldsymbol{x})+w_{\mathrm{s}}f_{\mathrm{s}}(\mathrm{Decoder}(\boldsymbol{x}))+w_{\mathrm{d}}f_{\mathrm{d}}(\boldsymbol{x}) \tag{8}$$

$$\mathrm{s.t.}\,f_{\mathrm{p}}(\boldsymbol{x})=|F^{\gamma}(\boldsymbol{x})-y_{\mathrm{exp}}^{\gamma}|-F^{\delta_w}(\boldsymbol{x}) \tag{9}$$

$$f_{\mathrm{s}}(\mathrm{Decoder}(\boldsymbol{x}))=f_{\mathrm{s}}(\mathrm{Decoder}(\boldsymbol{x}^-))+f_{\mathrm{s}}(\mathrm{Decoder}(\boldsymbol{x}^+))=p^-+p^+ \tag{10}$$

$$f_{\mathrm{d}}(\boldsymbol{x})=\sum_{z=1}^{m}\|\boldsymbol{x}-\boldsymbol{x}_z\| \tag{11}$$

其中,$f_{\mathrm{p}}(\boldsymbol{x})$ 与离子液体强化的酸烃界面性质有关,使优化过程中候选解强化的酸烃界面性质逼近预期界面性质 $y_{\mathrm{exp}}^{\gamma}$;$f_{\mathrm{s}}(\mathrm{Decoder}(\boldsymbol{x}))$ 与结构合理性有关,其中 $\mathrm{Decoder}(\boldsymbol{x})$ 表示与潜在向量对应的离子结构;$f_{\mathrm{d}}(\boldsymbol{x})$ 与离子液体优化方向有关,期望决策变量 $\boldsymbol{x}$ 趋

向数据集中已存在的 m 个优秀解 $\boldsymbol{x}_z$；w_p、w_s 和 w_d 分别为对应函数的权重系数。

基于 SHADE 算法在潜在空间的搜索过程由 5 个主要过程组成, 即: 初始化-变异-交叉-选择-参数更新。在初始化阶段, 初始化 p 个群体 (离子液体组合), 每个个体在潜在空间 L 中随机生成。使用 $x^n_{\min}$、$x^n_{\max}$ 表示个体的每个维度的最小值和最大值。用 $x^n_{i,j}$ 表示第 i 个体(i_{th})的第 j 维度(j_{th})位置, 搜索空间 $U(x^n_{\min}, x^n_{\max})$ 满足随机生成的均匀分布 $x^n_{i,j}\ U(x^n_{\min}, x^n_{\max})$, 以确保每个个体的每个维度的值都在该范围内。

在变异阶段, 用 $x^n_{i,g}$ 表示 g 代中群体的 i_{th} 个体的位置。在每个 g 代中, 随机选择的互不相似的个体通过差异突变产生突变个体 $v_{i,g}$ 即:

$$v_{i,g} = x_{r1,g} + u_F(x_{r2,g} - x_{r3,g}) \tag{12}$$

其中, $x_{r1,g}$, $x_{r2,g}$, $x_{r3,g}$ 表示在 g_{th} 代群体中随机选择的互不相同的个体, 而 u_F 表示比例因子。

在交叉阶段, 基于个体的随机指数(j_{rand})和交叉概率(u_{CR})生成测试个体 $u_{i,g}$。如式 (13) 所示, 当 $j_{\text{rand}} \leqslant u_{\text{CR}}$ 时, 或者当前维度 $j = j_{\text{rand}}$ (至少保证有一个维度交叉), 则选择变体个体 $v_{i,g}$ 的 j_{th} 分量($v_{i,j,g}$)作为测试个体 $u_{i,g}$ 的 j_{th} 分量($u_{i,j,g}$), 在其他情况下则保留当前个体 $x_{i,g}$ 的 j_{th} 分量($x_{i,j,g}$)。即:

$$u_{i,j,g} = \begin{cases} v_{i,j,g}, & j_{\text{rand}} \leqslant u_{\text{CR}} \text{ or } j = j_{\text{rand}} \\ x_{i,j,g}, & \text{Otherwise} \end{cases} \tag{13}$$

在选择阶段, 基于 RDKit 分析候选种群解的离子结构并得到对应的结构约束惩罚值, 结合适应度函数($O(x)$)选择是否接受测试个体。当测试个体 $u_{i,g}$ 的适应度 $O(u_{i,g})$ 小于当前最小适应度 $O(x_{i,g})$ 时, 则保留该测试个体, 即:

$$x_{i,g+1} = \begin{cases} u_{i,g}, & O(u_{i,g}) \leqslant O(x_{i,g}) \\ x_{i,g}, & \text{Otherwise} \end{cases} \tag{14}$$

如果测试个体 $u_{i,g}$ 被接受, 则将相应的 u_F 和 u_{CR} 进行存储。

在参数更新阶段, 基于历史均值进行正态分布采样来调整 u_F 和 u_{CR}:

$$u_F \sim \mathcal{N}(\overline{u}_F, \text{std}),\ u_{\text{CR}} \sim \mathcal{N}(\overline{u}_{\text{CR}}, \text{std}) \tag{15}$$

其中: std 通常取 0.1, 参数 u_F 和 u_{CR} 根据其历史均值进行正态分布采样, 以适应新的搜索方向。

2　结果与分析

2.1　离子液体与酸烃界面性质分析

为了解离子液体对酸烃界面性质的影响, 分析离子液体结构和界面性质 (δ_w 和 γ) 的关系。图 3 示出了不同离子液体组合对应的归一化后界面厚度 (δ_w/δ_0) 和归一化界面张力 (γ/γ_0), 数据集离子液体最终包含 30 种阴离子和 40 种阳离子, 共有 1 200 对离子液体组合。不同的离子液体组合对于界面厚度与界面张力的影响差异显著。绝大多数离子液体可以加宽 δ_w 并降低 γ, 这意味着离子液体添加剂可以为硫酸烷基化反应提供更大的反应面积和更好的两相

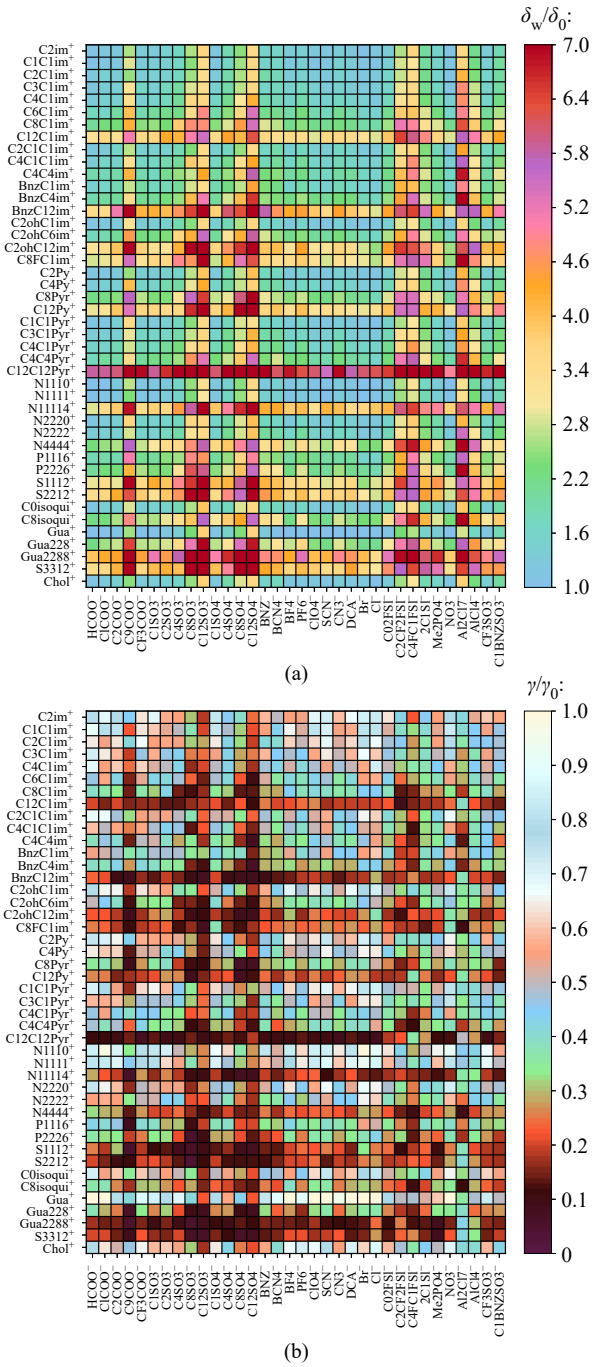


图 3　不同离子液体归一化界面厚度 (a) 和归一化界面张力 (b)

Fig. 3　Normalized interfacial thickness (a) and normalized interfacial tension (b) of different ionic liquids

分散性,有助于提升烷基化产物的质量;具有较短烷基链的阳离子与较小尺寸的阴离子会导致 δ_w 降低,γ 升高;总的来说,δ_w 和 γ 呈现负相关的规律。

2.2 酸烃界面性质预测

为了预测离子液体对酸烃界面性质的影响,构建基于多种描述符的机器学习模型,将离子液体的结构特征与酸烃界面行为相关联。所选描述符包括DFT、FP、One-hot以及CDDD,作为输入特征用于模型训练。从数据集中随机抽取75%的样本作为训练集,剩余25%作为测试集,以确保模型能在充足的训练数据上学习,并通过独立测试集验证其泛化能力。

为准确预测界面厚度和界面张力这两个关键性质,分别构建了独立的预测模型。选择RF算法作为基础模型,因为其在处理复杂非线性关系方面表现优越,并对特征选择具有较高鲁棒性。在模型训练过程中,计算相关系数 R^2 评估模型的拟合效果,R^2 越接近于1,表明模型的预测能力越强。图4直观地展示了在不同描述符情况下模型的预测性能,为进一步探讨各种描述符对模型预测性能的影响提供了依据。为了确保模型评估结果的稳健性与可靠性,采用基于ShuffleSplit的交叉验证方法。该方法每次将数据集划分为75%的训练集和25%的测试集,共进行10次随机划分。与仅进行1次随机划分的方法相比,ShuffleSplit交叉验证方法能够更全面地评估模型性能,有效减少因数据划分随机性所导致的偏差,这一特性使其特别适用于数据量较小的情况,可充分利用有限的数据资源,从而为模型的性能评估提供更为准确的依据。表1中包含了交叉验证相关系数(R^2_{CV})、平均绝对误差和均方根误差等关键指标的数值,以比较基于不同描述符的模型预测性

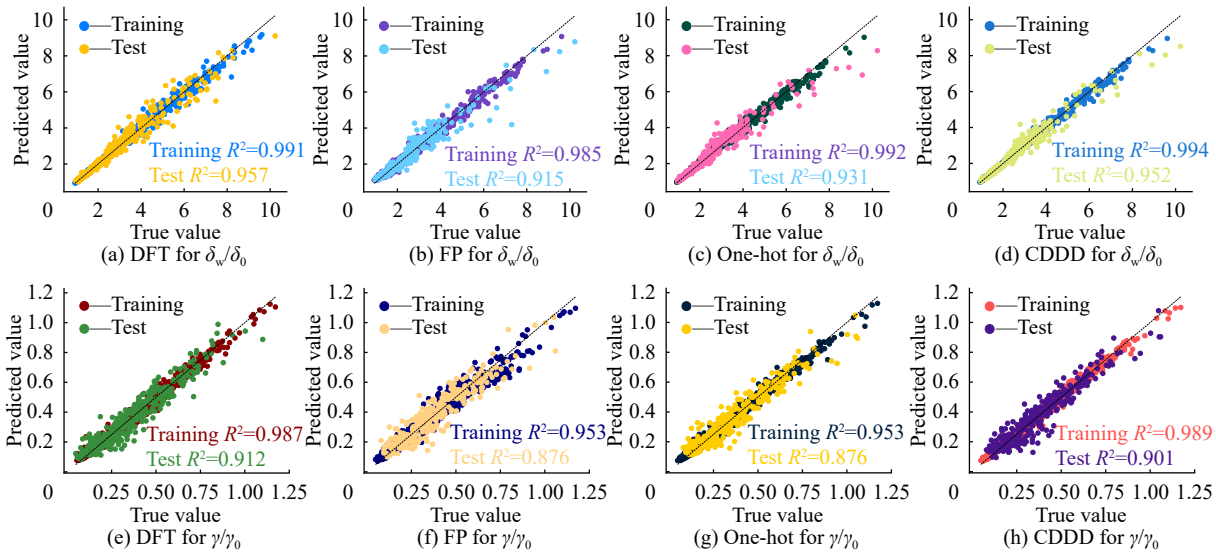


图4 不同界面性质基于4种描述符的训练和测试数据的相关性

Fig. 4 Correlation plots of different interfacial properties based on training and test data from four descriptors

表1 不同界面性质基于4种描述符的训练和测试数据的预测结果

Table 1 Prediction results of different interfacial properties based on training and test data from four descriptors

Data	Descriptor	δ_w/δ_0			γ/γ_0		
		R^2_{CV}	MAE	RMSE	R^2_{CV}	MAE	RMSE
Training	DFT	0.946	0.093	0.157	0.904	0.016	0.023
	FP	0.934	0.151	0.198	0.890	0.032	0.044
	One-hot	0.927	0.093	0.149	0.896	0.018	0.023
	CDDD	0.952	0.077	0.123	0.908	0.016	0.022
Test	DFT	—	0.219	0.343	—	0.044	0.057
	FP	—	0.268	0.436	—	0.051	0.067
	One-hot	—	0.228	0.380	—	0.044	0.058
	CDDD	—	0.200	0.333	—	0.046	0.060

能。计算这些评估指标所用的界面厚度和界面张力数据均为归一化后的数据，归一化处理有助于消除不同量纲和数量级对模型评估的影响，使模型的性能评估更加公平和准确。表 1 结果表明，CDDD、DFT、FP 和 One-hot 描述符在预测酸烃界面性质时各有优缺点。DFT 描述符虽精确但计算成本高，且难以直接映射回具体的离子结构，限制了其在逆向设计中的应用。FP 描述符计算效率高且可逆性好，但精确性和泛化能力有限，难以处理复杂分子结构。One-hot 描述符存在维度爆炸问题，影响计算效率和模型泛化能力。相比之下，CDDD 描述符不仅计算效率高、可逆性强，还能捕捉分子结构的内在规律，具有较强的泛化能力，适用于复杂数据集的处理和逆向设计。因此，CDDD 描述符成为本研究中预测和优化任务的理想选择。

基于 CDDD 描述符，构建 4 种机器学习（ML）预测模型，即支持向量机（SVM）、GBM、RF 和多层感知机（MLP），这些模型在训练和测试数据上的表现通过 R^2 值的相关性图进行了展示，如图 5 所示，R^2_{CV}、MAE 和 RMSE 列于表 2 中。结果表明，GBM 和 RF 模型在关联离子液体的结构特征与酸烃界面厚度和界面张力方面表现良好，这表明它们能够有效地捕捉离子结构与酸烃界面性质之间的关联，并为预测结果提供了较高的准确性。

2.3　离子液体合理性设计

本文对数据集中的离子液体强化的酸烃界面性质设置了严格的筛选标准：$\delta_w/\delta_0 > 6$ 和 $\gamma/\gamma_0 < 0.1$，该判据可确保筛选出的离子液体添加剂能显著地增宽界面厚度和降低界面张力。在此基础上，在初始数据集中的 1200 对离子液体组合中仅能筛选出

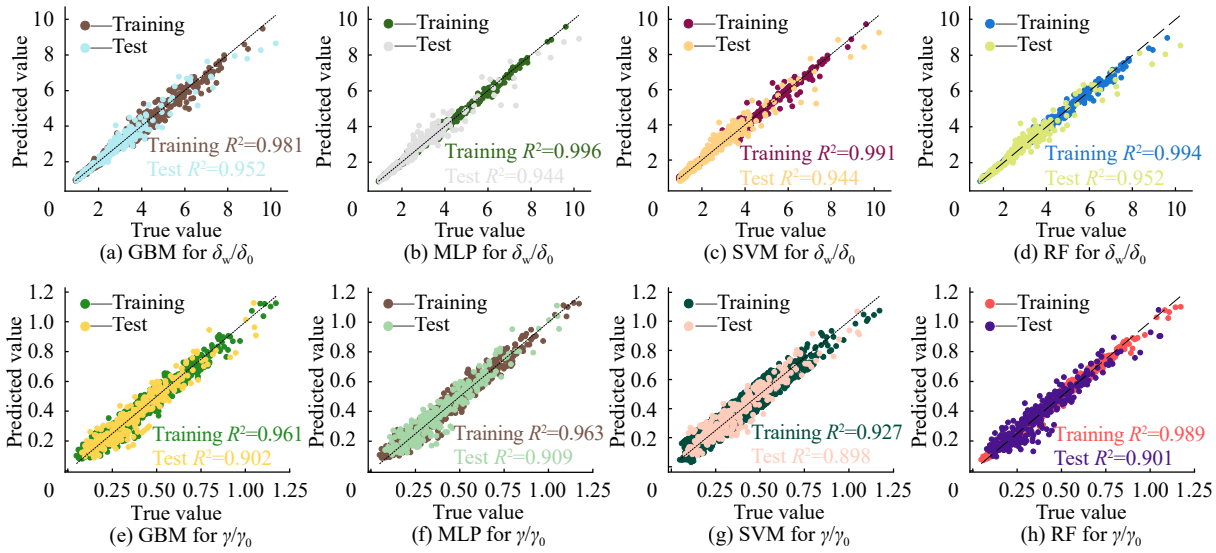


图 5　不同界面性质基于 4 种机器学习模型的训练和测试数据的相关性

Fig. 5　Correlation plots of different interfacial properties based on training and test data from four ML models

表 2　不同界面性质基于 4 种机器学习模型训练和测试数据的预测结果

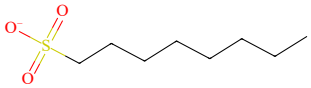
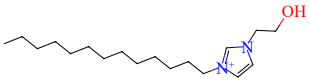
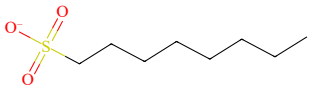
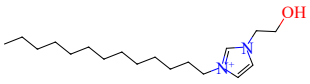
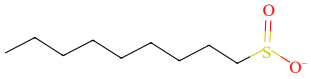
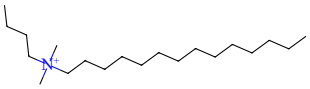
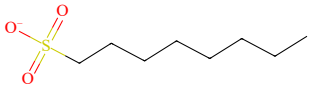
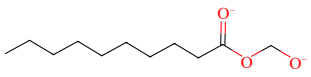
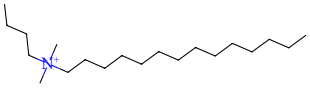
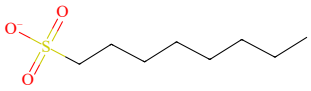
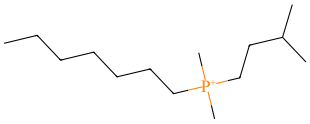
Table 2　Prediction results of different interfacial properties based on training and test data from four ML models

Data	Model	δ_w/δ_0			γ/γ_0		
		R^2_{CV}	MAE	RMSE	R^2_{CV}	MAE	RMSE
Training	GBM	0.955	0.136	0.212	0.923	0.030	0.039
	MLP	0.963	0.067	0.097	0.922	0.031	0.039
	SVM	0.954	0.100	0.149	0.902	0.045	0.055
	RF	0.952	0.077	0.123	0.908	0.016	0.022
Test	GBM	—	0.196	0.317	—	0.046	0.059
	MLP	—	0.210	0.345	—	0.044	0.057
	SVM	—	0.227	0.345	—	0.048	0.061
	RF	—	0.200	0.333	—	0.046	0.060

33对符合标准的离子液体组合。为了扩展符合条件的离子液体组合，采用了离子液体的合理化设计。为此，随机初始化500个种群作为算法的初始解，并采用SHADE算法进行迭代优化。这种方法增强了搜索效率，最终优化结果如表3所示，成功地生成了多对符合标准的离子液体组合，这些组合在现有数据集中未被记录，显著扩展了符合标准的离子液体组合的可行域。

表3 SHADE算法在潜在空间生成的离子液体及其预测的界面性质

Table 3 Ionic liquids generated in potential space by the SHADE algorithm and their predicted interface properties

Anion	Cation	δ_w/δ_0	γ/γ_0
		7.332	0.0865
		7.001	0.0774
		6.728	0.0990
		6.445	0.0700
		6.425	0.0968
		6.377	0.0694

为了直观地展示基于SHADE算法的优化方法的优化效率，将基于SHADE算法的优化方法与Winter等[18]基于PSO算法的优化方法进行比较，结果如图6所示。结果表明随着迭代次数的增加，PSO更容易陷入局部最优，SHADE算法能够搜索到更优的离子液体。

为了验证生成的离子液体对酸烃界面性质的影响是否符合已知的界面厚度和界面张力关系，进行了一系列的对比分析。首先，如图7所示，将生成的离子液体强化的酸烃界面性质与数据集中已有的离子液体强化的酸烃界面性质进行对比，对生成离子液体强化的酸烃界面性质关系进行了多项式拟合，并对关键区域（$\delta_w/\delta_0>6.0$ 和 $\gamma/\gamma_0<0.10$）进行了放大观察。结果表明，生成的离子液体强化的酸烃界面性质关系与已有数据集离子液体强化的酸烃界面性质关系规律相吻合，特别是在相同的界面张力条件下，这些离子液体的添加能有效增加界面厚度，验证了方法的有效性。

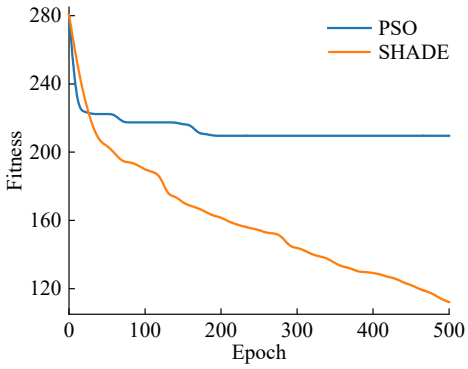


图6 优化过程中不同方法适应度值的对比

Fig. 6 Comparison of fitness values among different methods during the optimization process

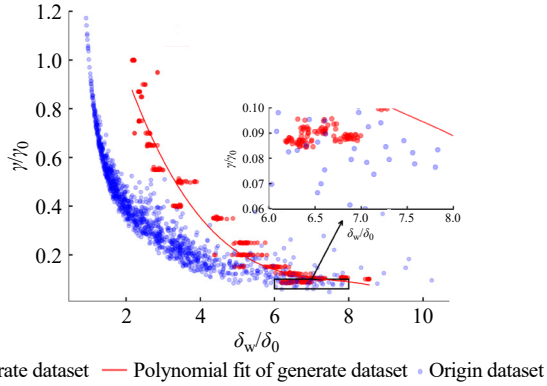


图7 生成集和数据集界面性质关系

Fig. 7 Relationship between the interface property of the generated sets and datasets

为了可视化模型生成的离子液体与原始数据集中离子液体之间的结构关系, 进一步采用加权整体不变分子(Weighted Holistic Invariant Molecular, WHIM)描述符。WHIM 描述符是一种包含分子大小、形状、电荷等信息的 114 维向量[30], 能够全面描述分子特性。通过 RDKit 软件计算这些描述符, 捕捉了分子结构的关键信息。使用主成分分析(Principal Components Analysis, PCA)对这些高维数据进行降维处理, 以便于分析和可视化。图 8 示出了基于本文条件下生成的 328 对离子液体组合以及数据集中符合标准的 33 对实际的离子液体组合的 PCA 分析结果。对图中红色密集部分的数据进行了局部放大, 以便于详细观察。从 PCA 结果中可以看出, 生成的离子液体在结构和性质上与实际的离子液体紧密相连, 散布在其周围。这种分布情况表明, 本文方法不仅能够在宏观上生成具有强化已知界面性质的离子液体, 也能在微观结构上保持与实际离子液体高度的相似性。

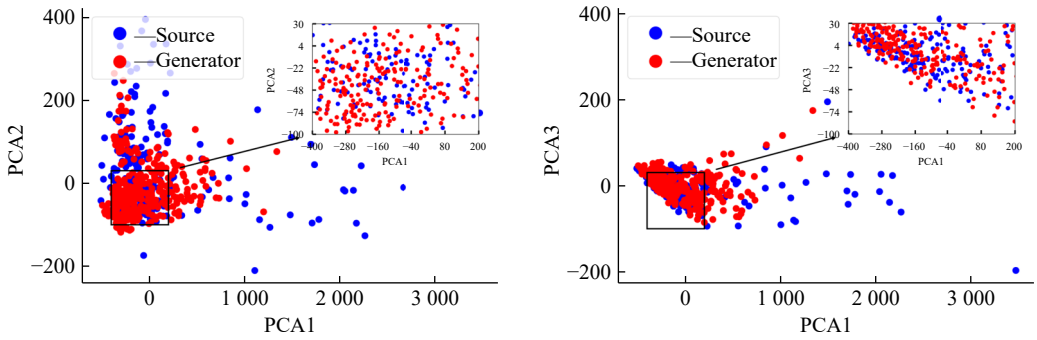


图 8　生成离子液体与数据集离子液体在 WHIM 描述符空间中的 PCA 分析(PCA1、PCA2 和 PCA3 分别代表前 3 个主成分轴)

Fig. 8　PCA analysis of generated ionic liquids and dataset ionic liquids in the WHIM descriptor space (PCA1, PCA2 and PCA3 represent the first three principal component axes, respectively)

3　结束语

本文提出了一种融合机器学习与进化优化算法的离子液体设计方法, 实现了酸烃界面性质的高效预测与具有酸烃界面性质增强能力的离子结构生成。通过系统比较离子表征方式, 选取 CDDD 嵌入向量作为离子结构输入, 结合 RF 模型构建前向预测框架, 实现了从结构到界面性质的精准映射。进一步借助 SHADE 算法对潜在设计空间进行优化, 并利用 RDKit 进行约束离子结构的合理性, 构建了"结构-性质-性能"联动的反向设计流程。结果表明, 所生成离子在界面性质上优于原始数据集, 在保持合理结构相似性的同时实现了性能提升, 体现了该方法在性能导向分子设计中的有效性与针对性。未来工作将致力于替代现有的结构约束工具(如 RDKit), 引入更高效的分子合理性评估机制, 以提升优化效率与设计能力。同时将进一步拓展目标性质维度, 发展多目标优化策略, 并探索图神经网络等更具表达力的建模框架, 为复杂界面体系下的离子液体设计提供更通用、高效的技术支撑。

参考文献:

[1] PÖHLMANN F, SCHILDER L, KORTH W, *et al*. Liquid phase isobutane/2 - butene alkylation promoted by hydrogen chloride using Lewis acidic ionic liquids[J]. ChemPlusChem, 2013, 78(6): 570-577.

[2] KORE R, SCURTO A M, SHIFLETT M B. Review of isobutane alkylation technology using ionic liquid-based catalysts: Where do we stand?[J]. Industrial & Engineering Chemistry Research, 2020, 59(36): 15811-15838.

[3] LEI Z, CHEN B, KOO Y M, *et al*. Introduction: Ionic liquids[J]. Chemical Reviews, 2017, 117(10): 6633-6635.

[4] DOKOOHAKI M H, ZOLGHADR A R, GHATEE M H, *et al*. Aqueous solutions of binary ionic liquids: Insight into structure, dynamics, and interface properties by molecular dynamics simulations and DFT methods[J]. Physical Chemistry Chemical Physics, 2020, 22(47): 27882-27895.

[5] BARDESTANI R, PATIENCE G S, KALIAGUINE S. Experimental methods in chemical engineering: Specific surface area and pore size distribution measurements: BET, BJH, and DFT[J]. The Canadian Journal of Chemical Engineering, 2019, 97(11): 2781-2791.

[6] INGHAM J, DUNN I J, HEINZLE E, *et al*. Chemical Engineering Dynamics: An Introduction to Modelling and Computer Simulation[M]. Weinheim, Germany: John Wiley & Sons, 2008.

[7] GARDAS R L, COUTINHO J A P. Group contribution methods for the prediction of thermophysical and transport properties of ionic liquids[J]. AIChE Journal, 2009, 55(5): 1274-1290.

[8] ZHANG X, LIU Z, WANG W. Screening of ionic liquids to capture CO_2 by COSMO - RS and experiments[J]. AIChE Journal, 2008, 54(10): 2717-2728.

[9] ZHENG W, MA Z, SUN W, *et al*. Target high-efficiency ionic liquids to promote H_2SO_4-catalyzed C4 alkylation by machine learning[J]. AIChE Journal, 2022, 68(7): e17698.

[10] CAO L, ZHU P, ZHAO Y, *et al*. Using machine learning and quantum chemistry descriptors to predict the toxicity of ionic liquids[J]. Journal of Hazardous Materials, 2018, 352: 17-26.

[11] LOW K, KOBAYASHI R, IZGORODINA E I. The effect of descriptor choice in machine learning models for ionic liquid melting point prediction[J]. The Journal of Chemical Physics, 2020, 153(10): 1103-1113.

[12] KUROKI N, SUZUKI Y, KODAMA D, *et al*. Machine learning-boosted design of ionic liquids for CO_2 absorption and experimental verification[J]. The Journal of Physical Chemistry B, 2023, 127(9): 2022-2027.

[13] PAQUET E, VIKTOR H L. Molecular dynamics, Monte Carlo simulations, and Langevin dynamics: A computational review[J]. BioMed Research International, 2015, 2015(1): 183918.

[14] HARVEY M J, DE FABRITIIS G. High-throughput molecular dynamics: The powerful new tool for drug discovery[J]. Drug Discovery Today, 2012, 17(19/20): 1059-1062.

[15] REUTLINGER M, RODRIGUES T, SCHNEIDER P, *et al*. Multi-objective molecular *de novo* design by adaptive fragment prioritization[J]. Angewandte Chemie International Edition, 2014, 53(16): 4244-4248.

[16] DEY F, CAFLISCH A. Fragment-based *de novo* ligand design by multiobjective evolutionary optimization[J]. Journal of Chemical Information and Modeling, 2008, 48(3): 679-690.

[17] YUAN Y, PEI J, LAI L. LigBuilder 2: A practical *de novo* drug design approach[J]. Journal of Chemical Information and Modeling, 2011, 51(5): 1083-1091.

[18] WINTER R, MONTANARI F, STEFFEN A, *et al*. Efficient multi-objective molecular optimization in a continuous latent space[J]. Chemical Science, 2019, 10(34): 8016-8024.

[19] TANABE R, FUKUNAGA A S. Improving the search performance of SHADE using linear population size reduction[C]//2014 IEEE Congress on Evolutionary Computation (CEC). USA: IEEE, 2014: 1658-1665.

[20] ZHANG J, SANDERSON A C. JADE: Adaptive differential evolution with optional external archive[J]. IEEE Transactions on Evolutionary Computation, 2009, 13(5): 945-958.

[21] PHILIPPI F, PUGH D, RAUBER D, *et al*. Conformational design concepts for anions in ionic liquids[J]. Chemical Science, 2020, 11(25): 6405-6422.

[22] PHILIPPI F, WELTON T. Targeted modifications in ionic liquids–from understanding to design[J]. Physical Chemistry Chemical Physics, 2021, 23(12): 6993-7021.

[23] YAN C, LI G. The rise of machine learning in polymer discovery[J]. Advanced Intelligent Systems, 2023, 5(4): 2200243.

[24] GAO W, COLEY C W. The synthesizability of molecules proposed by generative models[J]. Journal of Chemical Information and Modeling, 2020, 60(12): 5714-5723.

[25] WANG J, HSIEH C Y, WANG M, *et al*. Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning[J]. Nature Machine Intelligence, 2021, 3(10): 914-922.

[26] YANG F, VAN HERWERDEN D, PREUD'HOMME H, *et al*. Collision cross section prediction with molecular fingerprint using machine learning[J]. Molecules, 2022, 27(19): 6424.

[27] AL-SHEHARI T, ALSOWAIL R A. An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques[J]. Entropy, 2021, 23(10): 1258.

[28] WINTER R, MONTANARI F, NOÉ F, *et al*. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations[J]. Chemical Science, 2019, 10(6): 1692-1701.

[29] RIGATTI S J. Random forest[J]. Journal of Insurance Medicine, 2017, 47(1): 31-39.

[30] TODESCHINI R, GRAMATICA P. New 3D molecular descriptors: The WHIM theory and QSAR applications[J]. Perspectives in Drug Discovery and Design, 1998, 9: 355-380.

Machine Learning-Based Acid-Hydrocarbon Interface Prediction and Ionic Liquids Design

TIAN Yifan[1], GAO Weiqun[2,3], LU Jingyi[1], ZHENG Weizhong[2,3], SUN Weizhen[2,3]
(1. School of Information Science and Engineering; 2. State Key Laboratory of Chemical Engineering and Low-Carbon Technology; 3. School of Chemical Engineering, East China University of Science and Technology, Shanghai 200237, China)

Abstract: Ionic liquids, with their green chemistry attributes and tunable properties, have garnered significant interest as potential additives in sulfuric acid-catalyzed C4 alkylation. Given the vast combinatorial possibilities of anion-cation pairs, traditional experimental screening methods are inefficient in exploring extensive chemical spaces, thus falling short. To address this challenge, we employed machine learning techniques—most notably the random forest (RF) algorithm—to establish correlations between the structural features of ionic liquids and their enhanced acid-hydrocarbon interfacial properties. This approach enabled the development of predictive models based on diverse descriptors, advancing our understanding and facilitating the selection of ionic liquids for specific applications.Furthermore, to streamline the rational design of novel ionic liquid combinations, we utilized continuous and data-driven molecular descriptors (CDDD) derived from the SMILES codes of the compounds. These descriptors were fed into the Success-History based Adaptive Differential Evolution (SHADE) algorithm, which efficiently navigates and decodes the potential space to identify promising candidates. Substructure constraints were also integrated to ensure the rationality and feasibility of the generated structures.Focusing on key parameters such as interfacial thickness (δ_w) and tension (γ) in C4 alkylation, the constructed predictive models achieved a determination coefficient (R^2) of 0.952 for interfacial thickness and 0.901 for interfacial tension on the test set, indicating high predictive accuracy. Additionally, through optimization via the SHADE algorithm, 328 ionic liquid combinations meeting the requirements for interfacial thickness and tension were successfully generated, significantly expanding the feasible range of qualifying ionic liquid combinations. This work not only enhances the capability to predict acid-hydrocarbon interfacial properties but also provides novel methodologies and insights for the rational design of ionic liquids.

Key words: ionic liquid; machine learning; acid-hydrocarbon interface property; rational design; evolutionary algorithm

（责任编辑：张欣）