基于机器学习的 SF_6 替代气体绝缘强度预测*

刘伟[1)] 凌梦旋[2)] 廖红[1)] 林涛[1)] 吴文婷[1)] 程龙玖[2)†]

1) (国网安徽省电力有限公司电力科学研究院, 合肥 230601)

2) (安徽大学化学化工学院, 合肥 230601)

(2025 年 9 月 8 日收到; 2025 年 12 月 2 日收到修改稿)

绝缘强度 (E_r) 是筛选和评估六氟化硫 (SF_6) 替代气体的关键指标. 本研究基于机器学习方法, 构建了 SF_6 替代气体的 E_r 预测模型. 首先, 收集了 88 个气体分子相对 SF_6 的 E_r 数据并计算了这些分子的全局参数和静电势参数, 并将其作为描述符. 采用 5 种经过五折交叉验证和超参数优化后的机器学习方法, 在 E_r 数据及描述符之间建立 E_r 预测模型. 研究表明, 自适应增强回归模型表现突出, 其决定系数达到了 0.90, 平均绝对误差和均方根误差分别为 0.17 和 0.18. 结合 Shapley 加性解释量化了描述符特征对 E_r 的贡献, 发现极化率是影响 E_r 的主要因素. 最后, 利用 SF_6 及已知的 6 种环保替代气体对自适应增强回归模型的 E_r 进行了验证, 其绝对误差均在 0.02—0.33 之间, 进一步证实了该模型的可靠性. 本研究有望为筛选 SF_6 替代气体提供一种可行的路径.

关键词: 绝缘强度预测, 机器学习, SF_6 替代气体, 密度泛函理论

DOI: 10.7498/aps.75.20251229 **CSTR:** 32037.14.aps.75.20251229

1 引 言

六氟化硫 (sulfur hexafluoride, SF_6) 因其较高的绝缘灭弧性能, 成为气体绝缘金属封闭开关设备、绝缘断路器、输电线路的首选绝缘介质[1]. 然而, 早在 1997 年, 《京都议定书》就将 SF_6 列为 6 种主要温室气体之一, 其全球变暖潜能值 (global warming potential, GWP) 约为 CO_2 的 25200 倍[2]. 数十年来, 国内外对于寻找 SF_6 的环境友好型替代气体进行了大量的研究与探索. 研究表明, CO_2, N_2 等作为自然界的常规气体, 来源广泛, 环境友好, 但其绝缘灭弧性能远低于 SF_6, 限制了它们在高压电力设备中的应用[3]. 采用 CO_2, N_2 等作为缓冲气体与 SF_6 混合可降低 SF_6 的用量, 但仍不能从根本上达到 SF_6 零排放目标[4]. 近几十年来, 新型环保绝缘替代气体 (如 C_4F_7N, $C_5F_{10}O$, CF_3SO_2F 等) 因其独特的绝缘性能而备受关注, 但部分替代气体, 因成本较高, 合成工艺复杂, 且难以兼顾绝缘、液化、环保和安全等性能, 被限制使用[5–7]. 因此, 寻求替代 SF_6 的环境友好型气体仍然是电网绿色发展亟待解决的重要问题.

绝缘强度 (dielectric strength, E_r) 作为衡量气体绝缘性能的核心指标, 是筛选和评估 SF_6 替代气体的关键因素, 决定了气体在高电压条件下的绝缘能力, 影响着电力设备的安全性、可靠性和经济性. 目前实验室获取 E_r 的方法有气体击穿特性实验[8]、稳态汤逊实验[9] 和局部放电实验[10] 等. 然而, 这些实验获取绝缘性能的方法不仅费时费力, 而且价格昂贵, 不适合批量的进行新型 SF_6 替代气体的筛选. 自 20 世纪 90 年代以来, 随着密度泛函理论等第一性原理方法的成熟, 该理论已经被广泛应用于 SF_6 替代气体 E_r 的预测[11–22]. 此外, 机器学习作为以数据驱动为核心的预测方法, 能够自动学习数据中的非线性关系, 在处理复杂体系时表现出色[23–29]. 不仅如此, 机器学习也可以通过优化算法

* 安徽省自然科学基金 (批准号: 2208085UD16) 资助的课题.

† 通信作者. E-mail: clj@ustc.edu

和参数调整实现高精度预测, 并结合交叉验证等技术有效避免过拟合问题, 从而在新数据上展现出良好的泛化能力. 目前机器学习也被应用于 SF_6 替代气体 E_r 的预测, 并取得一定进展. 如表 1 所示, Rabie 等 [23] 和刘关平等 [18] 分别采用多元线性回归 (multiple linear regression, MLR) 和电子定域化函数表面分析的方法预测 E_r, 结果表明, 两种模型对非极性分子预测效果良好, 决定系数 (R^2) 均达到了 0.90 以上, 但在极性分子中预测效果并不理想. 在 Sun 等 [24] 报道的方法中, 随机森林 (random forest, RF) 方法预测 E_r 效果最佳但 R^2 也仅达到了 0.83. 此模型人为划分特定化学类别作为测试集, 虽然能够更直观地呈现测试结果, 但其在更广泛的数据上无法泛化, 缺乏通用性. 杨志强等 [11] 综述了大量定量构性关系 (QSPR) 模型, 其测试效果的 R^2 值范围为 0.71—0.98, 虽然整体表现良好, 但这些模型无法量化描述各个特征对模型的贡献程度, 仅能说明特征对模型存在影响. 衡盼盼等 [17] 和 Zhao 等 [20] 基于化学键组合对 E_r 进行预测其相关系数 (R) 达到了 0.97—0.988. 此方法虽然效果良好, 但却仅限于所应用的特定化学键的分子, 对于预测其他化学键的替代分子效果有待考究.

表 1 常见的 SF_6 替代气体的绝缘强度 (E_r) 预测方法及结果比较

Table 1. Prediction methods and results of E_r for common SF_6 replacement gases.

方法	数据集	数据类型	预测效果	文献
MLR	48	极性	$R^2 = 0.71$	[23]
MLR	19	非极性	$R^2 = 0.92$	[23]
MLR	67	极性+非极性	$R^2 = 0.50$	[23]
电子定域化函数表面分析	57	极性	$R^2 = 0.73$	[18]
电子定域化函数表面分析	15	非极性	$R^2 = 0.95$	[18]
RF	67	极性+非极性	$R^2 = 0.83$	[24]
QSPR	24—104	极性+非极性	$R^2 = 0.71$—0.90	[11]
化学键组合	63	极性+非极性	$R = 0.97$	[17]
化学键组合	47	极性+非极性	$R = 0.988$	[20]

本研究收集了 88 个实验测得的气体分子相对 SF_6 的 E_r 数据 [12,24,25,29–31], 并计算了这些分子的全局参数和静电势参数作为分子描述符. 采用 5 种经过五折交叉验证和超参数优化后的机器学习方法在 E_r 数据和分子描述符之间建立预测模型, 用 R^2、平均绝对误差 (mean absolute error, MAE) 和均方根误差 (root-mean-square error, RMSE) 作为评价指标确定最优预测模型. 在模型训练与测试阶段, 本研究采用两种数据划分策略: 一是随机分组划分, 二是将特定绝缘气体分子固定为测试集, 其余数据作为训练集. 两种方法所得误差均较小, 充分证明了模型出色的泛化能力. 此外, 本研究借助 SHAP 分析深入解析了每个特征对模型的贡献程度及特征重要性, 这不仅为后续开发 SF_6 替代气体的新分子提供了明确的方向, 还增强了模型的可解释性.

2 计算方法

用于描述气体分子的微观结构性质的描述符可分为全局参数和分子表面静电势参数 [32,33]. 众所周知, 气体分子的 E_r 主要受两个方面的影响: 一方面是气体分子对空间电荷的减速能力; 另一方面是气体分子吸附空间电荷、抑制自身产生空间电荷的能力 [21]. 因此, 我们计算了全局和分子表面静电势两种参数, 通过全局参数描述分子的整体性质, 表面静电势参数反映了分子电荷分布的局部特征. 为得到这些微观描述符的特征, 我们基于密度泛函理论, 采用 Gaussian 16 软件, 在 M062X/6-311+G (d, p) 理论水平下对所有分子进行了结构优化, 并计算频率确保所有分子为最稳定的构象. 结构优化采用的 Gaussian 16 软件的默认收敛标准: 最大受力 (maximum force) < 0.000450 Hartree/Bohr, 均方根受力 (root-mean-square force) < 0.000300 Hartree/Bohr, 最大位移 (maximum displacement) < 0.001800 Bohr, 方均根位移 (root-mean-square force displacement) < 0.001200 Bohr. 参考文献 [32,33] 通过密度泛函理论计算得来的全局参数和分子表面静电势参数被列举在补充材料表 S2 和表 S3 (online) 中, 其中表 2 仅列举了参数名称和单位, 便于后续相关描述. 另外, 分子表面静电势参数 [34] 是采用 Multiwfn[35] 软件进行了波函数分析, 以电子密度为 0.001 a.u.的等值面构建分子表面, VMD[36] 软件进行可视化得到的.

机器学习方法能够通过设计算法让计算机从数据中自动学习规律和模式, 从而实现对新数据的预测和决策. 本研究直接采用 scikit-learn 开源库中的两个主流分支: 基于 Bagging 的随机森林 (random forest, RF) 和极端随机树 (extra trees, ET) 方法, 以及基于 Boosting 的梯度提升回归 (gradient

表 2 全局参数和分子表面静电势参数的相关名称及单位

Table 2. The relevant names and units of global parameters and molecular surface electrostatic potential parameters.

全局参数	分子静电势参数	
分子量(M, g/mol)	等值面包裹的体积(V_A, A^3)	密度(ρ, g/cm^3)
电离能(VIE, eV)	总表面积(A_s, A^2)	总均值(V_s, kcal/mol)
亲合能(VEA, eV)	正电势表面积 (A_s^+, A^2)	负电势表面积 (A_s^-, A^2)
电负性(χ, eV)	正方差(σ_+^2, kcal/mol)	负电势均值(V_s^-, kcal/mol)
偶极矩(μ, D)	负方差(σ_-^2, kcal/mol)	正电势均值(V_s^+, kcal/mol)
极化率(α, A^3)	总方差(σ_{tot}^2, kcal/mol)	内部电荷分离指数Π
分子最高占据轨道能量(HOMO, eV)	非极性表面积(S_{NSA}, A^2)	极性表面积(S_{PSA}, A^2)
分子最低未占据轨道能量(LUMO, eV)	分子极性指数 (I_{MPI}, kcal/mol)	椭圆率O_{val}
HOMO和LUMO的能量差(HOMO-LUMO gap, eV)	电荷平衡度 v	总偏度O_s
	正电势偏度 O_s^+	负电势偏度 O_s^-
	静电相互作用趋势指数($v\sigma^2$, (kcal/mol)^2)	
	最小值($V_{s,min}$, kcal/mol)	最大值($V_{s,max}$, kcal/mol)

boosting regressor, GBR)、自适应增强回归 (ada boost regressor) 和直方图梯度提升回归 (hist gradient boosting regressor, HGBR) 共 5种机器学习方法 (细节见补充材料 (online)) 进行训练和测试. 采用随机分组的方式按照 9:1 划分训练集和测试集. 在训练过程中, 对不同机器学习方法均进行五折交叉验证和超参数优化确保模型在未见数据上具有良好的泛化能力. 另外, 采用以 R^2, MSE 和 RMSE 评估模型的准确性. 一般而言, 较高的 R^2 和较低的 MSE 和 RMSE 表明模型对 SF_6 替代气体 E_r 性能预测较好.

由于机器学习方法内部流程较为复杂, 通常被称为"黑盒子"模型. SHAP 分析方法的提出使机器学习模型的输出有了可解释性 [26]. 其是一种基于博弈论的模型解释方法, 主要通过计算每个特征对模型预测结果的贡献 (即 SHAP 值), 帮助理解模型的决策过程. 因此, 采用 SHAP 分析的方法揭示每个微观描述符对模型的贡献程度. 正的 SHAP 值表示特征对预测结果有正向贡献. 而负的 SHAP 值则表示特征对预测结果有负向贡献.

3 结果与讨论

3.1 数据集的构建

我们共收集在实验中已经测得 E_r 的气体分子, 共 88 个 (极性分子 68 个, 非极性分子 20 个), 包含卤代烃 (氟代烃、氯代烃和溴代烃等)、腈基 (—CN)、硝基 (—NO_2)、醚类 (—O—)、酯类 (—COO—)、醛类 (—CHO)、酮类 (—CO—)、硫醚类 (—S—)、磺酰类 (—SO_2—)、硫代酰胺类 (—CS—NH_2)、炔烃类 (—C≡C—)、烯烃类 (—C=C—) 等多种官能团. 这些气体分子除了基本 C, H 骨架外还包含大部分 O, N, F, S, Cl, Br 等电负性较强的元素, 尤其是含 F 化合物占比高达 72%. 这些绝缘气体分子的 E_r 均是相对于 SF_6 的 (SF_6 = 1).

影响分子绝缘性能的全局参数, 均可通过分析优化后基态分子或其得/失一个电子后形成的阳/阴离子的结构获得. 另外, 通过对波函数分析获取了分子的静电势参数并绘制出分子表面静电势图. 以 SF_6, C_4F_7N, CF_3SO_2F 以及 $C_5F_{10}O$ 为例, 从图 1 中可以直观地看出分子表面不同区域的电子分布特征, 其中蓝色区域为负静电势区域, 对正电荷有吸引力, 而红色区域为正静电势区域, 对负电荷有吸引力. SF_6 分子高度对称的 O_h 结构, 导致其电荷分布较为均匀. 此外, F 原子的电负性高于 S 原子, 所以中心 F 原子具有较高的负静电势, 而 S 原子中心位置则相反. 这种差异导致 SF_6 分子表面的静电势分布呈现明显的极性特征. SF_6 替代气体 (C_4F_7N, $C_5F_{10}O$ 和 CF_3SO_2F) 在保留强电负性的 F 原子同时, 还引入了比 S 电负性更高的 N, O 元素, 使其表面静电势分布更加复杂, 极性区域更为显著. 从 SF_6 及其常见的替代气体的静电势分布图来看, 气体分子表面静电势信息能够极大的影响分子的绝缘性能. 因此, 如支撑信息的表 S2 和表 S3(online) 所示, 我们共计算了 88 种 SF_6 替代气体分子的全局参数和分子静电势参数共计 32 个特征, 作为描述符用于机器学习的输入端.

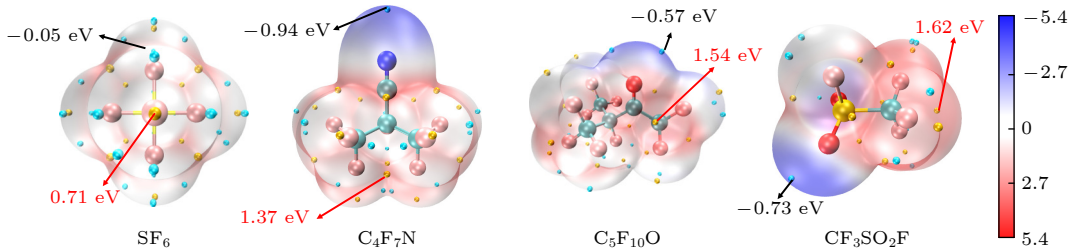


图 1 SF_6, C_4F_7N, $C_5F_{10}O$ 及 CF_3SO_2F 的表面静电势及极值点分布 (黄色为极大值点, 青色为极小值点, eV), 箭头指向为静电势的最大值点 (红色) 和最小值点 (黑色)

Fig. 1. Surface electrostatic potential and extreme point distribution of SF_6, C_4F_7N, $C_5F_{10}O$, and CF_3SO_2S (yellow dots represent maximum electrostatic potential points, and cyan dots represent minimum electrostatic potential points, eV), the arrow points to the maximum (red) and minimum (black) points for the electrostatic potential.

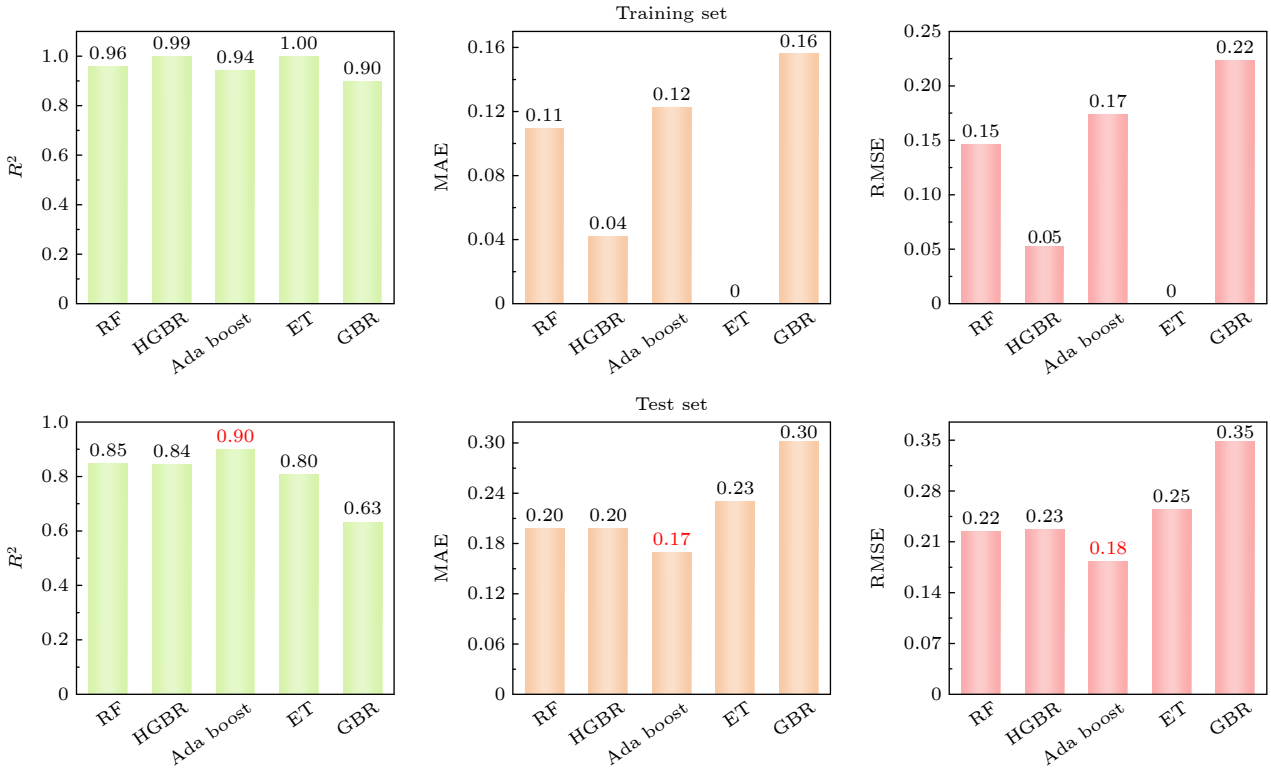


图 2 基于训练集/测试集中 E_r 的预测值和实验值对比, 利用 R^2 (绿色)、MAE(橙色) 和 RMSE(粉色) 评估 5 种机器学习模型的性能

Fig. 2. Performance evaluation of five machine learning models based on the comparison between the predicted/experimental values of E_r in the training/test sets and the experimental values using R^2 (green), MAE (orange), and RMSE (pink).

3.2 机器学习模型的构建

将 88 个 SF_6 替代气体分子的 E_r 数据及微观描述符作为输入, 构建了 5 种不同的机器学习模型, 即 RF, GBR, Ada boost, ET 和 HGBR. 为避免训练过程出现过拟合或欠拟合, 调整 5 个机器学习模型的超参数, 并进行了五折交叉验证来提高模型的泛化能力, 表 S1(online) 展现了 5 种机器学习模型最佳的超参数结果. 以 R^2, MAE 和 RMSE 作为评价指标的 5 种机器学习模型的训练和测试结果如图 2 所示. 从训练集中可以看出, 5 种模型的学习效果都极佳, R^2 均达到了 0.90 以上. ET 方法的 R^2 更是达到了 1.00, MAE 和 RMSE 为 0, 几近完美. 但训练集的训练结果并不能完全体现机器学习模型的学习能力, 需要从测试集的角度出发去评估 5 种模型的性能. 在测试集中, 模型的性能从高到低的排序依次为 Ada boost > RF > HGBR > ET > GBR. 其中, Ada boost 方法在测试集中效果最佳, R^2 达到了 0.90, MAE 和 RMSE 均低于 0.2. GBR 方法的拟合效果最差, R^2 仅有 0.63, MAE 和 RMSE 也相对较高. 而剩余 3 种模型 (RF, ET

和 HGR) 的 R^2 均在 0.8 左右, MAE 和 RMSE 的值控制在 0.20—0.25 之间. 针对训练集中效果极佳的 ET 方法, 其在测试集中的表现较差 (R^2 = 0.80, MAE = 0.23, RMSE = 0.25), 该现象被归结于训练过程中出现了过拟合的现象.

综上, 最佳的机器学习模型是 Ada boost, 其在训练集和测试集上的实验 E_r 和预测 E_r 之间的比较如图 3 所示, 详细数据 (实验 E_r 与预测 E_r) 分别列于表 S4 和表 S5 (online) 中. 从图 3 可以看出, 训练集 (由蓝色圆圈表示) 和测试集 (由橙色星号表示) 几乎所有的点都聚集在理想直线附近, 表明预测值与实验值相近. 我们还采用 1000 次蒙特卡罗交叉验证的方法来验证模型的可靠性. 其在迭代过程中训练集与测试集的性能对比曲线, 如图 S1(online) 所示, 训练集 R^2, MAE 和 RMSE 均值分别为 0.94, 0.13 和 0.17, 测试集 R^2, MAE 和 RMSE 均值分别为 0.86, 0.20 和 0.25. 另外, 在图 S1(online) 中还计算了过拟合程度 ($R^2_{\text{Train}} - R^2_{\text{Test}}$) 的分布, 其均值为 0.075, 以上结果表明模型未出现严重过拟合.

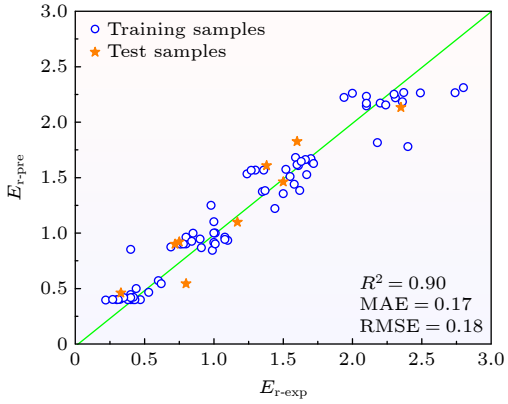


图 3 基于 Ada boost 模型对训练集和测试集中 E_r 的预测值 ($E_{r\text{-pre}}$) 和实验值 ($E_{r\text{-exp}}$) 的比较 (蓝色圆圈代表训练集, 橙色星号代表测试集, 右下角标注了预测值和实验值之间的误差: R^2, MAE 和 RMSE)

Fig. 3. Comparison of predicted ($E_{r\text{-pre}}$) values and experimental ($E_{r\text{-exp}}$) values of E_r in training and test sets based on Ada boost model (blue circles represent training set, orange asterisks represent test set, the errors between the predicted values and the experimental values: R^2, MAE, and RMSE are marked in the right corner).

为进一步验证 Ada boost 模型在实际替代气体筛选中的可信度, 我们从原始的 88 个样本中选取 SF_6 及 7 个常见的替代气体 (CF_3OCF_3, CF_3SO_2F, $CF_3(CF)_2CF_3$, $C_5F_{10}O$, C_4F_7N 和 $CFCl_2CFCl_2$) 作为新的测试集; 剩余 81 个样本为新的训练集并利用五折交叉验证与超参数调优训练 AdaBoost 模型. 如图 4 所示, 无论是训练集 (以蓝色圆圈表示) 还是测试集 (以红色星号表示) 都分布在对角线附近, 表明预测值与实际值较为吻合, 预测效果良好 (R^2 = 0.87, MAE = 0.15 和 RMSE = 0.19). 另外, 我们还对 7 个测试集的实验值和预测值进行详细对比, 并将其标记在图中. 结果显示, 它们的绝对误差均控制在 0.02—0.33 之间, 进一步证实 AdaBoost 模型的可靠性.

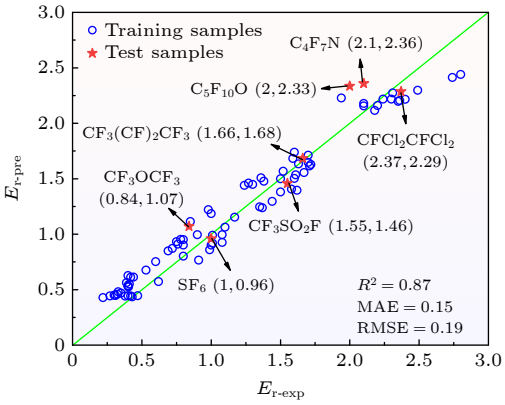


图 4 基于 Ada boost 模型对常见的 SF_6 替代气体的 E_r 的预测值 ($E_{r\text{-pre}}$) 和实验值 ($E_{r\text{-exp}}$) 的比较 (蓝色圆圈代表训练集, 红色星号代表测试集 (常见的 SF_6 替代气体), 括号的左侧为实验值, 右侧为预测值, 右下角标注了预测值和实验值之间的误差 R^2, MAE 和 RMSE)

Fig. 4. Comparison of predicted ($E_{r\text{-pre}}$) and experimental ($E_{r\text{-exp}}$) values of E_r by Ada boost model for common SF_6 replacement gases (blue circle represent training set, red asterisk represent test set (common SF_6 replacement gases), left side of the bracket is the experimental value, and the right side is the predicted value, the errors between the predicted values and the experimental values: R^2, MAE, and RMSE are marked in the right corner).

3.3 描述符对 E_r 的贡献解析

在优化后的 Ada boost 模型的基础上, 进一步利用 SHAP 分析进行解释, 深入探讨相关特征描述符与 E_r 之间的潜在关系. 图 5(a) 呈现了 Ada boost 模型预测 E_r 中有重要贡献的 9 个关键特征 ($\alpha > M > V_A > O_{val} > \text{HOMO} > A_s^- > A_s^+ > \text{VEA} > A_s$). 75.8% 的模型输出依赖于这 9 个特征的贡献, 而其余 23 个特征的占比仅为 24.2%. 图 5(b) 展示了模型的 SHAP 分析值的分布图. 图 5 中, 特征描述符依据其对模型预测结果的影响程度, 通过颜色编码进行区分. 红色代表高影响程

度, 意味着该特征对模型的预测结果有积极的推动作用; 而蓝色则代表低影响程度或消极影响, 这表明该特征对模型的预测结果影响较小或有消极影响. 对模型有重要贡献的 9 个关键特征中 α, M, $V_{\rm A}$, $A_{\rm s}^-$, $A_{\rm s}^+$ 和 $A_{\rm s}$ 特征 SHAP 值的正值区域呈现出红色, 表明这些特征对 $E_{\rm r}$ 的预测具有正向影响. 相对地, $O_{\rm val}$, HOMO 和 VEA 这些特征的负 SHAP 值对 $E_{\rm r}$ 的预测影响更为显著, 表明它们对预测结果具有逆向影响.

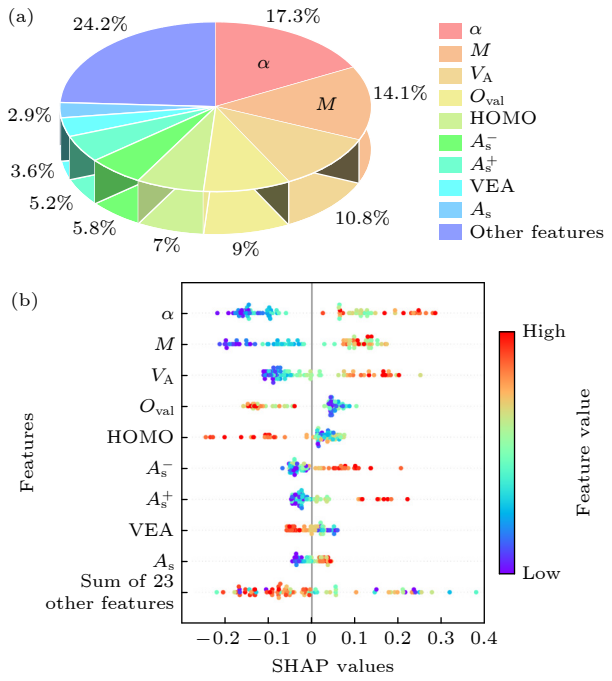


图 5 Ada boost 模型的 SHAP 分析结果 (a) 所有微观描述符 (全局参数和分子静电势参数) 对 $E_{\rm r}$ 的贡献度; (b) 所有微观描述符 (全局参数和分子静电势参数) 的 SHAP 值

Fig. 5. SHAP analysis of Ada boost model: (a) The contribution of all microscopic descriptors (global parameters and molecular surface electrostatic potential parameters) to $E_{\rm r}$; (b) the SHAP values of all microscopic descriptors (global parameters and molecular surface electrostatic potential parameters).

在有重要贡献的 9 个关键特征, 全局参数有 4 个, 其余为分子表面静电势参数. 在全局参数中, 最重要的特征为分子的极化率 α. 众所周知, 气体介质在受到电场作用时会导致其吸附电子的能力发生变化. 分子的极化率是识别和控制电荷传输的关键性参数, 能够影响分子中电荷的传输能力[37]. 因此 α 能够从根本上影响分子的绝缘性能. 具有高 α 的分子在电场作用下更容易变形, 电子云更容易被极化, 对 $E_{\rm r}$ 呈正相关作用. 早在 1982 年, Vijh[38] 的研究也发现, 气体的分子量 M 与 $E_{\rm r}$ 之间存在近似正相关的趋势, 我们的研究更加证实了这一观点. HOMO 能量高低决定分子给出电子的难易程度, 较低的 HOMO 能量意味着电子更紧密地束缚在分子内部, 分子在电场中更难被电离, 会对分子的绝缘能力有正向的积极作用, 与 Zhao 等[21] 的研究结果相符. 电子亲和能 VEA 是指分子获得一个额外电子时释放的能量, 而较低的 VEA 意味着分子在电场中更难捕获电子, 同样对 $E_{\rm r}$ 有积极作用.

对于分子表面静电势参数, 等值面包裹的体积 $V_{\rm A}$ 与电子散射过程相关, 能够反映电子与分子的碰撞[18], 较大的 $V_{\rm A}$ 能够降低自由电子的动能, 从而增强分子对电场的屏蔽能力, 对 $E_{\rm r}$ 有着积极的影响, 与 Zhao 等[20] 的研究相符合. 对于分子结构拓扑参数椭圆度 $O_{\rm val}$, 反映了分子表面电荷分布的不对称性. 偏离理想球面在一定程度上有利于提高 $E_{\rm r}$, 这一观点与 Zhang 等[25] 的研究一致. 而负静电势表面积 $A_{\rm s}^-$、正静电势表面积 $A_{\rm s}^+$ 和分子的总表面积 $A_{\rm s}$, 三者的趋势一致, 均对模型有正向的积极作用, $A_{\rm s}^+$ 和 $A_{\rm s}^-$ 同时增大证明表面电荷分布的不对称性越强, 在一定程度上有利于提高 $E_{\rm r}$. 本研究通过引入 SHAP 分析具体的量化了这些特征对模型的贡献程度, 还深入剖析了它们作用于 $E_{\rm r}$ 的具体机制. 为后续相关研究提供了更为清晰的理论依据.

4 结 论

本研究收集了 88 种已知 SF_6 替代气体的 $E_{\rm r}$ 实验值数据构建 $E_{\rm r}$ 数据集. 并基于密度泛函理论计算了这 88 个分子的全局参数和分子静电势参数作为分子描述符. 采用五折交叉验证和超参数调优后的 5 种机器学习方法 (RF, GBR, Ada boost, ET 和 HGBR) 在描述符和 $E_{\rm r}$ 数据之间建立 SF_6 替代气体 $E_{\rm r}$ 预测模型, 发现优化后的 Ada boost 模型表现突出: $R^2 = 0.9$, MAE = 0.17, RMSE = 0.18. 利用 SHAP 分析仔细解释了对模型贡献程度较高的 9 个特征 (α, M, $V_{\rm A}$, $O_{\rm val}$, HOMO, $A_{\rm s}^-$, $A_{\rm s}^+$, VEA, $A_{\rm s}$), 发现极化率 (α) 是影响 $E_{\rm r}$ 的主要因素. 为进一步探究此方法在预测 SF_6 替代气体 $E_{\rm r}$ 方面的性能. 对 SF_6 及其常见替代气体 (CF_3OCF_3, CF_3SO_2F, $CF_3(CF)_2CF_3$, $C_5F_{10}O$, C_4F_7N 和 $CFCl_2CFCl_2$) 的 $E_{\rm r}$ 进行预测并对比了其实验值和预测值. 结果显示它们的绝对误差均控制在 0.02—0.33

之间, 表现出良好的预测性能. 本模型无需对极性和非极性分子进行分组预测, 且预测效果良好; 在模型训练与测试过程中采用了两种划分训练集和测试集的方法证明了模型具有良好的泛化能力; 最后基于 SHAP 分析具体量化了这些特征对模型的贡献程度, 为后续开发 SF_6 替代气体的新分子提供了明确方向.

参考文献

[1] Pan B F, Wang G M, Shi H M, Shen J H, Ji H K, Kil G S 2020 *Appl. Sci.* **10** 2526

[2] Cui Z L, Yi L, Song X, Tian S S, Tang J, Hao Y P, Zhao X X 2024 *Sci. Total Environ.* **906** 167347

[3] Gao K L, Yang Y, Zhou W J, Wang B S, Ding W D, Lin X, Yan X L, Zhang B Y, Wang D B, Xiao S 2024 *CSEE J. Power Energy* **44** 7395 (in Chinese) [高克利, 杨圆, 周文俊, 王宝山, 丁卫东, 张晓星, 林莘, 颜湘莲, 张博雅, 王邸博, 肖淞 2024 中国电机工程学报 **44** 7395]

[4] Ma T, Wang Q G, Wang W M 2025 *Low Temp. Special. Gases* **43** 1 (in Chinese) [马腾, 王泉高, 王伟民 2025 低温与特气 **43** 1]

[5] Liu W, Dong W C, Song Y M, Zhao Y H, Wu P P, Huang X, Wang K, Cheng L J 2023 *Int. J. Quantum Chem.* **123** e27114

[6] Zhang X X, Li Y, Xiao S, Tang J 2017 *Environ. Sci. Technol.* **51** 10127

[7] Li Y, Pang Z Y, Zheng H B 2024 *IEEE Trans. Dielectr. Electr. Insul.* **31** 297

[8] Chen K, Lin X, Xu J Y, Xiong W, Song H F 2024 *High Volt.* **60** 99 (in Chinese) [陈凯, 林莘, 徐建源, 熊玮, 宋会发 2024 高压电器 **60** 99]

[9] Long Y X, Guo L P, Shen Z Y, Chen C, Zhou W J, Liu W 2019 *High Volt.* **45** 1064 (in Chinese) [龙云翔, 郭立平, 沈震宇, 陈成, 周文俊, 刘伟 2019 高电压技术 **45** 1064]

[10] Liu Y J, Ren M, Wang K, Chen X, Wang T T, Dong M 2025 *High Volt.* **51** 1206 (in Chinese) [柳玉洁, 任明, 王凯, 陈旭, 王腾腾, 董明 2025 高电压技术 **51** 1206]

[11] Yang Z Q, Zeng J J, Ma Y D, Wei T, Zhao B, Liu Y Z, Zhang W, Lv J, Li X W, Zhang B Y, Tang N, Li L, Sun D W 2023 *Chem. Ind. Eng. Prog.* **42** 4093 (in Chinese) [杨志强, 曾纪珺, 马义丁, 尉涛, 赵波, 刘英哲, 张伟, 吕剑, 李兴文, 张博雅, 唐念, 李丽, 孙东伟 2023 化工进展 **42** 4093]

[12] Yu X J, Hou H, Wang B S 2017 *J. Comput. Chem.* **38** 721

[13] Brand K P 1982 *IEEE Trans. Electr. Insul.* **17** 451

[14] Zhang B Y, Chen L, Li X W, Guo Z, Pu Y J, Tang N 2020 *IEEE Trans. Dielectr. Electr. Insul.* **27** 1187

[15] Meurice N, Sandre E, Aslanides A, Vercauteren D P 2004 *IEEE Trans. Dielectr. Electr. Insul.* **11** 946

[16] Zhang C H, Shi H X, Cheng L, Zhao K, Xie X Y, Ma H B 2016 *IEEE Trans. Dielectr. Electr. Insul.* **23** 2572

[17] Heng P P, Zhang M, Hou H, Wang B S 2024 *Chem. J. Chin. U.* **45** 61 (in Chinese) [衡盼盼, 张咪, 侯华, 王宝山 2024 高等学校化学学报 **45** 61]

[18] Liu G P, Yang S, Zhang N N, Wang H, Xiao J X 2022 *High Volt.* **48** 2208 (in Chinese) [刘关平, 杨帅, 张闹闹, 王航, 肖集雄 2022 高电压技术 **48** 2208]

[19] Zhou W J, Qiu R, Gao K L, Zheng Y, Hou H, Wang B S, Luo Y B, Yu J H 2023 *High Volt.* **49** 895 (in Chinese) [周文俊, 邱睿, 高克利, 郑宇, 侯华, 王宝山, 罗运柏, 喻剑辉 2023 高电压技术 **49** 895]

[20] Zhao Q L, Tong D J, Tu Y P, Zheng Z, Qiu X Y, Wang C 2024 *IEEE Trans. Dielectr. Electr. Insul.* **30** 49

[21] Li H M, Zhao S, Xiao D M 2024 *IEEE Trans. Dielectr. Electr. Insul.* **31** 2013

[22] Zhang M, Hou H, Wang B S 2023 *High Volt.* **9** 484

[23] Rabie M, Dahl D A, Donald S M A, Reiher M, Franck C M 2013 *IEEE Trans. Dielectr. Electr. Insul.* **20** 856

[24] Sun H, Liang L Q, Wang C L, Wu L, Yang F, Rong M Z 2020 *IEEE Access* **8** 124204

[25] Zhao G B, Kim H W, Yang C W, Chung Y G 2024 *J. Phys. Chem. A* **128** 2399

[26] Liu W, Zha J W, Ling M X, Li D, Shen K D, Cheng L J 2024 *Chem. Phys.* **588** 112447

[27] Štrumbelj E, Kononenko I 2014 *Knowl. Inf. Syst.* **41** 647

[28] Zhang Q, Zheng Y J, Sun W B, Ou Z P, Odunmbaku O, Li M, Chen S S, Zhou Y L, Li J, Qin B, Sun K 2022 *Adv. Sci.* **9** 2104742

[29] Luo H R, Gou Q Z, Zheng Y J, Wang K X, Yuan R D, Zhang S D, Fang W, Luogu Z, Hu Y Z, Mei H, Song B, Sun K, Wang J, Li M 2025 *ACS Nano* **19** 2427

[30] Zhang Q, Tan W, Ning Y Q, Nie G Z, Cai M Q, Wang J N, Zhu H P, Zhao Y Q 2024 *Acta Phys. Sin.* **73** 230201 (in Chinese) [张桥, 谭薇, 宁勇祺, 聂国政, 蔡孟秋, 王俊年, 朱慧平, 赵宇清 2024 物理学报 **73** 230201]

[31] Li X W, Zhao H, Murphy A 2018 *J. Phys. D. Appl. Phys.* **51** 153001

[32] Yang S, Wu S B, Chen H 2024 *Comput. Theor. Chem.* **1238** 114680

[33] Wu S B, Yang S, Chen H 2024 *J. Mol. Model* **30** 413

[34] Lu T, Chen F W 2012 *J. Mol. Graph. Model.* **38** 314

[35] Lu T, Chen F J 2012 *Comput. Chem.* **33** 580

[36] Humphrey W, Dalke A, Schulten K 1996 *J. Mol. Graph.* **14** 33

[37] Gillet A, Cher S, Tassé M, Blon T, Alves S, Izzet G, Chaudret B, Proust. A, Demont P, Volatron F, Tricard S 2021 *Nanoscale Horiz.* **6** 271

[38] Vijh A K 1982 *IEEE Trans. Dielectr. Electr. Insul. EI* **17** 84

Machine learning-based prediction of dielectric strength for SF_6 replacement gases*

LIU Wei [1)] LING Mengxuan [2)] LIAO Hong [1)] LIN Tao [1)]
WU Wenting [1)] CHENG Longjiu [2)†]

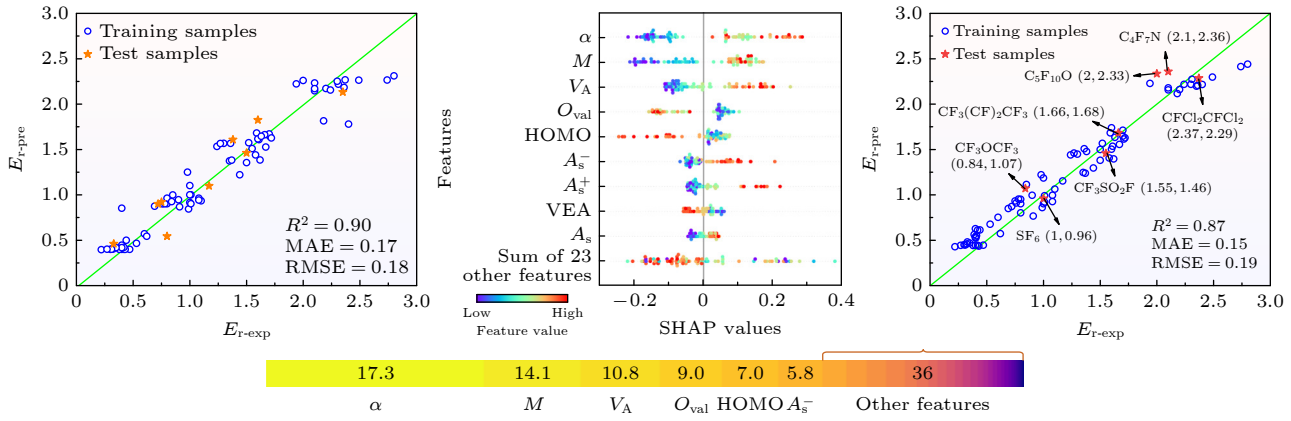
1) (*Electric Power Research Institute, Anhui Electric Power Co., Ltd., Hefei* 230601, *China*)

2) (*School of Chemistry and Chemical Engineering, Anhui University, Hefei* 230601, *China*)

(Received 8 September 2025; revised manuscript received 2 December 2025)

Abstract

Dielectric strength (E_r) is a critical factor in screening and evaluating SF_6 replacement gas. The traditional experimental methods of measuring E_r are not only extremely time-consuming but also very costly. In this work, an E_r prediction model for SF_6 replacement gases is constructed using machine learning methods. First, an exhaustive literature survey is performed to collect 88 high-quality experimental E_r values. Second, a total of 32 insightful microscopic descriptors are accurately calculated for each compound using density functional theory, including both global parameters and molecular electrostatic potential parameters. Furthermore, five state-of-the-art machine learning algorithms, which have been carefully modified through five-fold cross-validation and hyperparameter optimization, are used to train and test the 88 experimental E_r data and their relevant microscopic descriptors. Finally, the result shows that the Ada Boost regression model exhibits superior predictive performance, with a coefficient of determination of 0.90, a mean absolute error of 0.17, and a root mean square error of 0.18. Moreover, Shapley Additive exPlanations analysis is used to reveal the correlation between the microscopic descriptors and E_r. The results indicate that polarizability is the predominant factor significantly affecting E_r, which accounts for as high as 17.3%, followed by the molecular weight (14.1%). Specifically, molecules with high α are more prone to deformation under the action of an electric field, and their electron clouds are more likely to be polarized, which has a positive correlation with E_r. There is an approximately positive correlation between the molecular weight and the E_r of gases. To verify the reliability of the Ada Boost regression model for E_r prediction, the E_r of SF_6 and six known environmentally friendly replacement gases are tested within an absolute error of 0.02–0.33. This study provides a feasible method for accelerating the search for SF_6 replacement gases.



Keywords: dielectric strength prediction, machine learning, SF_6 replacement gas, density functional theory

DOI: 10.7498/aps.75.20251229 **CSTR:** 32037.14.aps.75.20251229

* Project supported by the Natural Science Foundation of Anhui Province, China (Grant No. 2208085UD16).

† Corresponding author. E-mail: clj@ustc.edu