doi:10.3969/j.issn.1673-9833.2024.05.005

基于机器学习的有机太阳能电池能级预测及分子设计

彭鑫裕[1]，雷　敏[1]，赵潇捷[2]，彭志嫣[2]

（1. 湖南工业大学 电气与信息工程学院，湖南 株洲 412007；
2. 湖南工业大学 轨道交通学院，湖南 株洲 412007）

摘　要：作为分布式可再生能源关键组成部分的有机太阳能电池，其效率的主要限制因素是分子的最高占据分子轨道（HOMO）和最低未占据分子轨道（LUMO）之间的能级差异。为了能降低有机太阳能电池的制造成本，提高有机太阳能电池的能量转换效率，提出利用机器学习分析有机太阳能电池的能级，指导分子设计。首先，利用机器学习的高效性和成本效益，筛选出 20 个关键特征，以深入分析其如何影响光伏器件的性能。随后，构建了 6 种不同的预测模型，对比发现其中基于梯度提升的 XGBT 模型在预测有机太阳能电池性能方面表现最佳，其决定系数为 0.8，并且其均方根误差仅为 0.2。最后，利用该模型有效地预测了有机太阳能电池的性能，并且通过对 HOMO 与 LUMO 的深入分析，成功识别出两种影响有机太阳能电池能级的关键分子结构。

关键词：机器学习；分布式新能源；有机太阳能电池；最高占据分子轨道；最低未占据轨道

中图分类号：TP206^{+}.1　　**文献标志码**：A　　**文章编号**：1673-9833(2024)05-0033-07

引文格式：彭鑫裕，雷　敏，赵潇捷，等 . 基于机器学习的有机太阳能电池能级预测及分子设计 [J]. 湖南工业大学学报，2024，38(5)：33-39.

Machine-Learning-Based Energy Level Prediction and Molecular Design of Organic Solar Cells

PENG Xinyu[1]，LEI Min[1]，ZHAO Xiaojie[2]，PENG Zhiyan[2]

（1. College of Electrical and Information Engineering，Hunan University of Technology，Zhuzhou Hunan 412007，China；
2. College of Railway Transportation，Hunan University of Technology，Zhuzhou Hunan 412007，China）

Abstract：The main limiting factor for the efficiency of organic solar cells, a key component of distributed renewable energy, is the energy level difference between the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) of molecules. In view of a reduction of the manufacturing cost of organic solar cells and an improvement of their energy conversion efficiency, machine learning is used to analyze the energy levels of organic solar cells and guide the molecular design. Firstly, based on the high efficiency and cost-effectiveness of machine learning, 20 key features are selected for a deeper analysis of how they affect the performance of photovoltaic

收稿日期：2023-12-16

基金项目：湖南省省市联合基金资助项目（2020JJ6071）

作者简介：彭鑫裕，男，湖南工业大学硕士生，主要研究方向为电力电子与电力传动，
E-mail：M22085800061@stu.hut.edu.cn

通信作者：雷　敏，女，湖南工业大学研究员，博士，硕士生导师，主要研究方向为复杂网络理论及其在电网稳定性分析中的应用，E-mail：leimin0606@hotmail.com

devices. Subsequently, 6 different prediction models are constructed and compared. It is found that the XGBT model based on gradient boosting is characterized with the best performance in predicting the property of organic solar cells, with a coefficient of determination of 0.8 and a root mean square error of 0.2. Finally, the performance of organic solar cells can be effectively predicted by using this model, and through an in-depth analysis of HOMO and LUMO, two key molecular structures that affect battery energy levels are successfully identified.

Keywords：machine learning；distributed new energy；organic solar cell；highest occupied molecular orbital (HOMO)；lowest unoccupied molecular orbital (LUMO)

1 研究背景

随着科学技术的不断发展，光伏技术 [1] 在现代能源系统中的重要性日益凸显，尤其是在分布式光伏系统应用方面。这些能源系统的独特之处，在于它们不仅能够在电力短缺时可作为独立电源，还能够通过与储能系统的结合，优化能源利用效率，实现能源供需动态平衡 [2]。在此背景下，有机光伏（organic photovoltaic，OPV）技术因为其独特的轻薄、柔性和低成本属性等而显得尤为重要。OPV 技术能适应多样化的应用场景，包括但不限于城市建筑物表面、移动装置和临时建筑等。此外，OPV 的简便生产工艺使其更易于大规模应用。随着此类技术的不断进步，OPV 转换效率显著提升，这不仅使得其在分布式光伏系统中的应用前景更为广阔，也预示着它将成为推动可持续绿色能源解决方案发展的关键技术之一。

在 OPV 技术中，分子的最高占据轨道（highest occupied molecular orbital，HOMO）和最低未占据轨道（lowest unoccupied molecular orbital，LUMO）能级特性是决定器件性能的关键因素。而 HOMO 和 LUMO 间的能级差异，即 HOMO-LUMO 能隙，对 OPV 电池的光电转换效率有着直接的影响。较大的 HOMO-LUMO 能隙能拓宽吸光光谱范围，使 OPV 电池能够有效地转换更广泛光谱范围的光子能量为电能，减少能量损失。这不仅提高了光电转换效率，还增强了光生载流子的移动能力，提升了 OPV 电池的稳定性和电流输出能力 [3]。在分子设计中，对这些能级进行精准调节至关重要。通过优化 HOMO 和 LUMO 能级，可以显著提升材料的光电转换效率，提升分布式光伏系统的发展速度。精准控制 HOMO 和 LUMO 能级，不仅能够提高 OPV 材料的性能，同时能够增强其在多样化应用场景中的适用性。从城市建筑的集成到临时建筑的应急供电，OPV 技术能够提供高效且可靠的能源解决方案，进而推动绿色能源的可持续发展。

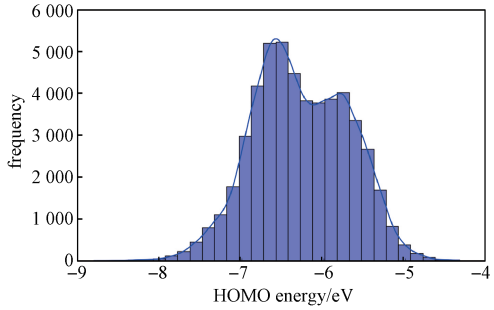
随着光伏技术的不断发展，机器学习作为人工智能领域的关键技术，展现出了其在能源领域的重要应用价值 [4−5]。通过使用先进的计算机算法和模型，机器学习能够从海量数据中自动提取有用的模式和规律，这一能力使得机器学习成为促进跨学科领域研究的强大工具，尤其是在那些缺乏物理或者化学专业背景的研究人员中。在 OPV 领域，机器学习的应用不仅有助于分析和理解新能源材料的性质，而且能够优化新能源器件的设计和性能 [6−9]。特别是在 OPV 的效率和稳定性研究中，机器学习的高效率和低成本特性，使得其成为替代传统试错法的理想选择。例如，文献 [10] 探索了不同的机器语言表达形式和算法对有机光伏材料分子结构与光伏特性之间关系建模的影响，并且提出了一种新的基于机器学习辅助分子设计及效率预测的材料开发流程。G. Brianna 等 [11] 使用包含了 1 225 个给体 / 非富勒烯受体对的新数据集，训练了一个集成机器学习模型来预测器件效率，并且使用遗传算法来探索高性能的非富勒烯受体和聚合物给体，利用它们的组合开发潜在的高效串联电池。这些研究均展示了机器学习在优化 OPV 材料结构与性能方面的潜力。尤其是在 OPV 制造中，面临效率和稳定性挑战，机器学习的应用不仅提高了研究效率，还为 OPV 的大规模制造和商业化应用开辟了新的道路。

基于以上分析，本研究拟充分利用机器学习的高效性和低成本的优势，探索和分析影响 OPV 性能的关键因素。通过 Morgan 指纹（Morgan fingerprint，fp）特征转换，筛选出 20 个关键的 fp 特征，并且构建了 6 种不同的机器学习模型，包括支持向量机、随机森林、梯度提升机等，以全面分析这些特征对 OPV 电池的 HOMO 和 LUMO 能级影响。通过深入分析，最终成功识别了两种影响 HOMO 和 LUMO 能级的关键分子结构，从而揭示了提升 OPV 性能的潜在途径。这些发现不仅为 OPV 材料的分子设计提供

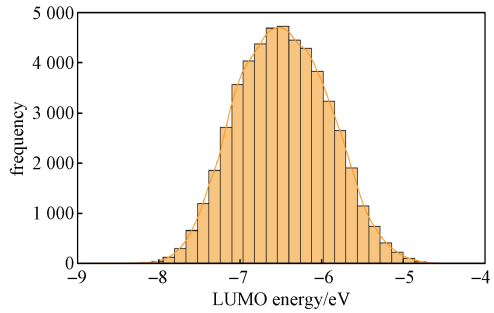
了新的方向，还为分布式新能源系统的进一步发展提供了有力的技术支持。

2　数据库与实验方法

本研究中选用的数据源自哈佛大学的清洁能源数据库 [12]，该数据库包含超过 50 000 种 OPV 分子材料详细信息。数据库中的 HOMO 与 LUMO 能级的概率分布如图 1 所示。



a) HOMO 概率分布



b) LUMO 概率分布

图 1　器件性能的概率分布图

Fig. 1　Probability distribution diagram of device performance

图 1 是基于核密度估计法得出的器件性能分布曲线，观察得知 HOMO 与 LUMO 的能级概率分布均近似正态分布，这为模型准确预测材料的最大潜能提供了丰富的数据基础。

本研究中共使用 6 种机器学习算法，分别是 Nu 支持向量（Nu support vector regression，NuSVR）、极端梯度提升（extreme gradient boosting regressor，XGBT）、随机森林（random forest regressor，RF）、装袋回归（bagging regressor，BR）、k 最近邻（k-nearest neighbors regressor，KNN）、贝叶斯岭回归（Bayesian ridge regressor，BRR）。与这些算法对应的模型均通过 Python 的开源库实现。其中，NuSVR 使用支持向量拟合数据，XGBT 基于梯度提升决策树技术优化预测结果。RF 和 BR 都是决策树的集成学习方法，分别通过平均或者投票机制以及有放回抽样来提高预测的准确性。KNN 算法基于训练数据之间的相似性进行预测，而 BRR 结合了贝叶斯方法和岭回归优化预测。

本实验中，使用 Morgan 指纹作为机器学习模型的输入。Morgan 指纹是一种通过考虑原子的连接性编码分子结构的方法，它迭代地分析分子中每个原子的局部环境，捕捉原子及其邻近原子的独特排列，并将此信息转换为 2 048 位的二进制字符串。在这些字符串中，不同的位点表示分子中是否存在特定的原子局部环境。这种表示法为机器学习模型提供了一种有效的输入特征，以便进行高效的计算处理。

为了系统比较不同机器学习模型在各数据集上的性能，本文采用图 2 所示的实验流程。

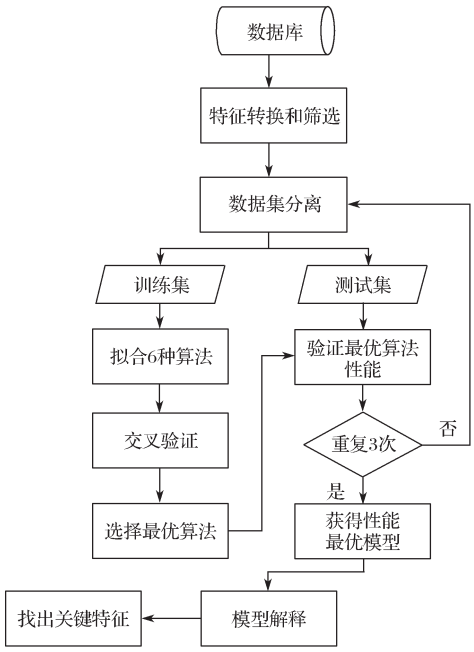


图 2　比较机器学习模型优劣的实验流程图

Fig. 2　Experimental flowchart for comparing the advantages and disadvantages of machine learning models

比较机器学习模型优劣的具体操作如下：首先，从数据库中提取信息，并将其转化为 Morgan 指纹，作为模型的输入特征。接着，通过特征工程去除方差小于 0.1 和高度相关的特征，然后将处理后的数据集拆分为训练集和测试集，比例为 8:2。随后，分别使用 6 种算法对训练集进行拟合，并且通过交叉验证选出最稳定和最准确的算法。最后，利用选定的最优算法对测试集进行 3 次自提取验证，以建立最优模型，并且对模型进行解释，以识别出对 HOMO 和 LUMO 水平影响最大的关键特征。

3　结果与讨论

3.1　特征工程

在本研究中，为避免由于数据集的大规模和高维

特性可能引起的机器学习模型混淆，对输入特征进行筛选，以优化 HOMO 和 LUMO 的预测模型。通过基于算法内置的特征重要性评分，从 2 048 个分子结构的特征中筛选出对 HOMO 和 LUMO 影响最为显著的特征。进一步，通过统计模型对特征间的相关性进行分析，剔除相关系数大于 0.75 的特征以减少共线性，同时去除方差小于 0.1 的低信息量特征，以提高预测效率和预测准确性。最终确定了 20 个关键特征，图 3 为 20 个特征间的相关性热力图。

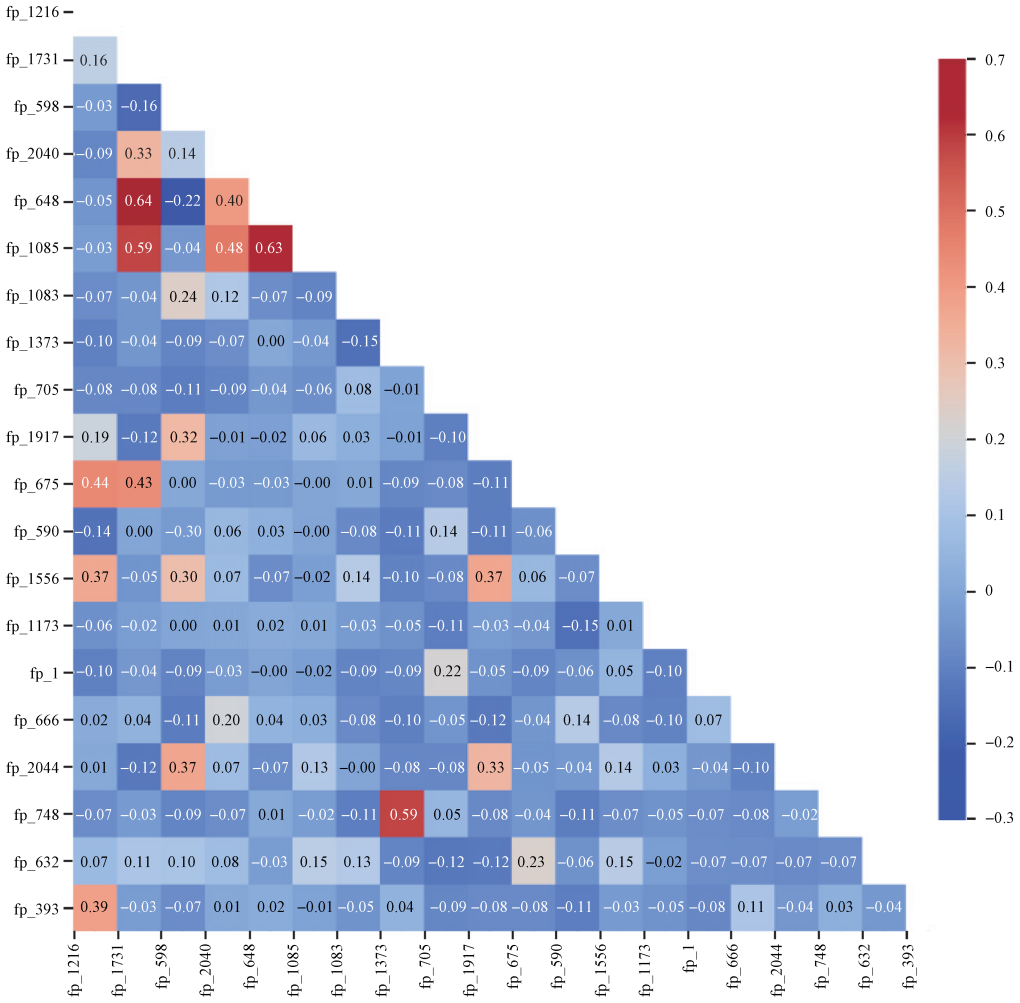


图 3 模型建立所需 20 个特征之间的相关性热力图

Fig. 3 A heatmap of correlation between the 20 features required for model building

由图 3 可见它们的相关性均显著低于 0.7，有效减少了模型维度，降低了过拟合风险，确保模型在新数据上的稳定性和准确性。这一筛选方法为 HOMO 和 LUMO 预测模型提供了一组高效、信息丰富的低维输入特征集，为实现精确预测提供了数据支撑。

3.2 算法选择与泛化性评估

本文中，为了预测 HOMO 和 LUMO，采用 6 种回归模型，模型性能的评估主要基于决定系数（R^2）、均方根误差（root-mean-square error，RMSE）和平均绝对误差（mean absolute error，MAE），这些指标综合反映了模型的拟合程度和预测准确性。所得 6 种机器学习算法拟合训练集时，预测 HOMO 和 LUMO 的 R^2、RMSE、time 指标和 5 折交叉验证时的不同性能表现见表 1。

根据表 1 中的数据，能够对模型性能进行定量比较。若 R^2 值较接近 1，则表明模型在数据拟合上表现良好，而较低的 RMSE 值和 MAE 值则指示了较高的预测准确度。由表 1 中的数据可以得知，在 HOMO 预测任务中，NuSVR 模型展现出了最优性能，具有较高的 R^2 值和较低的 RMSE 值。尽管 XGBT、RF 和 BR 模型的性能略逊于 NuSVR 的，但是它们也显示了相近的良好表现。对于 LUMO 预测任务，XGBT 模型表现最佳，有较高的 R^2 值和较低的 RMSE 值，而 RF 和 BR 虽然次之，但同样展现出不错的性能。相比之下，KNN 和 BRR 模型在两项任务上的表现较差，R^2 值较低，且 RMSE 值较高。

表 1　机器学习算法拟合训练集时的预测模型性能结果

Table 1　Prediction model performance of machine learning algorithms with fitting training sets

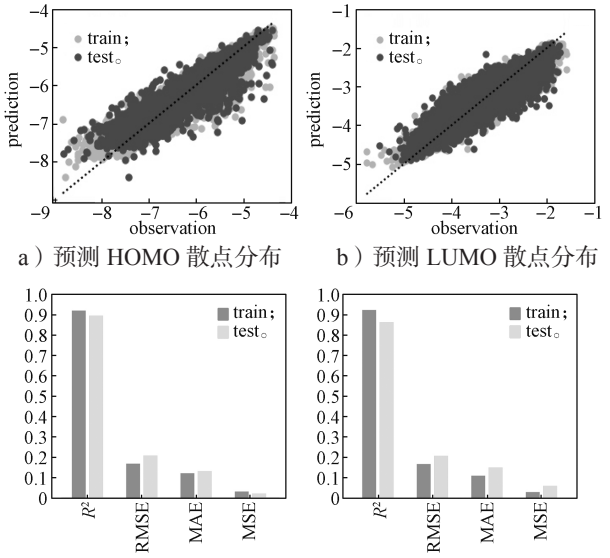
model	object	R^2		MAE 值 (5-fold)		RMSE 值	time/s
		train-set	mean (5-fold)	mean	standard		
NuSVR	HOMO	0.87	0.849	0.161	0.006	0.21	928.65
	LUMO	0.87	0.860	0.163	0.002	0.20	925.81
XGBT	HOMO	**0.85**	**0.832**	**0.145**	**0.001**	**0.22**	**2.62**
	LUMO	**0.84**	**0.840**	**0.149**	**0.002**	**0.21**	**2.67**
RF	HOMO	0.84	0.832	0.154	0.002	0.24	31.34
	LUMO	0.84	0.825	0.163	0.003	0.23	32.7
BR	HOMO	0.82	0.815	0.163	0.003	0.25	3.62
	LUMO	0.82	0.808	0.172	0.003	0.24	3.59
KNN	HOMO	0.73	0.716	0.215	0.003	0.31	1.07
	LUMO	0.72	0.713	0.223	0.002	0.30	1.36
BRR	HOMO	0.62	0.608	0.276	0.001	0.37	0.26
	LUMO	0.60	0.601	0.276	0.002	0.35	0.38

表 1 中还包含了 5 折交叉验证的结果，用以评估模型对于未见数据的泛化性能。在交叉验证中，平均值和标准差越小，表明模型在不同数据子集上的性能波动越小，具有更好的稳定性和泛化能力。对于 HOMO 的预测，NuSVR、XGBT 和 RF 模型在 MAE 值和 R^2 指标上表现最佳，但是 NuSVR 在交叉验证中的 MAE 标准差大于 XGBT 的对应值，显示出更大的性能波动。在预测 LUMO 方面，XGBT 模型在交叉验证中表现出最小的性能波动，暗示其优越的稳定性。除此之外，NuSVR 和 RF 模型在训练过程中需要较长的时间，而 XGBT、KNN 和 BRR 模型能够在约 4 s 内完成拟合。

综上，尽管 NuSVR 模型在某些方面表现出色，但由于其较高的计算成本和在交叉验证中的较大性能波动，本文最终选择 XGBT 作为对数据集最终拟合和优化的算法，并据此构建模型。

3.3　模型评估

完成对 XGBT 模型的优化之后，HOMO 预测模型在训练集和测试集上表现出了高度的相似性，如图 4 所示。具体来说，训练集的 R^2 值为 0.92，而测试集的 R^2 值为 0.87。同样，训练集和测试集的 RMSE 值分别为 0.17 和 0.21，MSE 值分别为 0.03 和 0.04，MAE 值分别为 0.12 和 0.14。这些数据表明 HOMO 模型在两个数据集上具有高度一致的性能。对于 LUMO 预测模型，其表现也相当出色，该模型中训练集和测试集的 R^2 分别为 0.91 和 0.85，RMSE 值分别为 0.17 和 0.21，MSE 值分别为 0.03 和 0.05，MAE 值分别为 0.12 和 0.15。此外，无论是 HOMO 模型还是 LUMO 模型，其散点分布图都显示训练集和测试集数据高度重合，进一步证明了模型的泛化能力和准确性。这些结果表明，经过优化的 XGBT 模型，不仅在训练集上表现出色，而且在未知数据的测试集上也能保持高度的准确度和稳定性，反映出该模型具有极高的泛化能力。



a）预测 HOMO 散点分布　b）预测 LUMO 散点分布

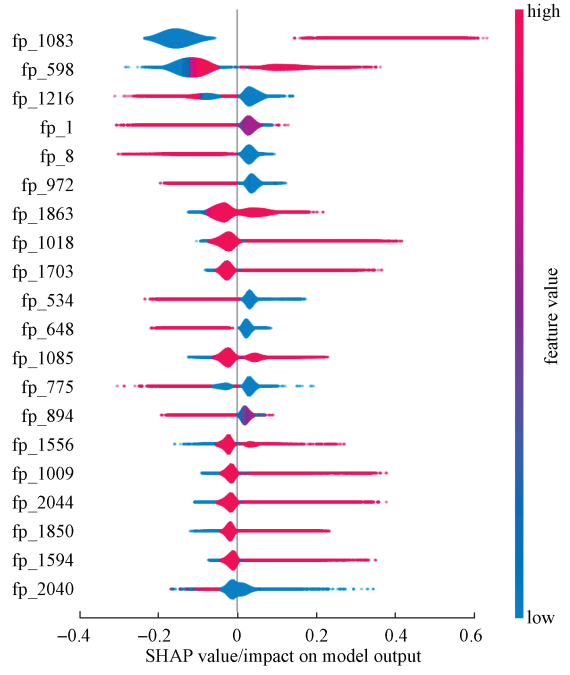
c）预测 HOMO 性能　d）预测 LUMO 性能

图 4　优化后 XGBT 模型预测 HOMO 和 LUMO 的散点分布和性能指标

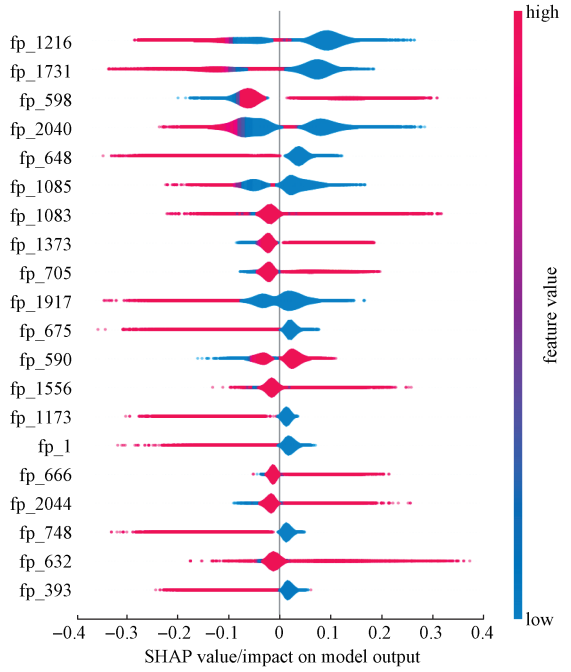
Fig. 4　Scatter distribution and performance indicators of HOMO and LUMO predicted by optimized XGBT models

3.4　模型解释

为了鉴定哪些特征关键地影响 HOMO 和 LUMO 的预测结果，并探索它们在这两个方面作用机制的异同，采用 SHAP（shapley additive explanations）值分析方法对预测模型进行全局分析，所得结果如图 5 所示。



a）预测 HOMO 模型



b）预测 LUMO 模型

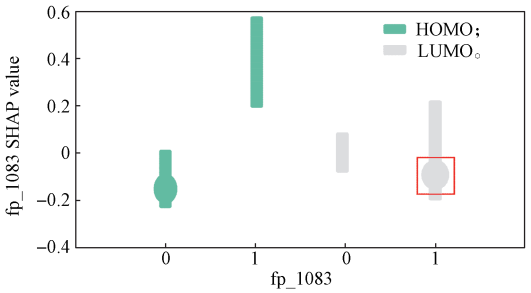
图 5 预测模型全局分析图

Fig. 5 Prediction model global analysis chart

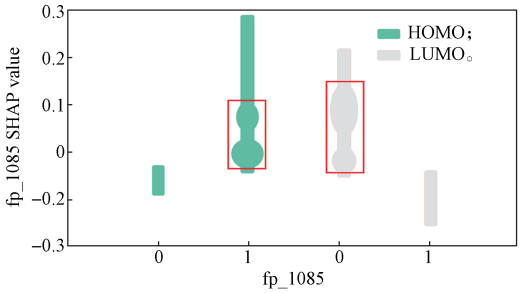
SHAP 分析方法通过色彩变化直观地展示了特征值的大小，横轴上的 SHAP 值显示了特征对模型预测结果的影响方向和强度，而图中的棒状隆起部分显示了每个特征值的概率密度。根据图 5，特征 fp_648、fp_1、fp_2044 在预测 HOMO 和 LUMO 方面表现出了一致的影响趋势，说明这些特征对两者水平的预测产生了相似的影响。相比之下，特征 fp_1083 和 fp_1085 在不同的预测模型中展现出不同的影响趋势，这表明它们在预测 HOMO 和 LUMO 时起到了不同的作用。

基于上述结果，课题组对两个关键特征 fp_1083 和 fp_1085 在模型中的具体影响进行了深入分析，所得结果如图 6 所示。图中横坐标为特征值，这些值指示了特定分子结构的存在；纵坐标显示了特征的 SHAP 值的波动范围，SHAP 值量化了每个特征对模型输出的影响程度。通过这种分析，可以精确评估这些特征对 HOMO 和 LUMO 能级影响的具体数值，从而深入理解它们在影响 HOMO 和 LUMO 方面的作用机制。

特别地，课题组观察到当分子中含有 fp_1083 结构时，HOMO 的能级提高了大约 0.2~0.6 eV，而 LUMO 的能级下降了约 0.1 eV。这表明 fp_1083 结构导致能级差增大了约 0.3~0.7 eV。另一方面，fp_1085 结构的存在使 LUMO 的能级下降了约 0.2 eV，去除该结构后 LUMO 的能级提高了约 0.1 eV，意味着 fp_1085 结构会使能级差减少约 0.2 eV。当两种结构共存时，能级差可扩大至 0.1~0.5 eV。



a）fp_1083 特征

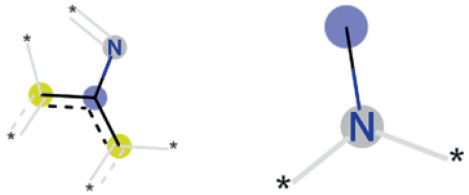


b）fp_1085 特征

图 6 预测模型特征局部分析图

Fig. 6 Local analysis chart of prediction model features

这一发现能够有效提升 OPV 的性能。能级差增大会导致光谱吸收范围扩展，这直接提高了光电转换效率。更宽的吸收范围能够促进光生载流子的有效移动和分离，从而增强了 OPV 电池的稳定性和电流输出能力。因此，通过精确控制分子结构，尤其是控制 fp_1083 和 fp_1085 的结构，能够优化 OPV 电池的性能，实现更高效的能量转换。通过最终分析，得出的两种能够改变器件能极差的分子结构，如图 7 所示。



a）fp_1083 分子结构　　b）fp_1085 分子结构

图 7 两种改变器件能级差的分子结构示意图

Fig. 7 Schematic representation of two molecular structures changing the energy level difference of the devices

4 结论

本文利用机器学习技术深入分析了有机太阳能电池（OPV）中分子特征对 HOMO 和 LUMO 能级的影响。基于哈佛大学的清洁能源数据库，筛选出 20 个关键特征，并使用包括 NuSVR、XGBT、RF、BR、KNN 和 BRR 在内的 6 种机器学习算法，对 OPV 性能的影响因素进行了全面分析。

1）在模型性能评估中，发现基于梯度提升的 XGBT 模型在预测 HOMO 和 LUMO 性能方面表现最优，其高决定系数和低均方根误差证明了模型高效的预测能力。

2）通过 5 折交叉验证，验证了模型的稳定性和泛化能力。XGBT 模型不仅在训练集上表现出色，而且在未知数据的测试集上也保持了高度的准确度和稳定性，显示出优秀的泛化能力。

3）通过 SHAP 值分析，成功识别了影响 OPV 能级的关键分子结构。特别是发现 fp_1083 和 fp_1085 的分子结构在调整能级差方面起着重要作用。这些发现为有机太阳能电池的分子设计提供了新的视角，对提高光伏器件的光电转换效率和稳定性具有重要意义。

综合以上结果，本文不仅展示了机器学习在分析和优化 OPV 材料方面的潜力，还为分布式新能源的发展贡献了重要的理论和实践指导。未来，这些发现将有助于进一步提高 OPV 技术的商业化应用前景，推动分布式可再生能源的整体发展。

参考文献：

[1] ULLAH F，CHEN C C，CHOY W C H. Recent Developments in Organic Tandem Solar Cells Toward High Efficiency[J]. Advanced Energy and Sustainability Research，2021，2(4)：2000050.

[2] IQBAL W，ULLAH I，SHIN S. Optical Developments in Concentrator Photovoltaic Systems：A Review[J]. Sustainability，2023，15(13)：10554.

[3] CHENG P，WANG J Y，ZHANG Q Q，et al. Unique Energy Alignments of a Ternary Material System Toward High-Performance Organic Photovoltaics[J]. Advanced Materials，2018，30(28)：e1801501.

[4] KRANTHIRAJA K，SAEKI A. Experiment-Oriented Machine Learning of Polymer：Non-Fullerene Organic Solar Cells[J]. Advanced Functional Materials，2021，31(23)：2011168.

[5] PADULA D，SIMPSON J D，TROISI A. Combining Electronic and Structural Features in Machine Learning Models to Predict Organic Solar Cells Properties[J]. Materials Horizons，2019，6(2)：343−349.

[6] SUN W B，ZHENG Y J，YANG K，et al. Machine Learning-Assisted Molecular Design and Efficiency Prediction for High-Performance Organic Photovoltaic Materials[J]. Science Advances，2019，5(11)：eaay4275.

[7] WU Y，GUO J，SUN R，et al. Machine Learning for Accelerating the Discovery of High-Performance Donor/Acceptor Pairs in Non-Fullerene Organic Solar Cells[J]. NPJ Computational Materials，2020，6：120.

[8] 王　钋，雷　敏，梁娇娇，等 . 基于 IPSO-GRU 的锂离子电池剩余使用寿命预测 [J]. 湖南工业大学学报，2022，36(4)：23−30.
WANG Po，LEI Min，LIANG Jiaojiao，et al. An IPSO-GRU-Based Prediction of Remaining Useful Life of Lithium-Ion Batteries[J]. Journal of Hunan University of Technology，2022，36(4)：23−30.

[9] 雷　敏，徐　波，华一飞，等 . 基于 SHEKF-GPM 融合的锂电池 SOC 估算 [J]. 湖南工业大学学报，2020，34(6)：10−15.
LEI Min，XU Bo，HUA Yifei，et al. State of Charge Estimation of Lithium Battery Based on SHEKF-GPM Fusion[J]. Journal of Hunan University of Technology，2020，34(6)：10−15.

[10] MEFTAHI N，KLYMENKO M，CHRISTOFFERSON A J，et al. Machine Learning Property Prediction for Organic Photovoltaic Devices[J]. NPJ Computational Materials，2020，6：166.

[11] BRIANNA G，GEOFFREY H. Screening Efficient Tandem Organic Solar Cells with Machine Learning and Genetic Algorithms[J]. The Journal of Physical Chemistry C，2023，127(13)：6179−6191

[12] STEVEN L，EDWARD P，GREGOR S，et al. The Harvard Organic Photovoltaic Dataset[J]. Scientific Data，2016，3：160086.

（责任编辑：廖友媛）