doi:10.6043/j.issn.0438-0479.202105015

卷积神经网络在新型非富勒烯受体分子生成与性能预测上的应用

杨新玉，彭师平，赵　仪 *

（厦门大学化学化工学院，固体表面物理化学国家重点实验室，福建省理论与计算化学重点实验室，福建 厦门 361005）

摘要：近年来，有机太阳能电池中非富勒烯小分子受体因其拓展了吸收光谱的范围、能够调节激子解离能量和具有灵活的给体-受体形貌等优点使得器件效率越来越接近产业化的目标。本文借助本组之前构建的分子生成和性质预测的卷积神经网络模型，来生成和筛选出具有满足高效解离激子的前线轨道能量（Y6 分子，最高占据分子轨道(HOMO)目标能量为-5.73 eV 和最低未占据分子轨道(LUMO)目标能量为-3.69 eV）的新型非富勒烯小分子受体。首先生成模型经数据库充分训练并利用小数据集进行微调后生成出 200 多个接近目标轨道能量的分子，然后利用预测模型进一步筛选并预测分子片段对轨道能量的贡献，接着将这些分子与数据库中具有相近前线轨道能量的分子共同聚类挑选出具有不同化学空间的 10 个新型受体分子，最后通过从头算验证了轨道能级、分子片段对轨道贡献预测的准确性，并给出了分子光吸收的振子强度。我们也进一步利用生成和预测模型提供了具有另一组轨道能量（HOMO 能量为-5.10 eV 和 LUMO 能量为-3.10 eV）的 10 个非富勒烯小分子，性质预测与从头算结果一致，证明了生成和预测模型的鲁棒性和结果的可靠性。本文预测的分子也提供了设计具有高性能非富勒烯受体分子骨架的思路。

关键词：卷积神经网络；非富勒烯受体；前线分子轨道能量

中图分类号： O **641**　　**文献标志码：** A

有机太阳能电池(organic solar cells, OSCs)是一种将太阳光能转化为电能的器件，与无机太阳能电池相比，OSCs 具有材料来源广泛、工艺简单、轻便、生产容易的优点，在便携式电源[1]、可穿戴设备[2-3]、室内小型离网电子设备[4]等领域展示出了光明的产业前景，在过去的 20 多年间得到了迅速发展，近年来受到了广泛的关注。

OSCs 主要由电子给体和受体材料组成，其中给体常采用低能隙的聚合物或小分子，而受体多使用电子亲和性较大的富勒烯分子[5-7]。近年来，人们发现非富勒烯受体(non-fullerene acceptors, NFAs)的化学结构和电子亲和性可在大范围内进行调整，同时它还具有较大的近红外吸收范围、较好的能

收稿日期：2021-05-07　　录用日期：2021-06-30
基金项目：国家自然科学基金（22033006, 21833006，21773191）
*通信作者：yizhao@xmu.edu.cn

级匹配、较小的电压损失等特性[8]。NFAs 新型材料的不断发展推动了 OSCs 性能的迅速提高，目前其能量转换效率(power conversion efficiencies, PCE)已达到 18.69%[9]，激发了越来越多研究人员对高性能材料探索发现的热情。

近几年来，NFAs 的研究在中国科学家的推动下得到了蓬勃发展，电池的 PCE 也得到了显著提升。Lin 等[10]于 2015 年提出稠环电子受体概念和具有(A - D - A)结构的 ITIC 有机小分子，这类分子包含一个稠环给电子中心骨架和两个强吸电子端基，给体单元和受体单元之间发生强的分子内电荷转移，使得受体具有较窄的带隙和很强的可见光与近红外光吸收，同时端基单元的紧密堆积有利于电子传输，整个受体具有较高的迁移率，ITIC 与聚合物给体 PTB7-Th 共混的器件 PCE 达到了 6.8%。ITIC 打破了基于富勒烯受体的 OSCs 效率进一步提高的瓶颈，之后越来越多的研究由此展开，通过对中心给体单元、侧链及末端受体单元的扩大、取代、异构化等手段，OSCs 器件的 PCE 不断提升[11-13]。Yuan 等[14]于 2019 年发现了一种通过在中心部分引入苯并噻二唑吸电子单元的窄带隙 A - D - A' - D - A 型受体 Y6，PM6 与 Y6 共混得到的器件 PCE 高达 15.7%。在 2020 年，Liu 等[15]合成了给体 D18，D18 的空穴迁移率高达 1.59×10^{-3} $cm^2/(V\cdot s)$，之后，该课题组又设计了新的聚合物给体 D18-Cl，其中 D18 : Y6、D18-Cl : N3 和 D18 : N3 的器件 PCE 分别达到了 18.22%[15]、18.13%[16]和 18.56%[17]，D18-Cl : N3 : PC61BM (D : A1 : A2)型的器件 PCE 更是达到了 18.69%[9]。至今为止，A - D - A 或 A - D - A' - D - A 型的模式是 OSCs 的主流构筑方式，新型分子的出现将加快更高效率的实现。

上述受体材料表现出高性能的原因之一是给体和受体能级匹配，这可为电荷分离和转移提供有效驱动力，所以其直接影响着电池效率的提升。然而，大多数近红外吸收的有机分子的最低未占据分子轨道(lowest unoccupied molecular orbital, LUMO)和最高占据分子轨道(highest occupied molecular orbital, HOMO)能级很难与宽禁带给体的能级相匹配，对于正确选择给体和非富勒烯受体被认为是一项费时且复杂的任务。随着数据科学的发展，机器学习的应用推动多领域的变革，也影响着材料化学的研究，目前利用机器学习模型针对 OSCs 材料分子的前线分子轨道(frontier molecular orbital, FMO)能量的研究已取得巨大进展。首先，机器学习可以实现对 FMO 能量的预测，例如 Pereira 等[18]在由 111 000 个分子组成的数据集上训练随机森林等模型，在没有任何密度泛函理论(density functional theory, DFT)计算的情况下模型预测的 HOMO 和 LUMO 能量误差均小于 0.16 eV。通常，训练预测模型的数据来源于计算或实验，需通过校准来减少计算值与实验值的偏差，如 Lopez 等[19]在建立了 51 000 多个由碎片拼接而成的 NFAs 分子及其 HOMO、LUMO 能量的数据库后，利用 94 组实验值通过高斯过程回归模型校准了计算值，这使得 HOMO 能量的均方根误差(root mean square error, RMSE)由校准之前的 0.28 eV 降为校准之后的 0.17 eV，LUMO 能量的 RMSE 也从 0.45 eV 降至 0.26 eV。此外，FMO 能量等可作为描述符来预测器件 PCE，在获得更高预测精度的同时证明了其对 PCE 的重要影响[20-21]。这些研究在加快 NFAs 分子的筛选效率上起到了重要作用。目前利用机器学习对 NFAs

及其 FMO 的研究主要集中于提高预测精度和效率上，而利用机器学习对分子的结构与性质之间关系的研究却相对较少。

本文将利用课题组已提出的基于卷积神经网络(convolutional neural networks, CNN)构建的分子生成模型与性质预测模型[22]，使用生成模型快速得到多个具有特定 HOMO、LUMO 能量范围且结构差异性较高的 NFAs 分子，利用基于注意力机制的预测模型验证分子的 FMO 性质并得到分子碎片对性质的贡献。本研究能够在对非富勒烯有机小分子受体筛选的同时进行其结构与性质关系的研究，希望能够对新材料的发现带来一些启发。

1 数据库与模型

1.1数据库

用于训练神经网络模型的数据源于 Aspuru-Guzik 等于 2017 年提出的含 51 281 种潜在 NFAs 材料的数据集，该数据库中提供了每个分子的简化分子线性输入规范(simplified molecular input line entry specification, SMILES)[23]表示和 HOMO、LUMO 能量等值[19]。其中的分子是由包括萘二酰亚胺、苯并噻二唑和聚氟蒽二亚胺等 107 种常见基团拼接而成，每种分子碎片与其取代方式均通过文献或商用例子获得，其中，碎片共分为 13 种中心碎片(cores, C)，49 种侧位碎片(spacers, S)、45 种端位碎片(terminals, T)，分子的拼接方式有 T - S - C - S - T、T’ - S - C - S - T、T - C - T、T - S - T。数据库中每个分子 HOMO、LUMO 能量的计算大致分为 4 步：1)使用 RDKit[24]提供的构象生成器由 SMILES 编码生成 1500 个三维分子构象；2)对所有构象进行分子力场[25]优化，使用 OpenBabel[26]软件去除重复构象；3)按照最低能量原理对于每个分子挑选出 20 个构象，同时所选分子构象的能量与最低分子构象的能量差应不超过 5 kcal/mol，如果超过 5 kcal/mol，则剔除该构象（这种情况下，构象数少于 20 个），这些构象所组成的簇被认为是候选分子在固态中最具能量可行性的构象；4)用 BP86/def2-SVP 泛函基组对上述构象进行优化，之后用 B3LYP/def2-SVP 泛函基组做单点能、HOMO、LUMO 能量计算，提取具有最低能量的构象，将该构象的 HOMO 与 LUMO 能量视为该分子的 FMO 能量。我们对数据进行了简单的预处理，删除能隙值为负的不合理分子后，实际用于模型训练的数据数量为 50 656 个。

1.2分子表示

分子图(Graph)[27]和 SMILES 是分子生成模型常用的分子表示。用分子图作为输入时，分子中的原子和原子之间的键被表示成图的节点和边。分子图在对抗生成网络中的分子生成表现优异，然而，基于分子图的模型现今只能生成小分子。SMILES 通过使用一系列字符来表示分子结构，这些字符通过原子符号和拓扑特征来表示原子，如带有特殊字符的键和带有括号的支链。若没有固定顺序的

读取原子和键来生成 SMILES，特定分子可生成多个有效的 SMILES 字符串。为此，常使用规范化的 SMILES 保证分子的唯一性表示来克服同一分子生成字符串的多样性。生成和预测模型均基于 CNN 并使用 SMILES 作为分子表示，一方面，CNN 具有权值共享和可处理多个时间步长的特点，效率较高；另一方面，SMILES 表示已经被广泛使用在多种神经网络模型中[28-29]，一维 CNN 也可处理不同长度的 SMILES 表示。

1.3生成模型与性质预测模型

分子生成与性质预测模型均为本组之前基于一维 CNN 所建立的模型，更多模型信息可由 https://github.com/PSPhi/CNN-for-NFA[22]获得。CNN 利用卷积核（参与运算的矩阵）与节点的矩阵运算，可实现特征提取，主要应用于处理图像、视频、语音、音频等方面。由于 CNN 具有共享权重和平移不变性的特点，可同时处理多个时间步长，能够显著提高深度学习效率。这里我们运用的一维的 CNN 可处理不同长度的 SMILES 输入问题。对于生成模型，在训练过程中给每个输入的 SMILES 字符串加上起始字符“&”，给每个目标输出加上“\n”，我们的模型需要能够通过给定的起始字符，逐个生成后续字符直到终止字符“\n”被生成，从而完成一个 SMILES 字符串即分子的生成。预测模型在卷积网络之后的输出层加入了融合信息的注意力机制，通过注意力机制能够获得每个字符对于对应性质的重要性。

从数据库中挑选出 PCE 大于 0.5 并且信息完整的共 24 000 个分子，将这些分子随机划分成分别含有 20 000，2 000，2 000 个分子的训练集、验证集和测试集，分割后的数据集将用于生成模型和预测模型的训练。对于生成模型，训练好的模型所生成的分子中合理分子的比例高达 90%。预测模型对于测试集中分子 HOMO、LUMO 能量预测的平均绝对误差分别为 0.053 和 0.055 eV。

2 应　用

基于提出的分子生成模型和HOMO、LUMO能量的预测模型，下文将探索利用生成模型生成两组指定HOMO、LUMO能量的分子，并用预测模型对分子轨道能量做进一步预测来筛选分子，最后用DFT计算进行验证。这一工作可进一步拓展数据库的化学空间、为实验工作提供分子选取的思路。

实验上，聚合物给体D18与非富勒烯受体Y6和Y6侧链进行优化得到的N3共混的器件PCE分别达到了18.22%[15]和18.56%[17]，Y6也将电池效率推上了一个新台阶，因此选取Y6的FMO能量作为参考值，Y6的HOMO、LUMO实验值分别为-5.65和-4.10 eV[30]。由于本文所采用的模型均由数据库提供的计算值训练所得，并且，计算值与实验值之间存在计算方法的系统误差，所以本文使用与数据库一致的构象选取方式和泛函基组，采用BP86/def2-SVP泛函基组对构象进行优化并采用B3LYP/def2-SVP泛函基组做单点能、HOMO、LUMO能量计算，得到Y6的HOMO、LUMO能量计算值分别为-5.73

和-3.69 eV，结合数据库中的数据分布，我们选取计算值为-5.60和-3.60 eV这一组值作为在同等计算方式下分子生成和筛选的目标计算值。

分子的生成和筛选过程如图1所示，其中绿色部分表示的是使用原数据库训练生成模型和预测模型的过程，蓝色部分为给定的FMO能量目标值，橙色部分为本文所主要强调的针对目标值的分子生成和筛选过程。其中橙色部分的流程主要有4步：1)从原数据库中得到HOMO、LUMO能量在所选定值误差范围内的小数据集，记为D1；2)用D1分子集重新训练已由原数据库训练好的生成模型，实现对模型参数的微调，这样可使得模型倾向于生成目标能量值附近的分子，微调后的模型生成新分子集D2；3)对D2分子集中的分子进行处理，剔除重复、不合理以及原数据库中已有的分子，并通过预测模型从中筛选出HOMO、LUMO能量在误差范围内的分子，即可得到候选的新分子集D3；4)将由原数据库得到的D1分子集和新生成的D3分子集进行合并，使用RDKit软件包提供的最大最小聚类算法[31]，该算法通过从分子的SMILES表示中计算出分子指纹，再根据分子指纹距离的计算将分子划分到相应的类。聚类算法可从整个分子库中挑出一个多样性最高的子集，来最大程度地代表原始分子库的化学空间[32]。通过聚类可得到多样性最大的10个分子组成的集合D4，D4中的分子即为最终挑选出的分子。根据目标值挑选出的分子结构如图2中a1~a10所示，其中，a1~a7为新生成的分子，a8~a10为数据库中原有的分子，根据数据库的碎片种类可将分子划分成不同颜色碎片的组合。

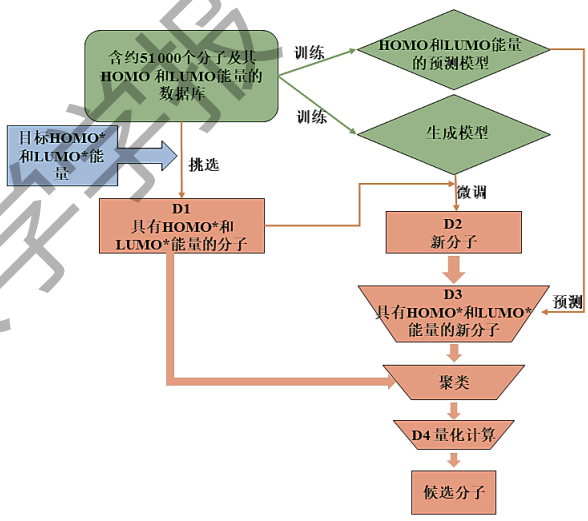

图 1 分子的生成和筛选过程示意图

Fig.1 Schematic diagram of molecular generation and screening process

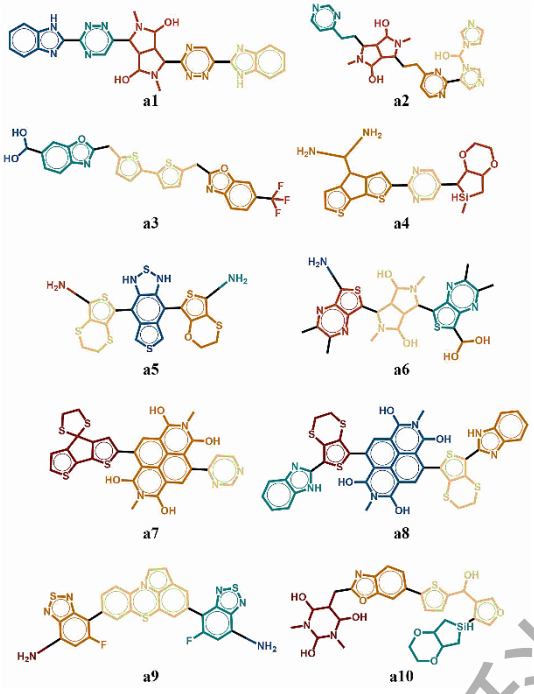

图 2 根据第一组目标值挑选出的分子结构图

Fig.2 Molecular structures selected according to the first set of target values

为了验证新生成的分子FMO能量是否在目标值误差范围内，进行了同等计算水平的计算，第一组的目标值得到的10个分子的HOMO、LUMO的计算值和通过预测模型得到的预测值如表1所示。

表 1 第一组分子的前线分子轨道能量的计算值和预测值

Tab.1 The calculated and predicted values of the front molecular orbital energies for the first group of molecules

分子	HOMO /eV			LUMO /eV		
	预测值	计算值	误差值	预测值	计算值	误差值
a1	-5.77	-5.77	0.00	-3.68	-3.60	-0.08
a2	-5.60	-5.60	0.00	-3.59	-3.53	-0.06
a3	-5.44	-5.46	0.02	-3.66	-3.53	-0.13
a4	-5.54	-5.55	0.01	-3.57	-3.42	-0.15
a5	-5.57	-5.60	0.03	-3.67	-3.58	-0.09
a6	-5.59	-5.63	0.04	-3.70	-3.64	-0.06
a7	-5.64	-5.61	-0.03	-3.61	-3.53	-0.08
a8	-5.53	-5.54	0.01	-3.57	-3.54	-0.03
a9	-5.58	-5.58	0.00	-3.60	-3.46	-0.14
a10	-5.65	-5.63	-0.02	-3.45	-3.53	0.08

从表1中可以看出，除了a3、a4、a9分子LUMO的预测值和计算值相差超过0.1 eV以外，其他的性质预测的绝对误差均小于0.1 eV，说明预测模型具有较高的准确度。同时，筛选出分子的HOMO、LUMO计算值与目标值的绝对误差均小于0.2 eV，有些分子如a2、a5、a6、a7的计算值甚至很接近目标值，经过后续的分子修饰，HOMO、LUMO的实验值可进一步调整以实现与给体分子的能级匹配。

为了比较所选分子的差异性，一方面，通过由分子的SMILES出发得到分子指纹，再根据分子指纹得到不同分子之间的相似度[33]，结果如图3所示。另一方面，结合预测模型可以获得每个分子的SMILES表示中每个字符对相应性质的贡献，因为数据库的分子是由碎片拼接而成，同样地，通过片段所含字符贡献的加和，我们可以得到组成每个分子的每个碎片对HOMO、LUMO性质的贡献程度。作为参考，使用Multiwfn[34]程序得到HOMO、LUMO在每个原子上的分布百分比，将片段所含原子的分布进行加和得到片段的分布百分比，这可以在一定程度上反应预测模型对片段贡献预测的准确性。使用预测模型和Multiwfn程序得到的对HOMO、LUMO能量贡献程度最大的片段和相应的比例如表2所示，其中片段的颜色与图2一致。

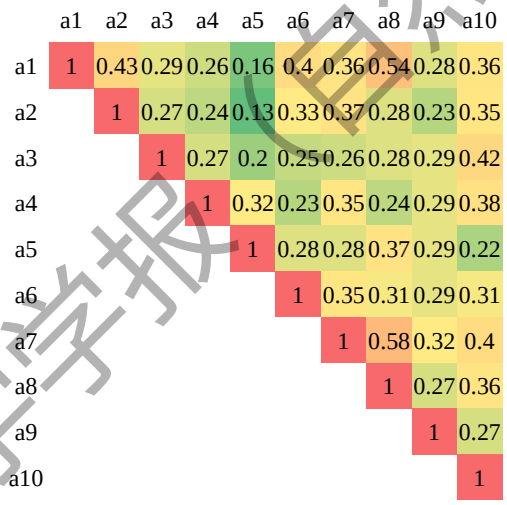

图 3 第一组分子的分子间相似度矩阵图

Fig.3 Intermolecular similarity matrix of the first group of molecules

表 2 使用预测模型和 Multiwfn 程序得到的第一组分子中对 HOMO、LUMO 能量贡献程度最大的片段和相应的比例

Tab.2 The fragments and corresponding proportions of the first group of molecules that contribute the most to the HOMO and LUMO energy obtained by using the prediction model and the Multiwfn program

分子	HOMO				LUMO			
	预测模型		Multiwfn		预测模型		Multiwfn	
	碎片	比例	碎片	比例	碎片	比例	碎片	比例
a1	红	0.74	红	0.44	红	0.47	红	0.41
a2	红	0.79	红	0.35	红	0.43	红	0.44
a3	黄	0.70	黄	0.47	黄	0.72	黄	0.58

a4	橙	0.45	橙	0.56	橙	0.99	橙	0.87
a5	蓝	0.90	蓝	0.39	蓝	0.96	蓝	0.75
a6	黄	0.85	黄	0.36	红	0.80	红	0.35
a7	红	0.97	红	0.63	橙	0.98	橙	0.77
a8	红	0.26	红	0.26	蓝	0.94	蓝	0.78
a9	黄	0.94	黄	0.59	橙	0.85	橙	0.36
a10	绿	0.58	绿	0.96	红	0.55	红	0.71

从图 3 相似度矩阵图中可以得到分子两两之间的相似度且相似度的大小可由颜色的深浅表示，颜色越绿表示相似度越低，分子的差异性越大，反之，颜色越红则表示分子越相似。图中，对角线表示分子与自身的相似度，即为 1，可以看到，只有两组分子 a1-a8 和 a7-a8 的相似度较大，为 0.54 和 0.58，其他分子间的相似度均小于 0.5 且大部分在 0.2~0.3，可见，挑选出的分子具有较大的差异性。对于第一组的每个分子，我们用预测模型和 Multiwfn 程序这两种方法获得了对 FMO 性质贡献最大的片段和对应的贡献值，如表 2 所示。表中，第二列与第四列、第六列与第八列描述的是用两种方法获得的对性质贡献最大的片段的颜色，它们是一致的，说明预测模型能够准确预测出最重要的片段；第三列与第五列、第七列与第九列的数值表示的是用两种方法获得的相应碎片的贡献值，它们之间有些存在较大差异。需要说明的是，Multiwfn 程序得到的碎片上 FMO 分布的比例仅为参考值，预测模型得到的是碎片对所预测轨道能量的重要程度，两者表示的性质相同但是计算方式不同，因此数值上存在差异。总得来看，具有相近 HOMO、LUMO 能量的分子可以具有不同的结构，且其中对两者影响最大的碎片也可不同，进一步说明了存在多种结构的受体可以与给体能级匹配。

为了进一步验证这些受体分子的吸光性能，计算其振子强度，如表 3 所示。

表 3 第一组分子的第一、二激发态能量和对应的振子强度

Tab.3 The first and second excited state energies and the corresponding oscillator intensities of the first group of molecules

分子	激发态 1		激发态 2	
	能量/eV	振子强度	能量/eV	振子强度
a1	2.03	0.87	2.21	0.00
a2	2.00	0.66	2.81	0.00
a3	2.06	2.21	2.71	0.00
a4	1.57	0.04	2.36	0.01
a5	1.73	0.21	2.03	0.01
a6	1.78	0.44	2.14	0.00
a7	1.77	0.27	2.59	0.04
a8	1.54	0.18	1.62	0.00

a9	1.76	0.16	1.93	0.03
a10	2.65	0.57	3.00	0.00

从表 3 中的能量和振子强度数据中可以看出，a1、a2、a3、a6、a10 分子的振子强度较大，具有较为优异的可见光吸收性能，可被后续修饰为潜在的受体材料。

此外，我们采用相同方式得到了 HOMO、LUMO 能量的计算值分别为 5.10 和 3.10 eV 的第二组 10 个分子，一方面，这些分子可以作为参考，另一方面，也可为新的给体提供思路。第二组分子的结构如图 4 的 b1~b10 所示，其中，b1~b8 为新生成的分子，b9 ~b10 为数据库中原有的分子。同样计算了新生成分子的 HOMO 和 LUMO 能量，第二组 10 个分子的 FMO 能量如表 4 所示。

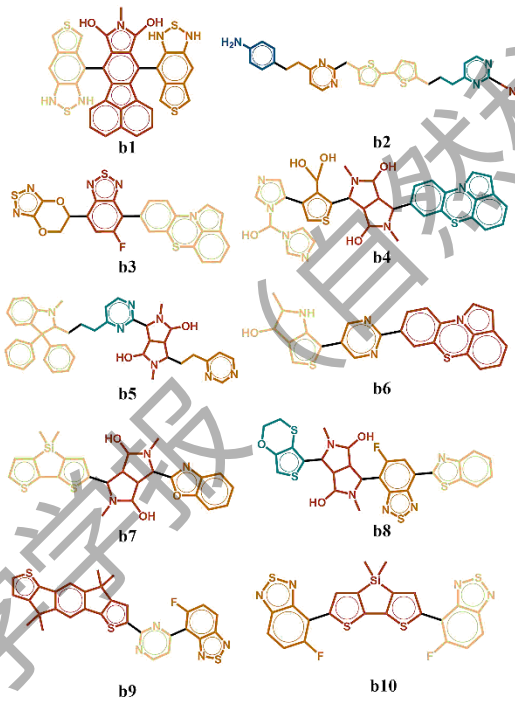

图 4 根据第二组目标值挑选出的分子结构图

Fig.4 Molecular structures selected according to the second set of target values

表 4 第二组分子的前线分子轨道能量计算值和预测值

Tab.4 The calculated and predicted values of the front molecular orbital energies for the second group of molecules

分子	HOMO /eV			LUMO /eV		
	预测值	计算值	误差值	预测值	计算值	误差值
b1	-5.22	-5.31	0.09	-3.22	-3.06	-0.16
b2	-5.20	-5.08	-0.12	-3.19	-3.22	0.03
b3	-5.24	-5.24	0.00	-3.16	-3.08	-0.08

b4	-5.25	-5.30	0.05	-3.07	-2.99	-0.08
b5	-5.13	-5.06	-0.07	-3.04	-3.11	0.07
b6	-5.13	-5.13	0.00	-3.04	-3.11	0.07
b7	-5.09	-5.13	0.04	-3.02	-2.93	-0.09
b8	-5.32	-5.27	-0.05	-3.19	-3.15	-0.04
b9	-5.20	-5.16	-0.04	-2.92	-2.87	-0.05
b10	-5.24	-5.22	-0.02	-2.98	-2.85	-0.13

表 4 的数据显示，仅有 b2 分子的 HOMO 和 b1、b10 分子的 LUMO 能量计算值与预测值的误差超过 0.1 eV。然而，第二组分子计算值与目标值偏离较大的分子比第一组多，如 b1、b4 的 HOMO 能量和 b9、b10 的 LUMO 能量，需要实验后续的修饰来进一步调整能级，如增加或减少吸电子、给电子和共轭基团等。同样，对其吸光性能进行验证，如表 5 所示。可以看出，b2、b4、b5、b7、b9 和 b10 分子具有较大的振子强度，有望成为光吸收能力优异的受体材料。

表 5 第二组分子第一、二激发态能量和对应的振子强度

Tab.5 The first and second excited state energies and the corresponding oscillator intensities of the second group of molecules

分子	激发态 1		激发态 2	
	能量/eV	振子强度	能量/eV	振子强度
b1	1.93	0.01	1.93	0.01
b2	2.01	1.63	2.12	0.44
b3	1.77	0.07	2.98	0.08
b4	2.02	0.43	2.46	0.26
b5	1.58	0.00	2.15	0.42
b6	1.78	0.07	2.42	0.18
b7	2.13	0.93	2.85	0.14
b8	1.74	0.25	2.41	0.02
b9	1.97	0.00	3.00	1.31
b10	2.10	0.68	2.27	0.01

3 结　论

利用卷积神经网络模型，我们生成并筛选出了 HOMO 和 LUMO 能量分别为-5.60 eV 和-3.60 eV、

-5.10 eV 和-3.10 eV 的两组受体分子，来匹配有机太阳能电池中不同给体分子所需的激子解离能。分析发现，尽管每组分子具有相同的 FMO 能量，但通过分子指纹的距离计算可说明它们的相似度具有较大差异，表明生成的分子覆盖了较广的化学空间。通过进一步的量子化学计算发现，这些分子中约 55%的分子具有较大的振子强度即较好的吸光能力。这些生成的具有给定 FMO 能量的分子可提供设计受体分子骨架的思路，这将加快新材料的发现和结构性质关系的研究。

参考文献：

[1] LI Y, Xu G, CUI C, et al. Flexible and semitransparent organic solar cells. Advanced Energy Materials, 2018, 8(7):1701791.

[2] CUI Y, HONG L, HOU J. Organic photovoltaic cells for indoor applications: opportunities and challenges[J]. ACS Applied Materials & Interfaces, 2020, 12(35):38815-38828.

[3] LANDERER D, BAHRO D, ROHM H, et al. Solar Glasses: A case study on semitransparent organic solar cells for self-Powered, smart, wearable devices[J]. Energy Technology, 2017, 5(11):1936-1945.

[4] BEUEL S, HARTNAGEL P, KIRCHARTZ T. The influence of photo-induced space charge and energetic disorder on the indoor and outdoor performance of organic solar cells[J]. Adv Theor Simul, 2021, 4(3): 2000319.

[5] Wu W P, Deng L L, Xiang L, et al. Theoretical insight into the stereometric effect of bisPC71BM on polymer cell performance[J]. Science Bulletin, 2016, 61(2):139-147

[6] Deng L L, Li X, Wang S, et al. Stereomeric effects of bisPC71BM on polymer solar cell performance[J]. Science Bulletin, 2016, 61(2):132-138.

[7] Liu T, Troisi A. What Makes Fullerene Acceptors Special as Electron Acceptors in Organic Solar Cells and How to Replace Them[J]. Advanced Materials, 2013, 25(7): 1038-1041.

[8] LIU W, XU X, YUAN J, et al. Low-bandgap non-fullerene acceptors enabling high-performance organic solar cells[J]. ACS Energy Letters, 2021, 6(2):598-608.

[9] JIN K, XIAI Z, DING L M. 18.69% PCE from organic solar cells[J]. Journal of Semiconductors, 2021, 42(6):060502.

[10] LIN Y, WANG J, ZHANG Z, et al. An Electron Acceptor Challenging Fullerenes for Efficient Polymer Solar Cells[J]. Advanced Materials, 2015, 27(7):1170-1174.

[11] JIANG Z, LI H, WANG Z, et al. Naphtho[1,2-b:5,6-b']dithiophene-based conjugated polymers for fullerene-free inverted polymer solar cells[J]. Macromolecular Rapid Communications, 2018, 39(14):e1700872.

[12] YUE Q, LIU W, ZHU X. n-Type molecular photovoltaic materials: design strategies and device applications[J]. Journal of the American Chemical Society, 2020, 142(27):11613-11628.

[13] XIAO Z, JIA X, LI D, et al. 26 mA/cm^2 J_{sc} from organic solar cells with a low-bandgap nonfullerene acceptor[J]. Science Bulletin, 2017, 62(22):1494-1496.

[14] YUAN J, ZHANG Y, ZHOU L, et al. Single-junction organic solar cell with over 15% efficiency using fused-ring acceptor with electron-deficient core[J]. Joule, 2019, 3(4):1140-1151.

[15] LIU Q, JIANG Y, JIN K, et al. 18% Efficiency organic solar cells[J]. Science Bulletin, 2020, 65(4):272-275.

[16] QIN J, ZHANG L, ZUO C, et al. A chlorinated copolymer donor demonstrates a 18.13% power conversion efficiency[J]. Journal of Semiconductors, 2021, 42(1):10501.

[17] JIN K, XIAO Z, DING L. D18, an eximious solar polymer![J]. Journal of Semiconductors, 2021, 42(1):10502.

[18] PEREIRA F, XIAO K, LATINO D A, et al. Machine learning methods to predict density functional theory B3LYP energies of HOMO and LUMO orbitals[J]. Journal of Chemical Information and Modeling, 2017, 57(1):11-21.

[19] LOPEZ S A, SANCHEZ-LENGELING B, DE GOES SOARES J, et al. Design principles and top non-fullerene acceptor candidates for organic photovoltaics[J]. Joule, 2017, 1(4):857-870.

[20] SAHU H , MA H . Unraveling Correlations Between Molecular Properties and Device Parameters of Organic Solar Cells Using Machine Learning[J]. Journal of Physical Chemistry Letters, 2019, 10(22):7277-7284.

[21] ZHAO Z W, Cueto M D, Geng Y, et al. Effect of Increasing the Descriptor Set on Machine Learning Prediction of Small Molecule-Based Organic Solar Cells[J]. Chemistry of Materials, 2020, 32(18):7777-7787.

[22] PENG S P, ZHAO Y. Convolutional neural networks for the design and analysis of non-fullerene acceptors[j]. Journal of Chemical Information and Modeling, 2019, 59(12):4993-5001.

[23] DAVID W. SMILES: A chemical language and information system[J]. Journal of Chemical Information and Modeling, 1988, 28(1):31-36.

[24] RINIKER S, LANDRUM G A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation[J]. Journal of Chemical Information and Modeling, 2015, 55(12):2562-2574.

[25] HALGREN T A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94[J]. Journal of Computational Chemistry, 1996, 17:490-519.

[26] GUPTA A, MULLER A T, HUISMAN B J H, et al. Generative Recurrent networks for De Novo Drug design[J]. Molecular Informatics, 2018, 37:1700111.

[27] BONCHEV D, ROUVRAY D H. Chemical graph theory: Introduction and fundamentals[M]. London: CRC Press, 1991: 1-300.

[28] IKEBATA H, HONGO K, ISOMURA T, et al. Bayesian molecular design with a chemical language model[J]. Journal of Computer-aided Molecular Design, 2017, 31(4):379-391.

[29] GOMEZ-BOMBARELLI R, WEI J N, DUVENAUD D, et al. Automatic chemical design using a data-driven continuous representation of molecules[J]. ACS central science, 2018, 4(2):268-276.

[30] ZHANG M, ZENG M, DENG X, et al. Simultaneously enhancing the Jsc and Voc of ternary organic solar cells by incorporating a Medium-Band-Gap acceptor[J]. ACS Applied Energy Materials, 2021, 4(4):3480-3486.

[31] ASHTON M, BARNARD J, CASSET F, et al. Identification of diverse database subsets using property-based and fragment-based molecular descriptions[J]. Quantitative Structure-Activity Relationships, 2002, 21(6):598-604.

[32] ESPOSITO C, WANG S, LANGE U E W, et al. Combining machine learning and molecular dynamics to predict P-Glycoprotein substrates[J]. Journal of Chemical Information and Modeling, 2020, 60(10):4730-4749.

[33] CERETO-MASSAGUE A, OJEDA, M J, VALLS C, et al. Molecular fingerprint similarity search in virtual screening[J]. Methods, 2015, 71:58-63.

[34] LU T, CHEN F. Multiwfn: a multifunctional wavefunction analyzer[J]. Journal of Computational Chemistry, 2012, 33(5):580-592.

Application of a convolutional neural network model in the generation and property prediction of novel non-fullerene acceptor molecules

YANG Xinyu, PENG Shiping, ZHAO Yi*

(State Key Laboratory of Physical Chemistry of Solid Surfaces, Fujian Provincial Key Laboratory of Theoretical and Computational Chemistry, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, China)

Abstract: In recent years, the efficiency of non-fullerene small molecule acceptors in organic solar cells is getting closer and closer to the goal of industrialization because of its expanded absorption spectrum, the ability to adjust the dissociation energy of the exciton, and the flexible donor-acceptor morphology. In this paper, we used a convolution neural network model proposed by our group recently for molecular generation and property prediction to generate and screen new non-fullerene small molecular acceptors with given frontier orbital energies (Y6 molecule, the highest occupied molecular orbital (HOMO) target energy of -5.73 eV and the lowest unoccupied molecular orbital (LUMO) target energy of -3.69 eV) for efficient exciton dissociation. After the generation model was fully trained in the database and fine-tuned with a small data set, more than two hundred molecules close to the target orbital energy were generated, and then the prediction model was used to further screen and predict the contribution of molecular fragments to the orbital

energy. After that, these molecules and molecules with similar frontal orbital energies in the database were clustered together and ten new acceptor molecules with different chemical spaces were selected. Finally, the accuracy of the prediction of the contribution of orbital energy and molecular fragments to the orbits was verified by ab initio calculations, and the oscillator intensity of the molecular light absorption was given. We also further used the generation and prediction model to provide ten non-fullerene small molecules with another set of orbital energies (HOMO energy of -5.10 eV and LUMO energy of -3.10 eV). The predicted properties were consistent with the ab initio results, which proves the robustness of the generation and prediction model and the reliability of the results. The molecules predicted in this paper also provide the design scheme of potential molecular frameworks with high-performance non-fullerene acceptors.

Keywords: convolutional neural network; non-fullerene acceptor; frontier molecular orbital energy