研究论文

DOI：10.11949/0438-1157.20201880

基于深度学习预测有机光伏电池能量转换效率

于程远，吴金奎，周利，吉旭，戴一阳，党亚固

（四川大学化学工程学院，四川 成都 610065）

摘要： 采用一种针对有机化合物提出的类语言分子描述符对哈佛清洁能源项目数据库（CEPDB）中29000个有机太阳能电池供体分子进行描述，分子将基于最近邻子图理论被分解成片段（词），并利用广度优先搜索算法将片段排列成一定的序列（句子），在每个片段的信息被嵌入一个数值向量后，每个分子可表示为一个信息矩阵。在此基础上，通过一个深层神经网络提取嵌入信息，并与对应材料的光电转换效率（PCE）关联，获得了决定系数（R^2）为0.97、均方误差（MSE）为0.16的预测结果。与现有方法的比较表明该方法在精度上具有竞争力。在建模过程中引入注意力机制，识别出了几个对PCE值具有决定性意义的分子片段，可为有机光伏材料的逆向设计提供指导信息。

关键词： 有机化合物；太阳能；类语言描述符；深度学习；预测；光电转换效率

中图分类号： TM 914.4　　**文献标志码：** A

文章编号： 0438-1157（2021）03-1487-09

Prediction of energy conversion efficiency of organic solar cells based on deep learning

YU Chengyuan, WU Jinkui, ZHOU Li, JI Xu, DAI Yiyang, DANG Yagu

(*School of Chemical Engineering*, *Sichuan University*, *Chengdu* 610065, *Sichuan*, *China*)

Abstract: A language-like descriptor for organic compounds was used to describe 29000 organic solar cell donor molecules collected from the Harvard Clean Energy Project Database (CEPDB). Inspired by the similarity between organic chemistry and natural language, these molecules were decomposed into fragments (words) based on the nearest neighbor subgraph theory, and these fragments were arranged into a certain sequence (sentences) by the breadth first search algorithm. After the information of each fragment was embedded into a numerical vector, each molecule can be represented by an information matrix. This matrix is a descriptor called g-FSI, which can reflect the composition and structure information of molecules. The descriptor was then parsed by a deep neural network to extract the embedded information and correlate to the corresponding PCE. The prediction model has obtained the prediction result in which the determination coefficient (R^2) is 0.97 and the mean square error (MSE) is 0.16. Compared with the existing research, this model is competitive in accuracy of prediction. The attention mechanism is introduced in the modeling process, and several molecular fragments that are decisive for the PCE value are identified, which can provide guidance information for the reverse design of organic photovoltaic materials.

Key words: organic compounds; solar energy; language-like descriptor; deep learning; prediction; power

收稿日期： 2020-12-20　　**修回日期：** 2020-12-27

通信作者： 周利(1987—)，女，博士，副教授，chezli@scu.edu.cn

第一作者： 于程远(1996—)，男，硕士研究生，843766990@qq.com

基金项目： 中央高校基本科研业务费专项资金(YJ201838)；国家自然科学基金项目(21776183，21706220)

引用本文： 于程远, 吴金奎, 周利, 吉旭, 戴一阳, 党亚固. 基于深度学习预测有机光伏电池能量转换效率[J]. 化工学报, 2021, 72(3): 1487-1495

Citation: YU Chengyuan, WU Jinkui, ZHOU Li, JI Xu, DAI Yiyang, DANG Yagu. Prediction of energy conversion efficiency of organic solar cells based on deep learning[J]. CIESC Journal, 2021, 72(3): 1487-1495

conversion efficiency

引 言

基于光伏技术的太阳能捕集是一种能够解决日益增长的全球能源需求的可持续手段。新型高效光伏材料的发现在世界范围内已成为学术界和工业界的热门话题[1-6]。其中，有机光伏（OPV）因其低成本、轻量化、机械灵活性和大面积制造潜力而备受关注[7-9]。尽管OPV有着许多的优点，但其发展仍然具有挑战性，并且很大程度上依赖于光电转换效率（PCE）的提高[10]。目前，新型OPV的开发主要采用实验驱动的试错法，在资源和时间上成本高且在探索新化学空间上有效性有限。

实验试错方法存在的这些缺点促使研究者们通过建立模型来指导OPV的开发。Scharber模型[11]从受体的最低未占据分子轨道（LUMO）和给体的最高占据分子轨道（HOMO）的能级来估计本体异质结太阳能电池的最大PCE，广泛用于光伏材料的性质预估。尽管这类模型通常过于简单化，无法解释有机太阳能电池的所有复杂物理化学行为，但其可对候选化合物可能达到的潜在最佳性能进行预判，为实验设计提供有价值的参考。最近，Green等[12]在半经验模型的分析指导下，实验合成的有机光伏材料PCE实现了新的突破（17.35%）。大量的研究工作也证明，正确反映化合物定量结构与性质关系的模型（QSPR模型）可为材料性能的改善提供有力支撑[13-16]。近年来，随着计算能力和机器学习算法的快速发展，高通量虚拟筛选（HTVS）正成为发现新型高性能材料的主流方法：包括建立化合物的理论/实验性能数据库，开发正向性能预测模型和逆向材料设计规则和算法。

针对有机光伏材料研发，哈佛大学建立了三个权威的开源数据库，包括哈佛清洁能源项目数据库（CEPDB）[17]、哈佛有机光伏数据集（HOPV15）[18]和非富勒烯小分子受体数据库（NFADB）[19]。CEPDB包含230万种有机化合物的HOMO/LUMO能量及其基于Scharber模型[11]计算出的PCE值。NFADB是包含50000个非富勒烯受体材料的集合，这些材料的PCE值为校准后的HOMO/LUMO能量计算值[19]。HOPV15为归纳总结已有文献报道的光伏材料实验数据和一定量子化学计算补充得到的小数据集[18]。尽管数据库中的计算数据基于一定假设，但也能一定程度上反映材料微观结构变化对其宏观性质的影响规律，指导实验设计原则的制定。Hachmann等[20]通过分析来自CEPDB的数据，确定了关键的分子构建块，并为潜在的高PCE值有机化合物候选材料提取了设计规则。

前向性能预测模型的开发通常由编码和映射两部分组成。编码过程将分子的组成、结构等信息转换为一系列称为描述符的数值，而映射过程通过确定合适的函数来映射描述符和需要预测的分子属性。Padula等[21]提出将电子特征和结构特征结合起来作为有机太阳能电池性能预测的描述符，基于此，机器学习模型的预测能力达到了r=0.7的水平。Sahu等[22]用量子化学计算得到的13个微观性质作为有机小分子PCE预测的描述符，建立了pearson系数为0.79的梯度提升树模型。随后，同一研究小组通过引入基态几何结构、阳离子和阴离子等新的微观属性作为描述符，进一步推进了研究，得到的模型pearson系数为0.78[23]。此外，Sun等[24]应用卷积技术从分子结构图中提取描述符，用于有机化合物PCE性能的估计，得到精度为91.2%的卷积神经网络模型。随着机器学习的发展，一些最新的算法如迁移学习也被用来提高材料性质预测值与实验值之间的一致性[25]。

尽管有机光伏材料领域的QSPR模型研究已取得较多成果，其未来发展仍面临挑战。首先，量子化学计算生成的描述符可以提供准确的结果，但这通常要求建模者具备深入的领域知识，并且计算成本高昂，限制了它在HTVS中的有效性。二是化合物的微观物理化学环境非常复杂，大多数易于获取的描述符容易遗漏重要的化学信息，导致预测结果不太理想；且许多机器学习模型都是“黑匣子”，其结果往往可解释性较低。本研究的开展动机正是在于此，力求在解决上述挑战上有所贡献。

受Cadeddu等[26]针对有机化合物和自然语言（英语）开展的相似性研究的启发，该工作在分子片段和文本片段的出现频率上论证了有机化学和自然语言（英语）之间的高度相似性，本文采用本课题组提出一种类语言的分子特征提取和表征策略，在此基础上建立预测有机化合物PCE值的深度学习模型，力求为高性能OPV的虚拟筛选提供支撑。首

先，将有机化合物的分子图分解为片段，并根据其相对位置和连接性对每个片段进行编号。通过将所涉及的分子片段标识为唯一的片段向量，有机化合物就可被表示为一个内嵌分子片段序列信息的类语言描述符。其次，构建自然语言处理算法来“理解”描述符，将分子信息与其潜在的PCE性能相关联。最后，使用已在自然语言处理领域成功获得广泛应用的神经网络解释器——注意力机制，来识别对有机化合物PCE性能有重要贡献的关键分子片段，提高模型的可解释性，并为具有更高光电转换效率的OPV材料设计提供支撑。

1 研究方法

本小节介绍研究组提出的结合Bi-LSTM网络、注意力机制和反向传播神经网络(BPNN)的深度神经网络(DNN)模型，用于构建有机化合物的分子结构与PCE值之间的映射关系。基于所提出方法建立QSPR模型主要包括以下四个步骤[27]。

(1)数据采集和预处理。搜集有机化合物的SMILES字符串和实验测量(或量子化学计算)PCE值，并进行数据预处理。

(2)分子预编码。通过分子SMILES字符串生成分子图，基于其构造片段的连接性和每个片段在预定义片段池中的位置，生成分子片段序列信息，如图1所示。

(3)编码-预测神经网络训练。基于分子片段描述符矩阵，将分子片段序列信息嵌入分子描述符；基于分子描述符和相应PCE值训练DNN。在训练过程中，不断优化描述符和DNN模型参数，提高模型性能，如图2所示。

(4)模型评估。利用测试数据集对所建立的QSPR模型的预测性能进行评价。

1.1 数据收集

从CEPDB收集29000个OPV供体分子的SMILES字符串及其理论PCE值[17]。对数据进行预处理，去掉存在异常、缺失以及不符合实际值的数据项。基于式(1)所示的Z评分标准化对PCE值进行预处理，以加快模型训练过程的收敛速度，提高模型精度。

$$y_i' = \frac{y_i - \bar{y}}{s(y)} \tag{1}$$

其中，y_i是分子的PCE值，$\bar{y}$和$s(y)$分别是所收集的OPV数据集中所有PCE值的平均值和标准差。所搜集数据项在预处理前后的数据分布如图3所示。

1.2 有机化合物预编码

通过RDKit和Networkx将SMILES字符串转换成分子图[28-29]，分子图是分子到平面的投影，其中顶点代表原子，边代表化学键。如图1所示，为了更好地编码分子片段的连接性信息，采用最近邻子图[30]和广度优先搜索(BFS)算法[31]将分子(图)分解成片段(子图)，并将分子按相应的BFS顺序进行排序编码，便得到了相应分子的构成片段序列信息。需要指出的是，一般情况下，不同的有机化合物可以分解为不同数目的组成分子片段。为了确保分子的片段序列具有相同的维数，需进行最大分子，即组成片段数最多的分子的识别，并将其组成片段的计数设为片段序列的维数。在生成其他分子的片段序列信息时，以零值填充多出部分，以确保维度的一致性。

其中，每一分子片段基于最近邻子图法由最近邻顶点和距离当前顶点一跳内的边切割获得，换句话说，每个分子片段所反映的是顶点原子与其最近一个化学键所构成化学环境的总和。由于OPV分子中只有少量的原子和键，因此，该方法比一般的

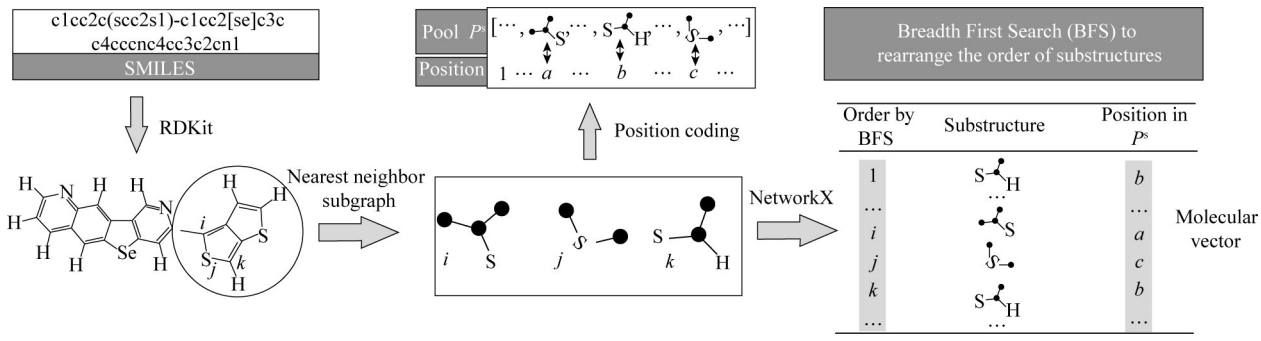


图1 基于分子图的嵌入[27]

Fig.1 Embedding based on a given molecular graph[27]

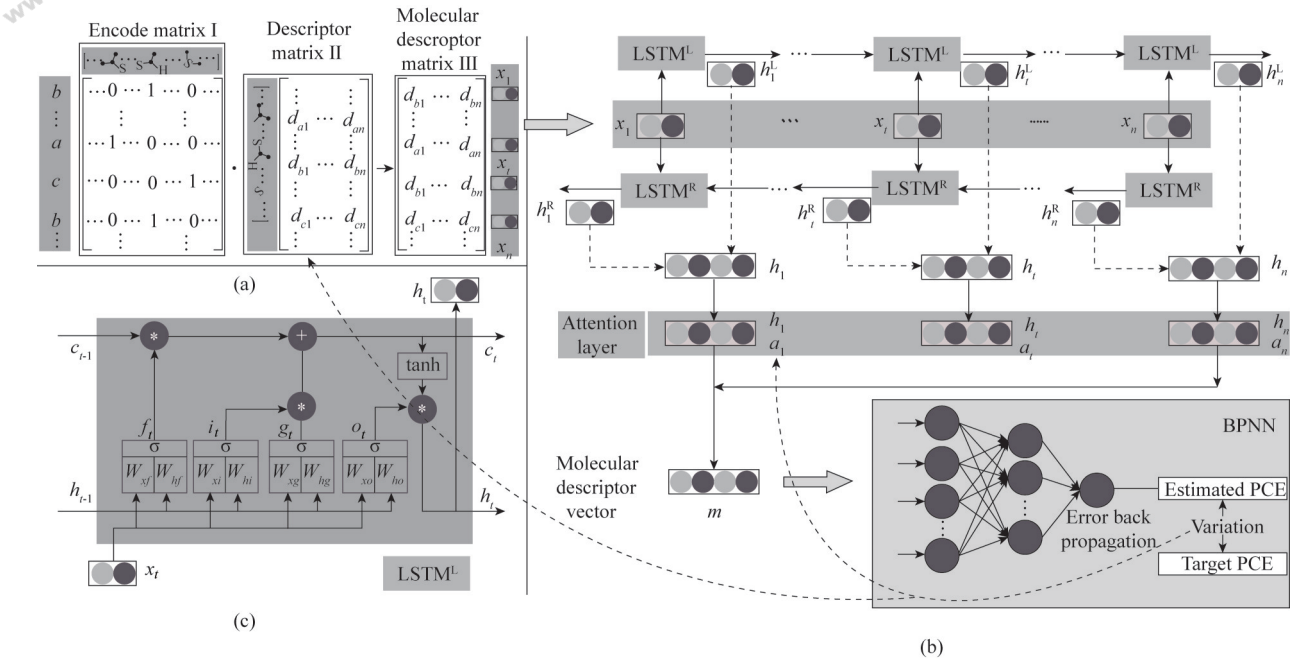


图2 预测网络结构示意图[27]

Fig.2 Structure of predictor network

(a) 嵌入过程,每一个嵌入矩阵代表一个分子,m是嵌入向量的维度,本案例m=128;(b) LSTM单元的结构图;(c) 拥有注意力机制的Bi-LSTM网络

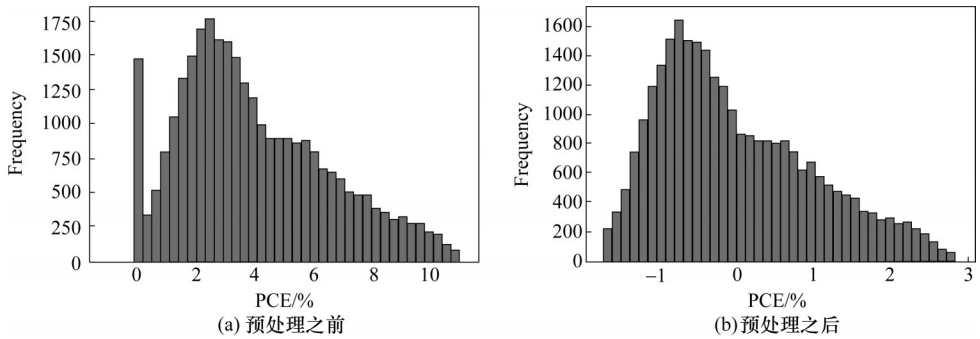


图3 预处理前后的数据分布

Fig.3 Data distribution before and after data preprocessing

化学信息嵌入法更高效。此外,BFS遍历算法考虑了分子结构片段的连通性信息,可提高后续构建的QSPR模型的精度。预编码之后,每个化合物就被表示为一个唯一的序列。

1.3 类语言描述符的生成和预测模型训练

如图2(a)所示,将所得到的分子片段序列信息转化为一个one-hot矩阵(矩阵Ⅰ),再基于分子片段池中所有分子片段的嵌入向量(即分子片段描述符)构成片段嵌入矩阵(矩阵Ⅱ),矩阵Ⅰ和矩阵Ⅱ的乘积得到相应化合物的分子描述符(矩阵Ⅲ),简称为g-FSI[27]。接着,将分子描述符传递到Bi-LSTM网络。为了更有效地提取片段序列信息,描述符的每一行(片段向量)同时由一个正向LSTM单元和一个反向LSTM单元处理,处理后的片段信息分别表示为$\overrightarrow{\boldsymbol{h}_t}$和$\overleftarrow{\boldsymbol{h}_t}$。处理后的信息继续进入后续的正向和反向LSTM单元进行信息提取。$\overrightarrow{\boldsymbol{h}_t}$和$\overleftarrow{\boldsymbol{h}_t}$的信息组合继续被输入到深度神经网络的注意力层[32],在注意力层上,引入标准化权重$\boldsymbol{\alpha}_t$表示每一分子片段的重要性,以提高模型的性能和可解释性。在注意力层之后,将经过Bi-LSTM和注意力机制层处理后的信息(记为$\boldsymbol{M}$)送入BPNN网络,进行PCE值预测。在训练过程中,以均方误差(MSE)作为损失函数评估回归模型性能,并将模型误差向后传播以更新片段嵌入矩阵以及Bi-LSTM和注意力层中的参数。选择随机梯度下降算法优化损失函数,使均方误差最小。下面两小节内容将对Bi-LSTM网络和注意力机制进行更详细的阐述。

1.3.1 Bi-LSTM网络用于分子片段尺度的特征提取和信息集成 Bi-LSTM网络是一种具有处理长序列能力的增强型递归神经网络(RNN),能同时考虑分子片段序列信息中嵌入的前向和后向上下文信息[33-34],被广泛应用于序列数据处理中,如,无约束手写体识别、机器翻译、图像字幕等[35-37]。对于一个给定的分子p,对应的描述符为$[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_t,\cdots,\boldsymbol{x}_n]$(这里$n$表示分子片段序列信息的维度,$\boldsymbol{x}_t$为分子片段向量),如图2(b)所示,当前分子的每个分子片段向量都将作为一个前向和一个后向LSTM单元的输入,处理后的片段信息——$\overrightarrow{\boldsymbol{h}_t}/\overleftarrow{\boldsymbol{h}_t}$被传递到下一个LSTM单元。对于每个LSTM单元,引入自适应机制来决定前一个单元传递的前一个片段信息的保存程度,并存储当前片段信息输入的特征[34]。

如图2(c)所示为一个正向LSTM单元,单元中引入三个门控单元来控制信息流:遗忘门$\boldsymbol{f}_t$,输入门$\boldsymbol{i}_t$和输出门$\boldsymbol{o}_t$,并通过$\boldsymbol{g}_t$控制该单元的状态$\boldsymbol{c}_t$和隐藏单元的状态$\boldsymbol{h}_t$。门控单元和单元状态的信息处理过程服从方程式(2)~式(6),通过一系列的权重矩阵$\boldsymbol{W}_{xi},\boldsymbol{W}_{hi},\boldsymbol{W}_{xf},\boldsymbol{W}_{hf},\boldsymbol{W}_{xc},\boldsymbol{W}_{hc},\boldsymbol{W}_{xo},\boldsymbol{W}_{ho}$和偏置参数$\boldsymbol{b}_i,\boldsymbol{b}_f,\boldsymbol{b}_c,\boldsymbol{b}_o$来处理当前单元的信息输入$\boldsymbol{x}_t$以及前一单元生成的状态信息$\overrightarrow{\boldsymbol{h}_{t-1}}$。门控单元处理之后的信息用于更新当前单元状态$\boldsymbol{c}_t$及其隐藏状态输出$\overrightarrow{\boldsymbol{h}_t}$[38],计算如式(7)所示。

$$\boldsymbol{i}_t=\sigma\left(\boldsymbol{W}_{xi}\boldsymbol{x}_t+\boldsymbol{W}_{hi}\overrightarrow{\boldsymbol{h}_{t-1}}+\boldsymbol{b}_i\right) \tag{2}$$

$$\boldsymbol{f}_t=\sigma\left(\boldsymbol{W}_{xf}\boldsymbol{x}_t+\boldsymbol{W}_{hf}\overrightarrow{\boldsymbol{h}_{t-1}}+\boldsymbol{b}_f\right) \tag{3}$$

$$\boldsymbol{g}_t=\tanh\left(\boldsymbol{W}_{xc}\boldsymbol{x}_t+\boldsymbol{W}_{hc}\overrightarrow{\boldsymbol{h}_{t-1}}+\boldsymbol{b}_c\right) \tag{4}$$

$$\boldsymbol{c}_t=\boldsymbol{i}_t\boldsymbol{g}_t+\boldsymbol{f}_t\boldsymbol{c}_{t-1} \tag{5}$$

$$\boldsymbol{o}_t=\sigma\left(\boldsymbol{W}_{xo}\boldsymbol{x}_t+\boldsymbol{W}_{ho}\overrightarrow{\boldsymbol{h}_{t-1}}+\boldsymbol{b}_o\right) \tag{6}$$

$$\overrightarrow{\boldsymbol{h}_t}=\boldsymbol{o}_t\tanh\left(\boldsymbol{c}_t\right) \tag{7}$$

本文采用的Bi-LSTM网络包含n个正向LSTM单元和n个反向LSTM单元。经信息处理后,得到一组隐藏状态,用于前向和后向信息提取。Bi-LSTM网络的信息提取如式(8)和式(9)所示。

$$\overrightarrow{\mathrm{LSTM}}\left([\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_t,\cdots,\boldsymbol{x}_n]\right)=[\overrightarrow{\boldsymbol{h}_1},\overrightarrow{\boldsymbol{h}_2},\cdots,\overrightarrow{\boldsymbol{h}_t},\cdots,\overrightarrow{\boldsymbol{h}_n}] \tag{8}$$

$$\overleftarrow{\mathrm{LSTM}}\left([\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_t,\cdots,\boldsymbol{x}_n]\right)=[\overleftarrow{\boldsymbol{h}_1},\overleftarrow{\boldsymbol{h}_2}\cdots,\overleftarrow{\boldsymbol{h}_t},\cdots,\overleftarrow{\boldsymbol{h}_n}] \tag{9}$$

将由反向单元获得的隐藏状态$[\overleftarrow{\boldsymbol{h}_1},\overleftarrow{\boldsymbol{h}_2}\cdots,\overleftarrow{\boldsymbol{h}_t},\cdots,\overleftarrow{\boldsymbol{h}_n}]$和由正向单元获得的隐藏状态$[\overrightarrow{\boldsymbol{h}_1},\overrightarrow{\boldsymbol{h}_2},\cdots,\overrightarrow{\boldsymbol{h}_t},\cdots,\overrightarrow{\boldsymbol{h}_n}]$拼接形成新信息向量$\boldsymbol{h}_t$,作为DNN后续注意力层的输入。

$$\boldsymbol{h}_t=\left[\overrightarrow{\boldsymbol{h}_t}\oplus\overleftarrow{\boldsymbol{h}_t}\right] \tag{10}$$

1.3.2 注意力机制用于分子尺度上的特征提取和信息集成 从微观化学环境角度,并非每一分子片段对有机化合物的PCE性能都具有相同的贡献。因此,采用注意力机制来跟踪对PCE性能有重要影响的分子片段。在分子片段尺度提取的特征通过与标准化的重要性权重向量相乘,合并为分子尺度的特征向量。经信息处理后的分子片段信息$\boldsymbol{h}_t$输入一个单层MLP(多层感知器)得到$\boldsymbol{u}_s$,其中,引入了权重向量$\boldsymbol{W}_s$和偏置参数$\boldsymbol{b}_s$,进一步通过softmax函数计算得到标准化的重要性权重$\boldsymbol{\alpha}_t$。然后,通过计算信息向量的加权和得到处理后的分子信息$\boldsymbol{M}$。计算公式如下[39]:

$$\boldsymbol{u}_t=\tanh\left(\boldsymbol{W}_s\boldsymbol{h}_t+\boldsymbol{b}_s\right) \tag{11}$$

$$\boldsymbol{\alpha}_t=\frac{\exp\left(\boldsymbol{u}_t^{\mathrm{T}}\boldsymbol{u}_s\right)}{\sum\exp\left(\boldsymbol{u}_t^{\mathrm{T}}\boldsymbol{u}_s\right)} \tag{12}$$

$$\boldsymbol{M}=\sum_t\boldsymbol{\alpha}_t\boldsymbol{h}_t \tag{13}$$

其中,$\boldsymbol{u}_t^{\mathrm{T}}$是在网络训练过程中随机初始化,在模型训练过程中将基于分子片段信息向量不断学习优化。

1.4 模型验证

基于测试数据集,对所建立的QSPR模型的预测性能、竞争力进行评估,并与其他预测模型进行比较,评价所得模型外部竞争力。

以上所有的模型训练和评估步骤都是通过Python语言编写完成,并在Windows和Linux平台上部署。同时,神经网络算法的实现基于开源的深度学习框架Pytorch[40],并基于2个GTX-1080Ti GPU实现模型训练。

2 结果与讨论

2.1 实验超参数设置

将收集到的CEPDB数据集随机划分为训练集、验证集和测试集,其中,验证集用于模型训练过程中模型超参数优化的验证,测试集用于最终的模型评估。

采用网格搜索优化模型超参数,包括损失函数优化器的选择、学习速率、隐层和隐层单元的个数。选择Adam[41]作为损失函数优化器,学习率0.001。经过模型训练和验证过程,得到QSPR模型。最终优化模型的BP神经网络包含3层,每层32个隐藏单

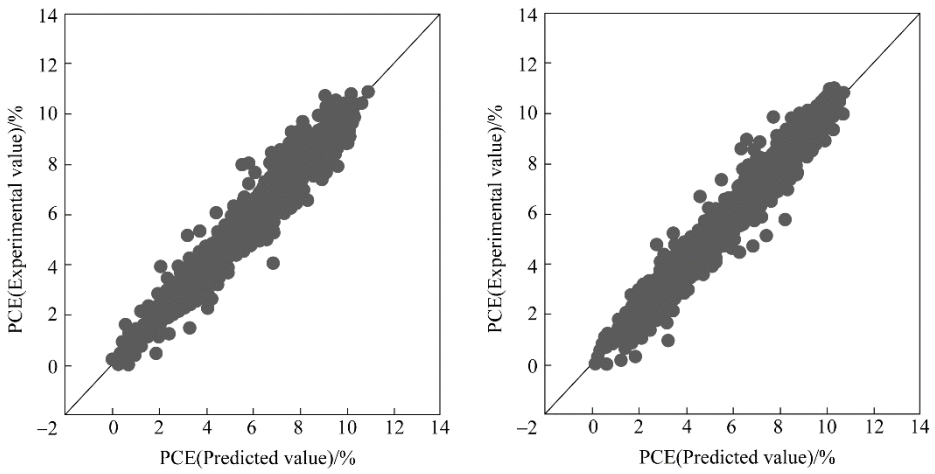


图4 真实值和QSPR模型预测值的散点图

Fig.4 Scatter plots for the predicted – experimental value with the QSPR model

元。利用测试集对得到的QSPR模型进行评价，模型预测值与计算理论值之间的比较如图4所示。对于验证和测试数据集，可以看到，预测结果沿对角线分布紧密。计算得到模型的决定系数(R^2)为0.97，验证集和测试集的预测均方误差(MSE)分别为0.17和0.16。可以得出所得QSPR模型具有较高精度的结论。

2.2 模型的竞争性

进一步验证所提出方法的合理性和可靠性，基于同一数据集，应用其他几种分子描述符和机器学习算法建立QSPR模型。所选择的分子描述符包括扩展连通型指纹ECFP [42]和Mol2vec[43]；选择的机器学习算法为ANN和RF。对于ECFP，设定位向量长度为2048，同时，将Mol2vec的嵌入维度固定为300，基于此，ECFP和Mol2vec将分子描述为固定长度的向量，生成的描述符可以直接用作ANN和RF的输入。共得到5个QSPR模型，其性能比较列于表1。

表1 测试集的预测精度

Table 1 Prediction accuracy of the testing set

Measure	MSE	R^2
Mol2vce+ RF	0.98	0.83
ECFP + RF	0.65	0.89
Mol2vec + ANN	1.07	0.81
ECFP + ANN	0.42	0.92
g-FSI+ Bi-LSTM + attention	0.16	0.97

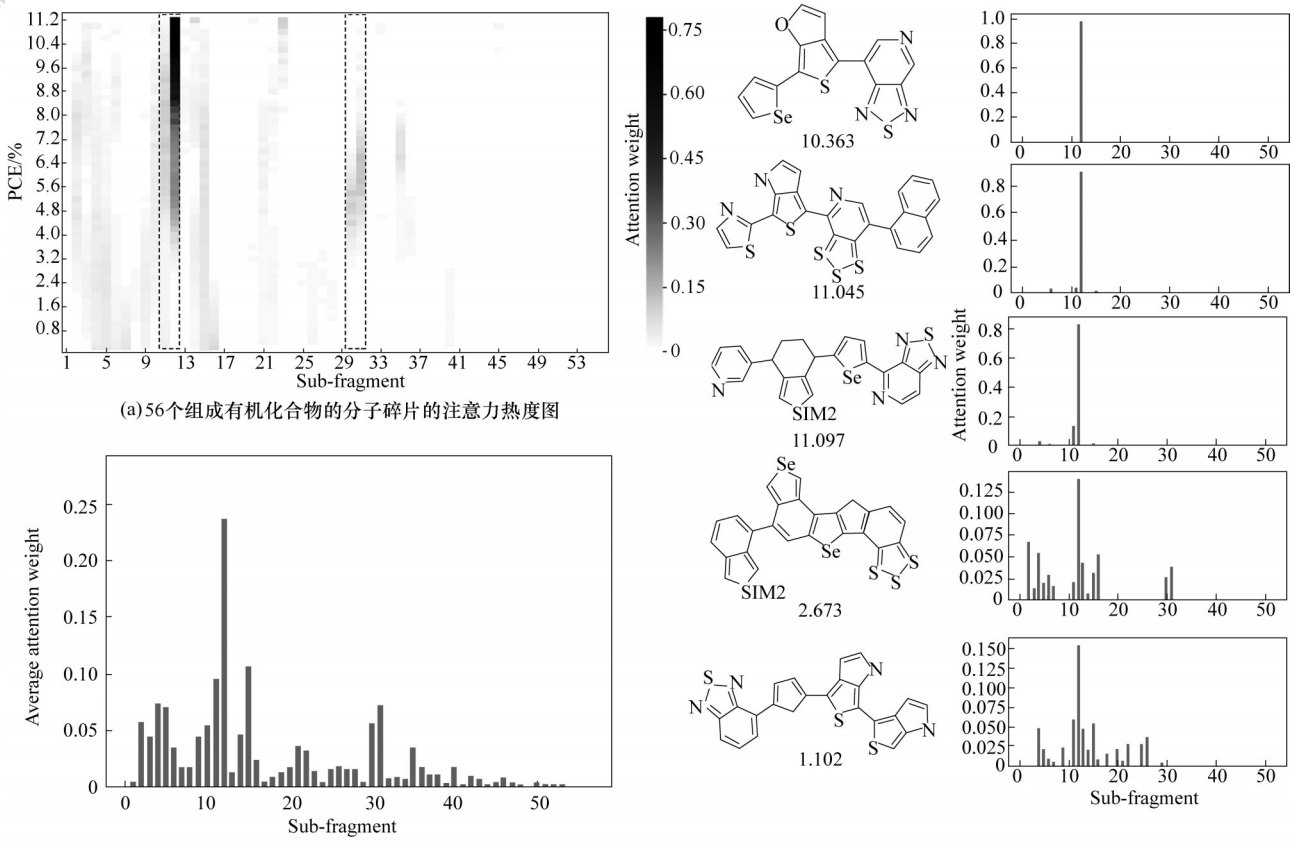
从三个分子描述符g-FSI、ECFP和Mol2vec的比较可以看出，基于g-FSI和ECFP所得到预测结果的决定系数高于以Mol2vec作为描述符时的结果，均大于0.9，且MSE较低。说明g-FSI和ECFP能够更好地满足当前研究的需要。本质上，g-FSI和ECFP是基于分子片段信息的相同类型的分子描述符，所以均取得了较好的表现；相比于ECFP，g-FSI同时考虑了分子的片段信息和序列信息，这也是使得g-FSI预测效果更好的重要原因。

同样是受到自然语言处理技术启发而产生的Mol2vec却在预测任务中取得了不理想的结果。Mol2vec其本身是利用大量有机分子作为语料库通过Word2vec预训练得到分子片段嵌入向量的一种无监督方法，其特点在于学习到的嵌入向量是稠密的。但从分子片段向整个分子过渡的过程中，采用了直接加和平均的方法，该过程势必带来分子整体信息的损失，尤其是分子的序列信息被彻底忽略，这些因素的共同作用使Mol2vec表现不佳。

2.3 基于注意力机制对重要分子片段的分析

对于材料设计，QSPR模型预测结果的可解释性不亚于其预测精度[44]。与专家的经验直觉或经验类似，模型“学习”过程中获得的信息对具有更佳性能的OPV材料设计具有指导意义。本节通过“学习”过程，根据注意力机制赋予每个分子片段的注意权重，分析有利于有机化合物潜在PCE性能的重要分子片段。

对于分子片段，其对于有机化合物光电转换性能越重要，在模型训练中获得的注意力权重越大。图5(a)给出了基于训练数据集获得的56个片段在具有不同PCE值的有机化合物中的注意力权重的热力图。颜色越深，注意力权重越高。可以看到，在大多数有机化合物中，有两个分子片段群“备受



(a) 56个组成有机化合物的分子碎片的注意力热度图

(b) 不同分子碎片的平均注意力权重图

(c) 5个有机太阳能电池给体材料分子的注意力权重图

图5 注意力机制的可视化

Fig.5 Visualization of the attention mechanism

关注”,即分子片段11~12和30~31。如图5(b)给出了56个分子片段的平均注意力权重,这些分子片段的平均权重值也高于其他分子片段。

为便于进一步分析,从测试集中提取5种含有片段12的化合物,如图5(c)所示。通过比较这五种化合物的分子片段组成,可以得到十分有趣的结果。对于前三个PCE值大于10%的分子,其性能主要受片段12的影响。除片段12外,片段4~6以及片段11在高PCE的分子中也有着不同程度的作用。而在PCE值介于3%~7%范围内,分子性能还将受到片段30~31的强烈影响;基于此,可以大胆地推断片段12以及片段30~31之间的协同作用将会导致分子的PCE值趋向于平庸化,在分子设计中需要避免同时引入片段12与片段30~31。而对于其余两种PCE值低于3%的化合物,不难发现,片段12依旧占据主导作用,但是其权重系数均小于0.15,远低于在高PCE分子中的权重(大于0.8)。此外,片段群4~7和13~16也具有相对活跃的表现,但更明显的是PCE较低的两组在注意力权重分布上比其他三组更为平均。

3 结 论

光伏技术被认为是解决21世纪能源短缺和环境危机的最有前途的途径之一。发现具有高光电转换效率的化合物已成为推动该技术发展的关键任务之一。受有机化学与自然语言的相似性启发,本文采用一种类语言的分子描述符描述有机化合物,建立深度学习模型,以实现高精度的PCE值预测。在分子描述过程中,将由原子和键组成的分子片段信息嵌入到数值向量中,并根据分子片段的序列信息将相关向量聚合成矩阵。研究已表明,片段(词)的位置信息对分子(句)的性质预测(意义理解)具有重要意义,故采用Bi-LSTM对分子描述符进行处理,使嵌入的分子片段序列信息能够被完全“理解”。然后,将处理后的信息传递给BPNN,实现PCE值的预测。在此过程中,应用注意力机制帮助识别分子片段的重要性,提高预测精度。模型评价结果表明,与其他几种分子描述符和机器学习算法

相比，该模型具有更高的预测精度和竞争性。此外，所建立的方法能在一定程度上揭示分子片段对分子PCE性能的影响，可以为OPV的逆向设计提供依据。

本研究中的描述符生成和性质映射过程都是自动完成的，避免了人为干预。换句话说，深度学习方法能够从SMILES中提取和学习重要的知识，因此不需要建模者提供深入的领域知识。此外，在所用方法的“学习”过程中，能够识别出具有决定性作用的片段，表明所采用的方法能够为OPV的逆向设计提供有指导意义的信息。虽然本研究的重点是OPV的PCE值预测，但是该方法可以进一步扩展到有机材料的其他重要性质的预测。

参考文献

[1] Leijtens T, Eperon G E, Barker A J, et al. Carrier trapping and recombination: the role of defect physics in enhancing the open circuit voltage of metal halide perovskite solar cells[J]. Energy & Environmental Science, 2016, **9**(11): 3472-3481.

[2] Zheng B, Wang F, Dong S, et al. Supramolecular polymers constructed by crown ether-based molecular recognition[J]. Chemical Society Reviews, 2012, **41**(5): 1621-1636.

[3] Jošt M, Kegelmann L, Korte L, et al. Monolithic Perovskite Tandem solar cells: a review of the present status and advanced characterization methods toward 30% efficiency[J]. Advanced Energy Materials, 2020, **10**(26): 1904102.

[4] Kaltenbrunner M, White M S, Glowacki E D, et al. Ultrathin and lightweight organic solar cells with high flexibility[J]. Nature Communications, 2012, **3**(1): 770.

[5] Fukuda K, Yu K, Someya T. The future of flexible organic solar cells[J]. Advanced Energy Materials, 2020, **10**(25): 2000765.

[6] Meng L X, Zhang Y M, Wan X J, et al. Organic and solution-processed tandem solar cells with 17.3% efficiency[J]. Science, 2018, **361**(6407): 1094-1098.

[7] Jinno H, Fukuda K, Xu X M. et al. Stretchable and waterproof elastomer-coated organic photovoltaics for washable electronic textile applications[J]. Nature Energy, 2017, **2**(10): 780-785.

[8] Hedley G J, Ruseckas A, Samuel I D W. Light harvesting for organic photovoltaics[J]. Chemical Reviews, 2017, **117**(2): 796-837.

[9] Liu C, Wang K, Gong X, et al. Low bandgap semiconducting polymers for polymeric photovoltaics[J]. Chemical Society Reviews, 2016, **45**(17): 4825-4846.

[10] Chen C, Zuo Y, Ye W, et al. A critical review of machine learning of energy materials[J]. Advanced Energy Materials, 2020, **10**(8): 1903242.

[11] Scharber M C, Mühlbacher D, Koppe M, et al. Design rules for donors in bulk-heterojunction solar cells—towards 10% energy-conversion efficiency [J]. Advanced Materials, 2006, **18**(6): 789-794.

[12] Green M A, Dunlop E D, Hohl-Ebinger J, et al. Solar cell efficiency tables (Version 55) [J]. Progress in Photovoltaics: Research and Applications, 2020, **28**(1): 3-15.

[13] Gómez-Bombarelli R, Aguilera-Iparraguirre J, Hirzel T D, et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach[J]. Nature Materials, 2016, **15**(10): 1120-1127.

[14] Kim E, Huang K, Jegelka S, et al. Virtual screening of inorganic materials synthesis parameters with deep learning[J]. npj Computational Materials, 2017, **3**(1): 53.

[15] Wu S, Kondo Y, Kakimoto M, et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm[J]. npj Computational Materials, 2019, **5**(1): 66.

[16] Yamada H, Liu C, Wu S, et al. Predicting materials properties with little data using shotgun transfer learning[J]. ACS Central Science, 2019, **5**(10): 1717-1730.

[17] Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, et al. The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid[J]. The Journal of Physical Chemistry Letters, 2011, **2**(17): 2241-2251.

[18] Lopez S A, Pyzer-Knapp E O, Simm G N, et al. The Harvard organic photovoltaic dataset[J]. Scientific Data, 2016, **3**(1): 160086.

[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] Hachmann J, Olivares-Amaya R, Jinich A, et al. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry – the Harvard clean energy project[J]. Energy & Environmental Science, 2014, **7**(2): 698-704.

[21] 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.

[22] Sahu H, Rao W, Troisi A, et al. Toward predicting efficiency of organic solar cells *via* machine learning and improved descriptors [J]. Advanced Energy Materials, 2018, **8**(24): 1801032.1-1801032.9.

[23] Sahu H, Yang F, Ye X B, et al. Designing promising molecules for organic solar cells *via* machine learning assisted virtual screening [J]. Journal of Materials Chemistry A, 2019, **7**(29): 17480-17488.

[24] Sun W B, Li M, Li Y, et al. Material evaluation: the use of deep learning to fast evaluate organic photovoltaic materials[J]. Advanced Theory & Simulations, 2019, **2**(1): 1970001.

[25] Paul A, Jha D, Al-Bahrani R, et al. Transfer learning using ensemble neural networks for organic solar cell screening[C]// 2019 International Joint Conference on Neural Networks. Budapest, Hungary, 2019.

[26] Cadeddu A, Wylie E K, Jurczak J, et al. Organic chemistry as a language and the implications of chemical linguistics for structural and retrosynthetic analyses[J]. Angewandte Chemie International Edition, 2014, **53**(31): 8108-8112.

[27] Wu J K, Wang S H, Zhou L, et al. Deep-learning architecture in QSPR modeling for the prediction of energy conversion efficiency of solar cells[J]. Industrial & Engineering Chemistry Research, 2020, **59** (42): 18991-19000.

[28] RDKit: Open-Source Cheminformatics Software[CP/OL]. [2020-10-26]. http: //www.rdkit.org.

[29] Hagberg A A, Schult D A, Swart P J. Exploring network structure, dynamics, and function using networkx[C]//Varoquaux G, Vaught T, Millman J. Proceedings of the 7th Python in Science Conference. Pasadena, CA, USA, 2008: 11-15.

[30] Costa F, de Grave K D. Fast neighborhood subgraph pairwise distance kernel[C]//Fürnkranz J, Joachims T. Proceedings of the 27th International Conference on International Conference on Machine Learning. Madison, WI, USA: Omnipress, 2010: 255-262.

[31] Zhao Y, Hayashida M, Jindalertudomdee J, et al. Breadth-first search approach to enumeration of tree-like chemical compounds [J]. Journal of Bioinformatics & Computational Biology, 2013, **11** (6): 1343007.

[32] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[C]//Bengio Y, LeCun Y. 3rd International Conference on Learning Representations. San Diego, CA, USA, 2015.

[33] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, **9**(8): 1735-1780.

[34] Schuster M, Paliwal K K. Bidirectional recurrent neural networks [J]. IEEE Transactions on Signal Processing, 1997, **45**(11): 2673-2681.

[35] Graves A, Liwicki M, Fernández S, et al. A novel connectionist system for unconstrained handwriting recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2009, **31**(5): 855-868.

[36] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Ghahramani Z, Welling M, Cortes C, et al. Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. Cambridge, MA, USA: MIT Press, 2014: 3104-3112.

[37] Wang C, Yang H J, Bartz C, et al. Image captioning with deep bidirectional LSTMs[C]//Proceedings of the 24th ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2016: 988-997.

[38] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Berlin, Germany: Association for Computational Linguistics, 2016: 207-212.

[39] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Luxburg U V, Guyon I. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long. Red Hook, NY, USA: Curran Associates Inc., 2017: 6000-6010.

[40] Paszke A, Gross S, Massa F. PyTorch: an imperative style, high-performance deep learning library[C]//33rd Conference on Neural Information Processing Systems. Vancouver, Canada, 2019.

[41] Kingma D, Ba J. Adam: a method for stochastic optimization[C]// Bengio Y, LeCun Y. 3rd International Conference on Learning Representations. San Diego, CA, USA, 2015.

[42] Rogers D, Hahn M. Extended-connectivity fingerprints[J]. Journal of Chemical Information & Modeling, 2010, **50**(5): 742-754.

[43] Jaeger S, Fulle S, Turk S. Mol2vec: unsupervised machine learning approach with chemical intuition[J]. Journal of Chemical Information & Modeling, 2018, **58**(1): 27-35.

[44] Lipton Z C. The mythos of model interpretability[J]. Communications of the ACM, 2018, **61**(10): 36-43.