基于深度学习的有机太阳能电池能量转换效率 ~~QSPR~~ 预测模型~~的深度学习结构~~

吴金奎[1]，周利[1]

（[1]四川大学化学工程学院，四川省 成都市 610065；）

摘要：

本文提出了一种针对有机化合物的类~~似~~语言分子~~的有机化合物~~描述符，并在此基础上建立了一个预测太阳能电池功率转换效率（PCE）的深度学习模型，力求提供高精度的预测和良好的结果解释性。受有机化学与自然语言的相似性的启发，本文提出基于最近邻子图理论将分子分解成片段（词），并用广度优先搜索算法将片段排列成一定的序列（句子）。通过将每个片段的信息嵌入到一个数值向量中，分子可以表示为一个信息矩阵。这种矩阵被称为 g-FSI，它是一种嵌入片段序列信息的新型分子描述符。然后通过一个深层神经网络对描述子进行解析，提取嵌入的信息并与相应的 PCE 关联。在这个过程中，注意机制被用来识别对 PCE 重要的片段。三个权威数据集被用来训练和评估所提出的方法。结果与现有方法的比较表明，该方法在精度上具有竞争力。该方法还识别出了几个具有决定性意义的碎片，为 OPV 的逆向设计提供了指导信息

关键词：有机光伏材料；类语言描述符；深度学习；效率预测；太阳能电池

中图分类号：□□□□□□　　**文献标识码：**A

A deep learning architecture in QSPR modelling for the prediction of energy conversion efficiency of solar cells

Jinkui Wu[1]，Li Zhou[1]

（[1]*Sichuan University*, *Chengdu* 610065, *Sichuan*, *China*）

Abstract:

This paper proposes a language-like descriptor for organic compounds, and based on which a deep learning model for the prediction of power conversion efficiency (PCE), striving to provide highly accurate prediction and good result interpretability. Inspired by the similarity between organic chemistry and natural language, this work proposes to disassemble a molecule into fragments (words) based on the nearest neighbor subgraph theory, and arranges the fragments into certain sequence (sentence) by breadth first search algorithm. By embedding the information of each involved fragment into a numeric vector, a molecule can be represented as an information matrix. This matrix is called g-FSI, which is a novel molecular descriptor that embeds the fragment sequence information. The descriptor is then parsed by a deep neural network to extract the embedded information and correlate to the corresponding PCE. During this procedure, attention mechanism is applied to identify fragments that are important to PCE. Three authoritative data sets are used to train and assess the proposed method. Result comparisons between the proposed method and other existing approaches indicate that the method is competitive in accuracy. Several decisive fragments are also identified by the proposed method, providing instructive information for the reverse design of OPV

Key words: organic photovoltaics; language-like descriptor; deep learning; efficiency prediction; solar cells

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

联系人：周利，女，四川大学特聘副研究员，邮箱：chezli@scu.edu.cn. **第一作者：**吴金奎（1996—），男，硕士研究生.

展仍然具有挑战性，并且很大程度上依赖于功率转换效率（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。

这些缺点和失败的可能性促使社会发展模型来指导 OPV 的发展。Scharber 模型[11]广泛应用于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）的能级来估计大块异质结太阳能电池的最大 PCE。尽管这些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，但它可以提供一个有价值的指示，说明一种候选化合物可能实现的潜在性能。最近有报道称，在半经验模型分析的指导下，PCE 出现了一个新的峰值（17.35%）[12]。结果表明，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价值的支持[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。

在 OPV 领域，哈佛大学建立了三个权威数据库，并作为开源资源，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 50000 个非完全受主材料的集合，这些材料的 PCE 值来自校准的 HOMO/LUMO 能量[19]。HOPV15 是对文献中的光伏实验数据的一个小的整理，以及在一系列符合物上进行的相应的量子化学计算[18]。虽然数据是基于各种假设获得的，并且受到限制，但也可以从数据揭示的现象中得出见解和设计原则。Hachmann 等[20]通过分析来自 CEPDB 的数据，确定了关键的构建块，并为潜在的 OPV 候选对象提取了设计规则。

对于前向性能预测模型的开发，通常由编码和映射两部分组成。编码过程将化学组成和结构信息转换成一系列称为描述符的数值，而映射过程确定了一个理想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 的相关性。Sahu 等[22]用量子化学计算得到的 13 个微观性质作为小分子 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一些新的微观属性（如基态几何结构和阳离子和阴离子）作为描述符，进一步推进了研究，得到的模型相关性为 0.78[23]。此外，Sun 等[24]还应用卷积技术从分子结构图中提取描述子，用于有机化合物 PCE 性能的估计，得到了一个精度为 91.2%的卷积神经网络模型。随着机器学习的发展，一些最新的策略如转移学习也被用来提高预测与实验之间的一致性[25]。

虽然已经做出了许多努力，但未来的发展仍然面临挑战。首先，量子化学计算生成的描述符可以提供准确的结果，但这通常需要建模者提供深入的领域知识，并且计算成本高昂，这限制了它们在 HTVS 中的有效性。二是化合物的微观物理化学环境非常复杂。容易获取的描述词容易遗漏重要的化学信息，导致预测结果不太理想；许多机器学习模型都是“黑匣子”，其结果往往具有很低的可解释性。这项工作致力于缓解这一困境。

本文提出了一种新的分子特征提取和表征策略，并在此基础上建立了有机化合物 PCE 值预测的深度学习模型，力求为高性能 OPV 的虚拟筛选提供方便。它的灵感来源于 Cadeddu 等的工作[26]，作者们分别在分子片段和文本片段的频率上验证了有机化学和自然语言之间的相似性。首先，将每个有机化合物的分子图分解成片段，并根据其相对位置和连接性对每个片段进行编号。通过将每个涉及的片段表示为一个唯一的片段向量，有机化合物就可以被表示为一个考虑到片段序列信息的类语言描述符。第二，应用自然语言处理技术来“理解”描述符，并将分子信息与其潜在的 PCE 性能相关联。第三，广泛采用的神经网络解释器、注意力机制[27]来识别对有机化合物潜在 PCE 性能重要的分子片段，旨在提高模型的可解释性，为更好的 OPV 设计获得见解。

1 研究方法

在这一部分中，介绍了一种结合 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，用于确定有机化合物的分子结构与 PCE 值之间的相关性。用该方法建立可靠的 QSPR 模型的过程包括以下四个步骤。

i） 数据采集和预处理。收集了有机化合物的 PCE 值和 SMILES 字符串的实验测量或量子化学计算[28]。

ii）分子预编码。候选分子的 SMILES 字符串用于生成片段序列列表，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。（如图 2）

iii）编码器预测器网络培训。基于片段嵌入矩阵，进一步将片段序列嵌入到分子描述符中，然后利用分子描述符和相应的 PCE 值训练 DNN。在训练过程中，对描述器和 DNN 模型参数进行优化，以提高模型性能。（如图 3 所示）

iv）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了评价。

2.1 数据收集

从 CEPDB 收集 25000 个候选 OPV 的 SMILES 字符串和 PCE 值[17]。对收集到的数据进行处理，以消除有异常值和缺失值的数据项；并删除不现实的 PCE 值，如负值和零值。另外，采用 Z 评分标准化对 PCE 值进行预处理，加快了模型训练过程的收敛速度，提高了模型的精度。公式（1）给出了 Z 评分标准化的公式。

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

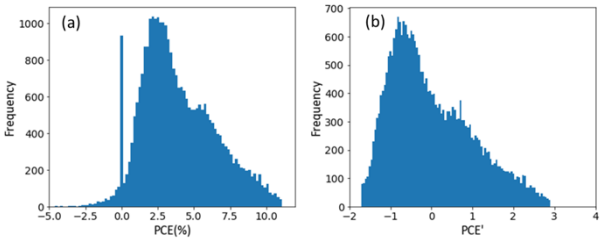

图 1：CEPDB 数据库的分布：（a）预处理之前（b）预处理之后

Figure 1: The distribution of data in CEPDB: (a) before data pre-processing, and (b) after data pre-processing.

其中，y_i是分子的 PCE 值，$\overline{y}$和$s(y)$分别是收集的 OPV 的所有 PCE 值的平均值和标准差。图 1 给出了在数据预处理之前和之后收集到的数据的分布。

2.2 有机化合物预编码

有机分子中含有与自然语言中句子片段的秩分布基本相同的片段[26]。受这一事实的启发和自然语言处理技术的启发，我们提出将一个有机分子分解成碎片，并根据其连接序列信息对其进行排列。这样，有机分子就可以转化成一系列的片段序列。这一过程如下所述。

SMILES 字符串通过 RDKit 和 Networkx 转换成分子图[31, 32]。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 2 所示，为了更好地编码片段的连接性信息，采用最近邻子图[32]和广度优先搜索（BFS）算法[33]将分子（图）分解成片段（子图）[33]，并按相应的 BFS 顺序排列片段。这样，就生成了一个有序的片段序列。值得一提的是，通常不同的有机化合物由不同数量的碎片组成。为了保证分子的片段序列具有相同的维数，识别出组成片段最多的分子，并将其组成片段的计数设为片段序列的维数。当为其他分子生成片段序列时，会填充零值以确保维度的一致性。

在这里，最近邻子图是由最近邻顶点和距离当前顶点一跳内的边所诱导的，这意味着每个原子表示，即碎片，是其最近化学环境的总和。由于 OPV 分子中只有少量的原子和键，因此它比普通的化学信息嵌入方法更为有效和高效。此外，Cadeddu 还证明了有机化合物片段与句子中的单词相似。BFS 遍历算法考虑了分子结构片段的连通性信息，提高了 QSPR 模型的精度。在预编码之后，每个化合物被表示为一个唯一的序列。

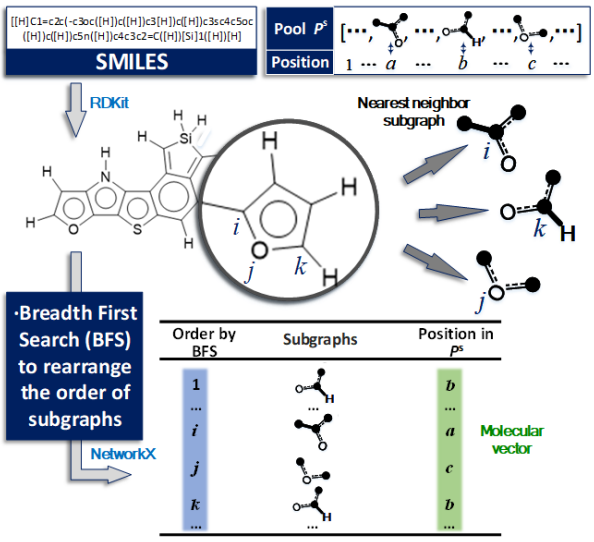

图 2：基于分子图的嵌入

Figure 2: Embedding based on a given molecular graph

1.3 片段序列嵌入类语言描述子和预测模型训练

生成的片段序列表示为一个热矩阵（表示为矩阵 I），如图 3（a）所示。包含池中所有相关片段的嵌入向量（每行）的片段嵌入矩阵（矩阵 II）是生成的矩阵 I 和矩阵 II 的乘积得到相应化合物的分子描述符（矩阵 III）。分子描述符随后被传递到 Bi

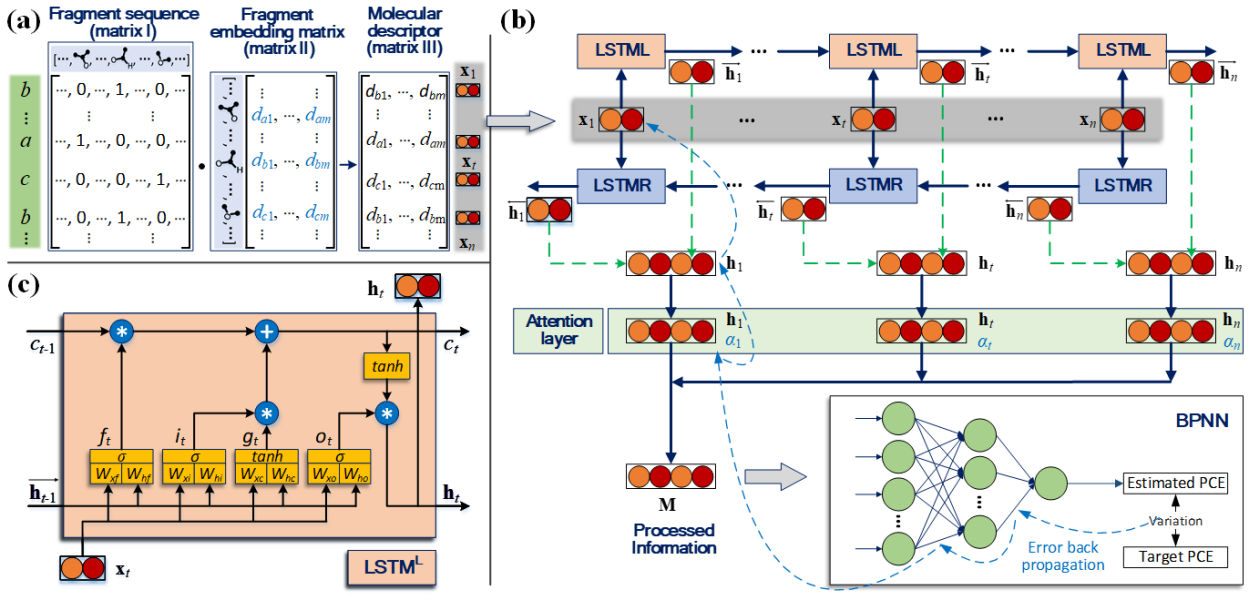

图 3: 预测网络：（a）嵌入过程，m 是嵌入向量的维度，本案例 m=128 每一个嵌入矩阵代表一个分子；（b）LSTM 单元的结构图；（c）拥有注意力机制的 Bi-LSTM 网络

Figure 3: Predictor network: (a) the Embedding procedure, and the m is the dimension of the embedded vectors, in this case, m = 128; (b) the LSTM cell sketch map; (c)Bi-LSTM network with attention mechanism.

LSTM 网络。为了有效地提取片段序列信息，描述符的每一行（片段向量）由正向 LSTM 单元和反向 LSTM 单元同时处理。处理后的片段信息分别表示为$\overrightarrow{\mathbf{h}_t}$和$\overleftarrow{\mathbf{h}_t}$，然后将这些信息馈送到后续的正向和反向 LSTM 单元中。$\overrightarrow{\mathbf{h}_t}$和$\overleftarrow{\mathbf{h}_t}$的组合构成了注意力层的输入[34]。在注意层上，标准化权重α_t用于指示片段的重要性并提高模型的性能。在注意层之后，将处理后的信息 **M** 送入 BPNN 网络，实现 PCE 值的预测。在训练过程中，以均方误差（MSE）作为损失函数来评估回归模型的性能，并将模型误差向后传播以更新片段嵌入矩阵以及 Bi-LSTM 和注意层中的参数。选择随机梯度下降算法作为损失函数优化算法，使均方误差最小。下面将更全面地描述 Bi LSTM 网络和注意力机制。

1.3.1 用于碎片级特征集成的 Bi-LSTM 网络

Bi-LSTM 网络被用来从分子描述子中提取更深层次的分子特征。它是一种具有处理长序列能力的增强型递归神经网络（RNN），它同时考虑了前向和后向序列中嵌入的上下文信息[35, 36]。它被广泛地应用于序列数据处理中，如无约束手写体识别[37]、机器翻译[38]、图像字幕[39]等。对于一个给定的分子p，对应的描述符为$[\mathbf{x}_1,\ldots,\mathbf{x}_t,\ldots,\mathbf{x}_n]$（这里$n$表示是片段序列维，$\mathbf{x}_t$是片段向量）如图 3（b）所示，当前分子的每个片段向量都是前向和后向 LSTM 单元的输入，处理后的片段信息$\overrightarrow{\mathbf{h}_t}$/$\overleftarrow{\mathbf{h}_t}$被传递到下一个单元。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[35]。

正向 LSTM 单元如图 3（c）所示。在单元中引入了三个门单元来控制信息流，一个遗忘门单元f_t，一个输入门单元i_t，以及一个输出门单元o_t。所有这些具有由单元g_t生成的当前信息的门控制单元c_t和隐藏单元h_t的状态。这些门单元和单元状态的信息处理过程服从方程(2-6)，其中一系列的权重矩$\mathbf{W}_{xi},\mathbf{W}_{hi},\mathbf{W}_{xf},\mathbf{W}_{xc},\mathbf{W}_{hc},\mathbf{W}_{xo},\mathbf{W}_{ho}$和偏置参数$\mathbf{b}_i,\mathbf{b}_f,\mathbf{b}_c,\mathbf{b}_o$用来处理当前单元格$\mathbf{x}_t$的信息输入，以及前一个单元格$\overrightarrow{\mathbf{h}_{t-1}}$生成的状态信息。门单元之后的处理信息随后被用于更新当前单元状态c_t及其隐藏状态输出$\overrightarrow{\mathbf{h}_t}$，[40]其公式在式（7）中给出。

$$i_t = \sigma\left(\mathbf{W}_{xi}x_t + \mathbf{W}_{hi}\overrightarrow{h_{t-1}} + \mathbf{b}_i\right) \quad (2)$$

$$f_t = \sigma\left(\mathbf{W}_{xf}x_t + \mathbf{W}_{hf}\overrightarrow{h_{t-1}} + \mathbf{b}_f\right) \quad (3)$$

$$g_t = \tanh\left(\mathbf{W}_{xc}x_t + \mathbf{W}_{hc}\overrightarrow{h_{t-1}} + \mathbf{b}_c\right) \quad (4)$$

$$c_t = i_t g_t + f_t c_{t-1} \quad (5)$$

$$o_t = \sigma\left(\mathbf{W}_{xo}x_t + \mathbf{W}_{ho}\overrightarrow{h_{t-1}} + \mathbf{b}_o\right) \quad (6)$$

$$\overrightarrow{\mathbf{h}_t} = o_t\tanh(c_t) \quad (7)$$

本文采用的双 LSTM 网络包含n正向 LSTM 单元和n反向 LSTM 单元。经过信息处理后，得到一组隐藏状态，用于前向和后向信息提取。Bi LSTM 网络的信息提取总结在式（8）和（9）中呈现。

$$\overrightarrow{LSTM}([\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_t,\ldots,\mathbf{x}_n]) = [\overrightarrow{\mathbf{h}_1},\overrightarrow{\mathbf{h}_2},\ldots,\overrightarrow{\mathbf{h}_t},\ldots,\overrightarrow{\mathbf{h}_n}] \quad (8)$$

$$\overleftarrow{LSTM}([\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_t,\ldots,\mathbf{x}_n]) = [\overleftarrow{\mathbf{h}_1},\overleftarrow{\mathbf{h}_2}\ \ldots,\overleftarrow{\mathbf{h}_t},\ldots,\overleftarrow{\mathbf{h}_n}] \quad (9)$$

之后反向单元格$[\overleftarrow{\mathbf{h}_1},\overleftarrow{\mathbf{h}_2}\ \ldots,\overleftarrow{\mathbf{h}_t},\ldots,\overleftarrow{\mathbf{h}_n}]$以及由前向单元格$[\overrightarrow{\mathbf{h}_1},\overrightarrow{\mathbf{h}_2},\ldots,\overrightarrow{\mathbf{h}_t},\ldots,\overrightarrow{\mathbf{h}_n}]$将拼接作为$\mathbf{h}_t$。获得的信息向量$\mathbf{h}_t$随后被用作后续注意层的输入。

$$\mathbf{h}_t = [\overrightarrow{\mathbf{h}_t}\oplus\overleftarrow{\mathbf{h}_t}] \quad (10)$$

1.3.2　分子水平上的注意神经网络特征提取

并不是所有的组分片段对有机化合物的 PCE 性能都有相同的贡献。因此，采用注意力神经网络来跟踪对 PCE 性能有重要影响的片段。在片段级提取的特征通过与标准化的重要性权重向量相乘，合并成分子级特征向量。处理后的碎片信息$\mathbf{h}_t$被送入一个单层 MLP（多层感知器）得到$\mathbf{u}_s$，其中引入了权重向量$\mathbf{W}_s$，并引入了偏压参数$\mathbf{b}_s$，然后，所得结果进一步用于通过 softmax 函数生成标准化重要性权重α_t。然后，得到信息向量的加权和作为处理后的分子信息 **M**。计算公式如下[27]：

$$u_t = \tanh(\mathbf{W}_s\mathbf{h}_t + \mathbf{b}_s) \quad (11)$$

$$\alpha_t = \frac{\exp(\mathbf{u}_t^T\mathbf{u}_s)}{\sum\exp(\mathbf{u}_t^T\mathbf{u}_s)} \quad (12)$$

$$\mathbf{M} = \sum_t \alpha_t\mathbf{h}_t \quad (13)$$

其中$\mathbf{u}_t^T$是在网络训练过程中随机初始化和共同学习的片段上下文向量.

1.4　模型验证

本文对所建立的 QSPR 模型的预测性能、竞争力和可移植性进行了评估，通过使用测试数据集来评估预测精度。与其他 8 个预测模型的比较，我们评估了该模型的外部竞争力。

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 平台上成功部署。同时，我们使用一个开源的深度学习框架 Pytorch[41]来实现神经网络，并在两个 GTX-1080Ti GPU 上加速了该过程。

2　结果与讨论

2.1　实验超参数设置

收集到的 CEPDB 数据集随机分为三个集：训练集、开发集和测试集。开发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[42]作为损失函数优化器，学习率为 0.001。经过模型训练和开发过程，得到了一个优化的 QSPR 模型。最终优化模型的 bp 神经网络包含 3 层，每层 32 个隐层单元。然后利用测试集对得到的 QSPR 模型进行评价。图 4 给出了由获得的模型预测的 PCE 值与计算的理论值之间的比较。对于开发和测试数据集，可以观察到，预测结果沿对角线分布紧密。计算得到的 QSPR 模型的 Pearson 相

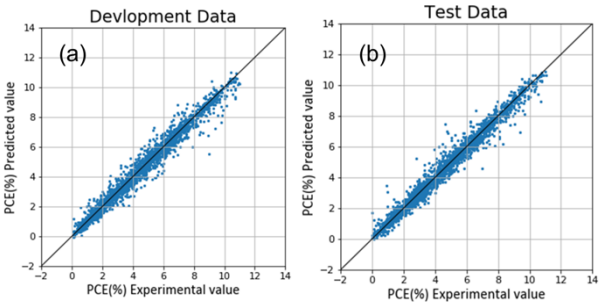

图 4：真实值和 QSPR 模型预测值的散点图

Figure 4: Scatter plots for the predicted - experimental value with the QSPR model

关系数为 0.98，开发集和测试集的预测均方误差分别为 0.13 和 0.12。因此，可以得出结论，所得到的 QSPR 模型具有较高的精度。

2.2　模型的竞争性

为了验证所提出方法的合理性和可靠性，在同一个数据集上还应用了其他几种分子描述符和机器学习方法来建立 QSPR 模型。另一个选择的分子

表 1：测试集的预测精度

Table 1: Prediction accuracy of the testing set

Index	Measure	MSE	R^2
1	Mordred + RF	0.56	0.90
2	sum-g-FSI + RF	0.26	0.95
3	ECFP + RF	0.26	0.95
4	Mordred + ANN	0.55	0.90
5	sum-g-FSI + ANN	0.27	0.95
6	ECFP + ANN	0.26	0.95
7	g-FSI + Transformer	0.85	0.85
8	g-FSI + RNN	0.28	0.95

9	g-FSI + Bi-LSTM + attention	0.12	0.98

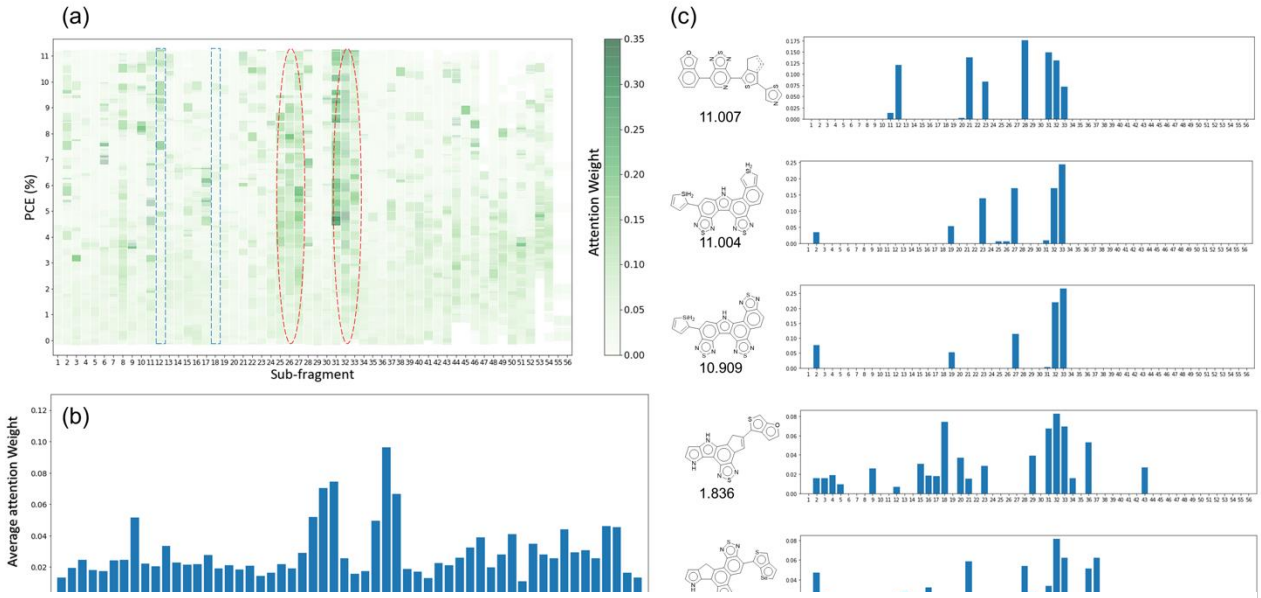

图 5: 注意力机制的可视化：（a）56 个有机太阳能电池碎片的注意力热度图；（b）不同碎片的平均注意力系数图；（c）五个有机太阳能电池分子的注意力权重图

Figure 5: Visualization of the attention mechanism: (a) attention alignment map for the 56 organic solar cell fragments; (b) average attention alignment of the various fragments; (c)attention alignment map for five concrete cases.

描述符包括扩展连接指纹（ECFP）[43]和 Mordred 描述符[44]；和选择的机器学习算法是普通 RNN、transformer[27]、ANN 和 RF。由于所提出的分子描述子（g-FSI）是用矩阵表示的，因此不能直接作为 ANN 和 RF 的输入。为了便于比较，通过对片段向量求和，将其转化为向量，并将得到的向量（sum-g-FSI）作为 ANN 和 RF 的输入。对于 ECFP 和 Mordred，它们都是固定长度的向量，因此生成的描述符直接用作 ANN 和 RF 的输入。共得到 9 个 QSPR 模型。表 1 给出了 QSPR 模型的性能比较。

从三个分子描述符 sum-g-FSI、ECFP 和 Mordred 的比较可以看出，sum-g-FSI 和 ECFP 的相关系数较高，为 0.95，MSE 较低，分别为 0.26 和 0.27，说明 sum-g-FSI 和 ECFP 能够更好地满足当前研究的需要。根据结果，sum-g-FSI 和 ECFP 表现性能相似。这是因为，本质上，它们是基于分子片段信息的相同类型的分子描述符。而对于 RF 和 ANN 这两种机器学习方法，本文给出了等效的建模能力。对于 sum-g-FSI 和 ECFP，由 RF 和 ANN 建立的模型计算出的相关系数、MSE 和平均绝对误差（MAE）几乎相等。虽然观察到细微的差异，但这很可能是射频建模不确定性的结果。

对于 sum-g-FSI 和 g-FSI，前者是后者的压缩表示。它忽略了分子的片段序列信息。然而，有趣的是，g-FSI 和传统 RNN 的结合并没有优于 sum-g-FSI 和 RF/ANN，这意味着传统的 RNN 未能有效地利用所提出的描述子中嵌入的碎片序列信息；这是因为传统的 RNN 不能有效地处理梯度消失或爆炸的问题。应用 Bi-LSTM，模型相关系数提高到 0.98，均方误差减小到 0.12。

至于 Transformer，本研究之所以选择它，是因为它也可以将矩阵作为输入，并且近年来在自然语言处理中表现出了广泛接受的性能，但是在这种情况下它并没有达到预期的效果。

2.3 注意力机制对重要碎片的分析

对于材料设计，结果的可解释性不亚于机器学习模型的预测精度[45, 46]。与专家的经验直觉和/或经验类似，模型“学习”过程中获得的信息对更好的 OPV 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 性能的重要片段。

对于分子片段，越重要，在模型训练中获得的注意权重越大。图 5（a）说明了在训练数据集中，56 个片段通过具有不同 PCE 的有机化合物获得的注意权重的热图。颜色越深，注意力权重越高。可

以观察到，在大多数有机化合物中，有两个片段群受到关注，即片段 25-27 和 31-33。如图 5（b）所示，这些片段的平均注意力权重也高于其他片段。

从测试集中提取出 5 种含有片段 32 和 33 的化合物（如图 5（c））所示。通过比较这五种化合物的碎片组成，可以得到有趣的观察结果。对于前三个 PCE 值大于 10.5%的分子，其性能主要受平均注意量较高的片段或相对较轻但多出现在高 PCE 化合物中的片段的影响。例如，第一种化合物的 PCE 性能主要受片段 12、21、23、28 和 31-33 的影响。除片段 31-33 外，其余 4 个片段在 PCE 高于 5%的化合物中均表现出较高的冲击力。更具体地说，如果你检查片段 12 的热图分布，你会发现上面 PCE=6%上方编码的颜色明显较暗。而对于其余两种 PCE 值低于 2%的化合物，不难发现另一种类型的片段也起了重要作用。以第四个化合物为例，其中一个重要的贡献片段是 18，其中在热图上显示的整体注意力权重分布大多为浅色。此外，PCE 较低的两组在注意权重分布上比其他三组更为平均。我们推断类似 18 的片段可能会降低化合物的潜在 PCE 性能。

3 结　论

光伏技术被认为是解决本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。本文提出了一种可解释的方法来建立精确可靠的有机光伏材料 QSPR 模型。受有机化学与自然语言的相似性启发，提出了一种类似语言的分子描述符来描述有机化合物，并建立了一个深度学习模型，实现了高精度的 PCE 值预测。将由原子和键组成的碎片嵌入到数值向量中，并根据碎片的序列信息将相关向量聚合成矩阵。人们认为，与句子相似，片段（词）的位置信息对复合词（句）的性质预测（意义理解）具有重要意义。采用 Bi-LSTM 对分子描述符进行处理，使嵌入的片段序列信息能够被完全“理解”。然后，将处理后的信息反馈给 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子和机器学习算法相比，该模型具有更高的预测精度和更好的可移植性。结果比较表明，片段的序列信息对有机化合物的 PCE 性能预测具有重要意义，改进的递归神经网络可以捕捉到这些信息。此外，所建立的方法能在一定程度上揭示碎片对分子 PCE 性能的影响，为 OPV 的逆向设计提供了依据。

描述符生成和属性映射过程都是自动完成的，避免了整个过程中的人为干预。也就是说，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，在所提出方法的“学习”过程中，识别出几个决定性的片段和相应的积木，表明所提出的方法能够为 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 OPVs 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能的预测。

参考文献

[1] T. Leijtens, G. E. Eperon, A. J. Barker, G. Grancini, W. Zhang, J. M. Ball, A. R. S. Kan-dada, H. J. Snaith, A. Petrozza, Energy & Environmental Science2016, 9, 11 3472.

[2] H. Dong, H. Zhu, Q. Meng, X. Gong, W. Hu, Chemical Society Reviews2012, 41, 5 1754.

[3] M. Jo˘st, L. Kegelmann, L. Korte, S. Albrecht, Advanced Energy Materials2020, 1904102.

[4] M. Kaltenbrunner, M. S. White, E. D. G lowacki, T. Sekitani, T. Someya, N. S. Sariciftci, S. Bauer, Nature Communications2012, 3, 1 770.

[5] K. Fukuda, K. Yu, T. Someya, Advanced Energy Materials2020, 20007651.

[6] L. Meng, Y. Zhang, X. Wan, C. Li, X. Zhang, Y. Wang, X. Ke, Z. Xiao, L. Ding, R. a. Xia, Science2018, 361, 6407 1094.

[7] H. Jinno, K. Fukuda, X. Xu, S. Park, Y. Suzuki, M. Koizumi, T. Yokota, I. Osaka, K. Takimiya, T. Someya, aNature Energy2017, 2, 10 780.

[8] G. J. Hedley, A. Ruseckas, I. D. Samuel, Chemical Reviews2017, 117, 2 796.

[9] Liu, K. Wang, X. Gong, A. J. Heeger, Chemical Society Reviews2016, 45, 17 4825.

[10] C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng, S. P. Ong, Advanced Energy Materials2020, 10, 81903242.

[11] M. C. Scharber, D. M¨uhlbacher, M. Koppe, P. Denk, C. Waldauf, A. J. Heeger, C. J. Brabec, Advanced Materials2006, 18, 6 789.

[12] M. A. Green, E. D. Dunlop, J. Hohl-Ebinger, M. Yoshita, N. Kopidakis, A. W. Ho-Baillie, Progress in Photovoltaics: Research and Applications2020, 28, 1 3.

[13] R. G´omez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D. G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. Baldo, R. P. Adams, A. Aspuru-Guzik, Nature Materials2016, 15, 10 1120.

[14] E. Kim, K. Huang, S. Jegelka, E. Olivetti, Npj Computational Materials2017, 3, 1

[15] S. Wu, Y. Kondo, M. aki Kakimoto, B. Yang, H. Yamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu, J. Shiomi, C. Schick, J. Morikawa, R. Yoshida, Npj Computational Materials2019, 5, 1.

[16] H. Yamada, C. Liu, S. Wu, Y. Koyama, S. Ju, J. Shiomi, J. Morikawa, R. Yoshida, ACS Cen-tral Science2019, 5, 10 1717.

[17] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. S´anchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, A. Aspuru-Guzik, The Journal of PhysicalChemistry Letters2011, 2, 17 2241.

[18] S. A. Lopez, E. O. Pyzer-Knapp, G. N. Simm, T. Lutzow, K. Li, L. R. Seress, J. Hachmann, A. Aspuru-Guzik, Scientific Data2016, 3, 1 160086.

[19] S. A. Lopez, B. Sanchez-Lengeling, J. de Goes Soares, A. Aspuru-Guzik, Joule2017, 1, 4 857. [3] M. Jo˘st, L. Kegelmann, L. Korte, S. Albrecht, Advanced Energy Materials2020, 1904102.

[20] J. Hachmann, R. Olivares-Amaya, A. Jinich, A. L. Appleton, M. A. Blood-Forsythe, L. R.Seress, C. Rom´an-Salgado, K. Trepte, S. Atahan-Evrenk, S. Er, S. Shrestha, R. Mondal, A. Sokolov, Z. Bao, A. Aspuru-Guzik, Energy Environ. Sci.2014, 7, 2 698.

[21] D. Padula, J. D. Simpson, A. Troisi, Materials Horizons2019, 6, 2 343.

[22] H. Sahu, W. Rao, A. Troisi, H. Ma, Advanced Energy Materials2018, 8, 24 1801032. [3] M. Jo˘st, L. Kegelmann, L. Korte, S. Albrecht, Advanced Energy Materials2020, 1904102.

[23] H. Sahu, F. Yang, X. Ye, J. Ma, W. Fang, H. Ma, Journal of Materials Chemistry A2019, 7, 29 17480.

[24] W. Sun, M. Li, Y. Li, Z. Wu, Y. Sun, S. Lu, Z. Xiao, B. Zhao, K. Sun, Advanced Theory andSimulations2019, 2, 1 1800116.

[25] A. Paul, D. Jha, R. Al-Bahrani, W.-k. Liao, A. Choudhary, A. Agrawal, Proceedings of theInternational Joint Conference on Neural Networks2019, 2019-July, July 1.

[26] A. Cadeddu, E. K. Wylie, J. Jurczak, M. Wampler-Doty, B. A. Grzybowski, AngewandteChemie International Edition2014, 53, 31 8108.

[27] W. David, Journal of Chemical Information and Computer Sciences1988, 28, 1 31.

[28] S. Aksoy, R. M. Haralick, Pattern Recognition Letters2001, 22, 5 563.

[29] RDKit: Open-source cheminformatics, http://www.rdkit.org, [Online; accessed 2-April-2020].

[30] A. A. Hagberg, D. A. Schult, P. J. Swart, In G. Varoquaux, T. Vaught, J. Millman, editors, Proceedings of the 7th Python in Science Conference. Pasadena, CA USA, 200811 – 15.

[31] F. Costa, K. De Grave, ICML 2010 - Proceedings, 27th International Conference on MachineLearning2010, 255–262.

[32] N. Meghanathan, InRouting Protocols and Graph Theory Algorithms for Mobile Ad Hoc Networks, 971–1411. IGI Global, 2017, URLhttp://services.igiglobal.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-5225-2227-0.les3.

[33] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align andtranslate, 2014.

[34] A. Graves, Neural Computation1997, 9, 8 1735.

[35] M. Schuster, K. Paliwal, IEEE Transactions on Signal Processing1997, 45, 11 2673.

[36] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, J. Schmidhuber, IEEE Transac-tions on Pattern Analysis and Machine Intelligence2009, 31, 5 855.

[37] I. Sutskever, O. Vinyals, Q. V. Le, Advances in Neural Information Processing Systems2014, 4, January 3104.

[38] C. Wang, H. Yang, C. Bartz, C. Meinel, InProceedings of the 2016 ACM on Multimedia Con-ference - MM '16. ACM Press, New York, New York, USA, ISBN 9781450336031, 2016988–997, URLhttp://dl.acm.org/citation.cfm?doid=2964284.2964299.

[39] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, B. Xu, InProceedings of the 54th AnnualMeeting of the Association for Computational Linguistics (Volume 2: Short Papers). Associ-ation for Computational Linguistics, Stroudsburg, PA, USA, ISBN 9781510827592, 2016207–212, URLhttp://aclweb.org/anthology/P16-2034.

[40] A. Vaswani, G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Advances in neural information processing systems2017, , Nips 5998.

[41] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alch´e-Buc, E. Fox, R. Garnett, editors, Advances inNeural Information Processing Systems 32, 8024–8035. Curran Associates, Inc., 2019, URLhttp://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[42] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014.

[43] D. Rogers, M. Hahn, Journal of Chemical Information & Modeling2010, 50, 5 742.

[44] H. Moriwaki, Y.-S. Tian, N. Kawashita, T. Takagi, Journal of REFERENCES Cheminformatics 2018, 10, 1 4.11

[45] S. Bang, P. Xie, H. Lee, W. Wu, E. Xing, 35th International Conference on Machine Learn-ing, ICML 20182019, 21386.

[46] Z. C. Lipton, Communications of the ACM2018, 61, 10 36.

格式要求：

（1）版心尺寸：宽 16.8 cm，高 23.7 cm。正文以下改为双栏排，栏宽：8 cm，栏间距：0.8 cm。

（2）全文按顺序包括：题目，作者姓名、单位，中文摘要，中文关键词，中图分类号，文献标识码，文章编号，英文题目，作者英文姓名、单位，英文摘要，英文关键词，正文，符号说明，参考文献。首页地脚注明联系人、第一作者及其简介（姓名、出生年、性别、学位、职称）、基金项目及其编号。

（3）须给出论文的中图分类号，按《中国图书馆分类法》确定。

（4）文中的层次编号用阿拉伯数字，并以“1”、“1.1”、“1.1.1”形式编排。引言不编号。

（5）图、表、公式依出现的顺序编号。

（6）图题、表题采用中英文对照，其他内容（包括分图题、图注、表注等）全部采用英文，Times New Roman，图题、表题小五号，其余六号。

（7）图序和图题置于图的下方，表序和表题置于表的上方。

（8）表的结构应简洁，具有自明性，采用三线表。表头物理量对应数据应纵向可读。

（9）公式中文字小五号。采用公式编辑器时，标准 9 磅，下标 6.5 磅，次下标 5 磅。

（10）物理量注意用斜体。组合单位用指数形式，如 $J\cdot kg^{-1}$，不用 J/kg 形式。数字与单位之间加空格。图表中物理量与单位间用斜线。

（11）符号说明按英文字母顺序排列，同一字母先排大写后排小写；希腊文接英文后排，也按字母顺序排列。

（12）参考文献以在正文中引用的先后顺序排列，序号加方括号。