DOI: 10.15849/IJASCA.230320.14

Int. J. Advance Soft Compu. Appl, Vol. 15, No. 1, March 2023
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Predicting Organic Photovoltaic Solar Cells Properties with Deep Transfer Learning and Multi-Head Self-Attention

Nassima Aleb
Jubail Industrial College
Kingdom of Saudi Arabia
alebn@rcjy.edu.sa

Abstract

Organic Photovoltaic solar cells (OPV) are the most promising technology for renewable energy. However, they suffer of a substantial weakness, that is their low Power Conversion Efficiency (PCE). This has motivated the scientific community to focus on the development of machine learning approaches for the rapid and efficient screening of OPV materials and the prediction of their properties. However, a huge obstacle faced by these methods is datasets scarcity. Two datasets are well-known in this domain: The first is the Harvard CEP (HCEP) dataset, which is a very large computational dataset which values correlate very poorly to experimental measurements. The second is the HOPV small dataset, that contains experimental accurate instances, but that is not large enough for machine learning techniques. The aim of our work is to develop a deep learning approach based on transfer learning and multi-head attention to take advantage from both datasets to predict efficiently OPV experimental properties. We use Depthwise Separable Convolutional networks and Simplified Molecular Input Line Entry System: SMILES as molecular representation. Multihead self-attention allows a profound learning of hidden patterns in SMILES sequences. Our approach shows very promising results. It demonstrates that recent advances in deep learning techniques can play a pivotal role in the systematic design of highly efficient OPV materials, which brings great promises for the future of global renewable energy needs.

Keywords: *Prediction of OPV devices properties, Transfer learning, Attention-based models, Multi-head self-attention, Depthwise Separable convolution models.*

1 Introduction

The global energy future is amongst the foremost worldwide concerns due to the high and ever increasing global needs. Nowadays, the most important source of energy is fossil fuel [1]. However, there are many good reasons to look for alternatives. Indeed, in addition to the scarcity of resources, CO2 pollution causes irremediable damage to the earth. Therefore, it becomes extremely imperative to seek alternative energy solutions, that are clean, cost effective and inoffensive. Renewable energy is the only candidate solution, since it is, as required: clean and harmless. However, it suffers from two limitations: first, it is very expensive, and it is difficult to generate quantities of electricity as large as those produced by fuel generating energies [2]. Renewable resources, including wind, solar cells, solar thermal, geothermal, marine and tides, still

Received 4 January 2023; Accepted 1 March 2023

represent a minor fraction of the overall energy supply [3]. At present, wind is the main alternative energy resource [2], even though wind and hydroelectric solutions require expensive installations and maintenance [4]. Therefore, in order to limit these costs, in the future it is imperative that a greater need for energy be supplied by other, cheaper forms of renewable energy. Solar power is an excellent choice for an abundant, reliable and environmentally friendly energy source and it offers a more sustainable and cost effective solution. Every day, the Earth receives a substantial amount of energy from the sun. Compared to the current energy consumption on Earth, the sun provides enough energy to meet our consumption many times over. Solar cells generating directly electricity are used to harness the sun's energy. The performance of these devices relies heavily on the utilized chemical compounds, specifically: donors and acceptors. However, the selection of the most effective compounds is a very long and laborious task, besides that it heavily relies on experts' knowledge and skills. To facilitate and accelerate this process, many computational approaches have been developed. Methods based on machine learning techniques are among the most popular due to their enormous success in many other fields. Our method belongs to this class; it is a transfer learning attention-based approach for OPV properties prediction. As it is well known, deep learning methods are data intensive, to be effective they require huge datasets. However, in the OPV domain, the most widely used dataset is the modestly sized Harvard OPV dataset: HOPV [5], which contains 344 samples. Despite the accurate quality of its data, it is clear that the mere use of this dataset cannot accomplish good prediction performance due to its small size. On the other hand, another large dataset, HCEP [6], was created by the Harvard Clean Energy Group through extensive Density Functional Theory (DFT) calculations. the Harvard HCEP dataset contains molecular structures and properties for 2.3 million candidate donor structures for OPV [7]. Nevertheless, its data are obtained by a large number of calculations using many approximations, making its values relatively imprecise. Therefore, this makes it unsuitable for prediction tasks despite its substantial size. For ML approaches to achieve high levels of performance, a crucial aspect is the choice of training data. Training using computationally determined PCEs has the advantage of large and standardized datasets with controllable and known degrees of freedom [31] but these PCEs correlate poorly to experimental measurements which can undermine their utility. Conversely, training using experimentally characterized PCEs is harder due to the small amount of available data. The aim of this paper is to take advantage of both datasets for OPV properties prediction, in order to improve the obtained performance.

Contribution : In this paper, we propose a new approach based on a transfer learning model, using multi-head self-attention with Depthwise Separable Convolutional networks for accurate prediction of OPV properties. Our work is motivated by the recent advancement achieved by the application of transfer learning and attention based-models in various fields. Thus, we define an approach that attempts to get the most out of transfer learning, attention mechanisms, and Depthwise convolution networks to provide an accurate prediction of OPV properties. The introduction of the Attention Mechanism in deep learning has improved the performance of various models in recent years, and continues to be an omnipresent component in state-of-the-art models. Our approach uses simply SMILES for molecules representation. In a first stage, the HCEP large dataset is used to learn the hidden features of donors SMILES sequences. Then when a suitable validation performance is achieved, the weights are frozen, and the resulting network is used as features extractor for the HOPV dataset which is the dataset of interest. Fully connected and prediction layers are added to the target network to refine its knowledge by considering the properties of the HOPV dataset. Thus, we achieve our main objective

to predict the properties of the HOPV dataset using the HCEP large dataset. The experiments show that our proposed method is very efficient, as our model shows very promising results. The rest of the paper is organized as follow: Section II introduces briefly some necessary background about the OPV technology, to ease the reading for the user unfamiliar to this domain; a literature review is also presented. In section III, we present the used datasets, and the architecture of our model. Section VI is devoted to the conducted experiments, it introduces the performance metrics, hyperparameters tuning, obtained results, and a comparison with baseline methods. Finally, the last section concludes the paper by highlighting the contributions and presenting some future perspectives.

2 Background on OPV Technology and Related Work

To facilitate the understanding of this document for a reader unfamiliar with the solar cells domain, we introduce a brief description of Organic Photovoltaic (OPV) technology. We briefly present solar cells successive generations, and explain, succinctly, how OPV solar cells work, how they recover energy, and what are their most important parameters. Afterwards, we present a review of the literature of related works.

2.1. Solar Cells and OPV Technology

Solar cells are used to harvest energy from sunlight. Their most important characteristic criterion is the Power Conversion Energy (PCE), that reflects the rate achieved by the device in converting the sunlight incoming power in electricity. To improve this factor, several solar cells technologies have been developed. Globally, there are three generations of solar cell technologies. The first is essentially fabricated of silicon. Its advantages consist in its good performance, measured by a PCE of around 15-20%. [8] as well as their great stability. However, they are rigid and their manufacture is very expensive in terms of time and prices. The second generation attempts to reduce these costs by using amorphous silicon. However, PCEs dropped a bit to 10-15%, [9] and their fabrication still involves vacuum processes and high temperature treatments. In addition, they are based on rare elements. The last generation is based on organic photovoltaic (OPV) technology. Organic or plastic (polymer) photovoltaic (OPV) technology has become very popular due to its flexibility [10]. In addition, manufacturing costs are the lowest due to the ease of fabrication of the devices and the lower cost of organic components compared to silicon [11]. Accordingly, the use of non-scarce materials, like polymers, allows an affordable largescale production. These characteristics make them a promising candidate for solving most of the problems present in other solar cell technologies. The chemical compounds involved in OPV are donors, usually made of semiconducting polymers, and acceptors. The donor electrons are excited by the photons absorbed from the sunlight. The generated exciton bond is broken in the donor-acceptor interface, resulting in an electron-hole pair, that is subsequently separated, and collected at the opposite electrodes. Electricity is generated as a consequence of the movement of the electrons from the cathode to the anode. Organic solar cell fabrication is basically simple. However, the relationship between materials properties, and performance is crucial. Given the huge number of potential materials and device configurations, a comprehensive experimentation, cannot guarantee finding the best performing materials. Six properties are fundamental for OPV: The power conversion efficiency (PCE), the highest occupied molecular orbital (HOMO), the lowest unoccupied molecular orbital (LUMO), the HOMO–LUMO gap (Gap), the open circuit potential (Voc), and, finally,

the short circuit density (Jsc). These parameters can be computed by Density functional theory (DFT), but the cost in computational resources and time is exorbitant, besides the usually needed calculations approximations. Consequently, developing new OPV materials is commonly based on materials scientist's expertise and intuition. Thus, an intensive effort of synthesis, characterization, and prototype device optimization, are necessary to concretize any new design idea. Furthermore, the main bottleneck in this task is that the search for candidate chemical compounds to create organic solar cells is iterative and very time-consuming [12]. Given sufficient and accurate data, Machine Learning (ML) can potentially model the complex relationships between materials, device properties, and OPV performance, which allows avoiding expensive and time-consuming experiments and quantum chemical calculations. Therefore, some recent methods have been developed for this purpose.

2.2. Related Work

Enhancing the efficiency of organic photovoltaic devices is essential to compete with the currently used silicon-based solar cells. However, arduous trial-and-error experimental processes, besides DFT calculations requiring substantial computational time that is not favorable to fast screening. Therefore, many computational methods have been developed to overcome these difficulties, by screening and predicting OPV properties of materials from their chemical characteristics. Machine Learning (ML) methods are one class of these methods that are often used to derive quantitative structure property relationships (QSPR) between the performance of the organic photovoltaic and the underlying properties of the materials, as they can make use of existing computational and experimental data and make predictions at a fraction of the cost. Machine learning techniques, with their noteworthy capability of identifying hidden patterns in data, allow to detect and take advantage of statistical correlations that are in concordance with theoretical laws of physics, with competitive accuracy and lower computational costs. This characteristic makes them a very promising candidate to accelerate materials design and discovery. As a results, various ML-based approaches have been developed for OPV properties prediction. Generally, we can classify these methods in traditional ML-based approaches and Deep Learning: DL-based ones. Some examples of the first class are : Riede et al. [13] developed a method based on PCA using a small size dataset consisting in 62 organic solar cells dataset. Their data was obtained by assessing parameters on seven different reactants. Olivares-Amaya et al. [14] employed a set of 200 organic materials descriptors obtained from 50 training molecules, by using ChemAxon Marvin code [15]. In another side, Mannodi-Kanakkithodi et al. [16] used fingerprints as chemical compound representation. They make use of kernel ridge regression (KRR) to transform the input fingerprint into higher dimensional space that is, subsequently, linearly related to a target property. They used well-known polymers computations alongside learned patterns extraction to accelerate the polymer design. Kanal et al. [17] presents another approach, for OPV properties prediction, that makes use of genetic algorithms for successive screening and refining research of acceptors and donors of suitable properties. A multiclass model based on random forest for PCE prediction was developed by Nagasawa et al. [18]. They created an experimentally tested database of polymer OPV devices containing approximately 1000 instances. However, even by using simultaneously: molecular fingerprints, highest occupied molecular orbital values, the bandgap, and the molecular weight for molecules description; their model obtained an accuracy of 48%. Another work is Sahu et al. [19], developing methods based on RF, and the well-known gradient boosting (GB) regression trees using 13 microscopic properties

(DFT) derived molecular descriptors to model PCE for 280 small OPV molecules. As reported in [19], the obtained results were not promising. Another method of this class are the models developed for PCE prediction by Padula et al. [20]. The authors modeled the photovoltaic properties of 249 organic donors–acceptors pairs using K-Nearest Neighbors (KNN), regression and Kernel Ridge Regression (KRR) methods trained on a combination of electronic and structural parameters. Pereira et al. [21] also developed ML- methods, consisting of Random Forest (RF) and Support Vector Machine (SVM), for PCE prediction. They used a dataset of 111,725 molecules fingerprint and modified distance descriptors. They reported that random forest obtained the best results. We conclude the list of ML-based methods with the work presented by Zhao et al. [22], that trained SVM, kNN, and KRR models with a variety of different DFT derived descriptors and 566 experimental PCEs. Recently, some DL-based methods have been developed. They differ mainly on the used models, and datasets. Pyzer-Knapp et al. [23] used a multi-layered perceptron (MLP) neural network for PCE prediction. They used the large-size HCEP [6] dataset to extract DFT-derived properties, they achieved good prediction accuracy for PCE and other molecular properties. Various deep learning models have also been applied, including the convolutional neural network of Sun et al. [24], which categorized organic photovoltaic candidates into two classes : low performance (<5% PCE) and high performance (5-10%PCE) materials, they developed five ML models (MLP, Deep Neural network (DNN), Convolutional Neural Network, RF, and Support Vector Machine (SVM)) that enables recognition of chemical structures and automatic classification for predicting the PCE of donors' materials. A different method is presented in Kaya et al. [25], they presented a screening of polymer molecules with artificial neural network (ANN) and random forest (RF) by using parameters such as PCE, molecular weight, energy levels, and electronic properties with digitized chemical structures. Meftahi et al. [26] developed BRANNLP: Bayesian Regularized Artificial Neural Network with Laplacian prior method. They used signature descriptors to generate models for the six OPV properties. Their models were trained on the HOPV dataset. Latest approaches have integrated the attention mechanism, originally introduced for recurrent neural networks for machine translation [27] in order to improve performance by focusing on local substructures that are relevant for the prediction task. For example, Wu et al. [28] used attention to couple a Bidirectional Long Short-Term Memory and multilayer perceptron (MLP) to predict PCE based on sequentialized molecular structure and fragment types. The authors achieved a very high degree of prediction accuracy on the HCEP and identified functional groups that contribute towards a molecule having higher PCE. However, other examples comprise the Attentive Fingerprints (FP) used by Xiong et al. [29] and Simple GNN [30] that is defining a simple Graph Neural Network for PCE prediction, based on descriptions of molecular structure obtained from SMILES strings and fingerprint analysis. Hence, as it is noticeable, there remain several limitations to the application of ML for the screening of OPV materials. The most penalizing one is the insufficiency or inadequacy of the used data. This is all the more serious, since the ML-based methods are data-driven, regarding both quantity and quality of data. This situation of data scarcity has forced researchers to create their own datasets, which are not large enough for ML techniques in addition that their quality has not been assessed, and they don't allow any comparison with other previous approaches. In another side, using HCEP dataset containing computationally determined PCEs has the advantage of large and standardized datasets with controllable and known degrees of freedom [31] but these PCEs correlate poorly to experimental measurements which can undermine their utility. All these drawbacks have serious repercussions on the performance of the results obtained which, still, need improvements.

3 Materials and Methods

We formulate the problem as a multi-output regression task that requires the model to predict the continuous values representing: PCE, HOMO, LUMO, Gap, Voc, and Jsc. We use molecules SMILES sequences as donors' representation. Hence, the SMILES are first submitted to a preprocessing phase. In this section, we introduce the used datasets, their preprocessing and the proposed method. We define two models: HCEPMODEL and HOPVMODEL, that are learning respectively from HCP, and HOPV datasets. A transfer learning approach is used to relate the knowledge, learned by the HCEPMODEL from HCEP dataset, to HOPV dataset by the HOPVMODEL. This approach of using another model as a feature extractor to a target dataset is known as transductive transfer learning, since we have similar tasks and different domains. The overall idea is about leveraging feature representations from a pre-trained model, that is trained on massive dataset. These models are integrated into the process of training a new model. The weights obtained from the HCEPMODEL are used directly to initialize the weights of the new HOPVMODEL. Including the pre-trained models in a new model leads to overcome the problem of reduced size dataset, decrease training time and generalization errors.

3.1 Datasets

Our approach uses two datasets that are: The Harvard Clean Energy Project dataset: HCEP [6], and The Harvard Organic Photovoltaic dataset: HOPV [5]. The HCEP dataset contains only computational results for approximately 2.3 million organic solar cell donor candidate materials. It provides substantial data for training machine learning techniques. The molecular structures in the HCEP are generated from 26 different building blocks [6]. Each molecule donor in the HCEP is assigned an ID, with the SMILES string, the PCE, the short circuit current density Jsc, the open circuit voltage Voc, the HOMO energy, the LUMO energy, and the HOMO-LUMO gap are reported as computed by the extensive DFT calculations described in [6]. The PCE values reported in the HCEP are computed using Scharber equations [32] Due to the wide use of approximations, it was noticed a poor correlation of HCEP values with experimental measurement. The second major dataset used in this work is the HOPV dataset, which consists of 344 different experimentally characterized organic solar cell donor structures that have been collected from various works in the literature by Lopez et al. [33]. For each molecule, the HOPV data set contains the SMILES and InChI strings that define the molecular structure. The experimental data in HOPV also includes measured photovoltaic characteristics of the organic solar cell: HOMO and LUMO energies of the donor, the electrochemical and optical gaps of the donor, as well as the PCE, Jsc, Voc. Besides the experimental properties, HOPV contains data from quantum chemical calculations. HOPV calculated properties are calibrated in a consistent way [26] and are known to correlate well with the experimental measurements. In our work, we use the HOPV dataset in its entirety, with both experimental and calculated values, as all its instances are consistent with the experimental measurements. However, it is clear that HOPV size is not suitable for effective deep learning. Thus, we make use of HCEP dataset which is much larger than HOPV, to learn intrinsic OPV properties. The learned skills are used to extract the most important features of HOPV and to predict new OPV candidates' properties.

3.2 Data Preprocessing

Our model uses The Simplified Molecular-Input Line-Entry System: SMILES [34] as molecule representation. SMILES is a line notation that uses printable characters for describing the structure of chemical elements. Our choice for SMILES as a molecular representation is motivated by the recent works in drug discovery, that showed their appropriateness for various prediction tasks. SMILES is a true language, although with a small vocabulary size and only few grammar rules [34]. In our model, SMILES are represented by a one-hot encoding, using integers to represent sequences characters. SMILES sequences have variable lengths; the maximal length is 190. The sequences that are shorter than the maximum length are 0-post-padded. Vocabulary size is equal to 64, that is the number of unique characters in the SMILES language.

3.3 Models Architecture

In this section, we describe our approach and present the architecture of the proposed models. We formulate the problem as a multi-output regression problem aiming to use SMILES donor molecules to predict continuous measures representing: PCE, HOM, LUMO, Gap, Voc, Jsc for the HOPV dataset. To achieve this goal, we first develop a model that is used to learn OPV donors features from the HCEP dataset: HCEPMODEL. First, a one-hot encoding is applied to each SMILES sequence. Then, we use an embedding layer to represent sequences' characters with high-dimensional (128-dimensional) dense vectors. As we said, we developed two models: the source model, HCEPMODEL, that is trained on the HCEP dataset, and the target model, HOPVMODEL, that is operating on the HOPV dataset to extract its features, refine them, and finally perform properties prediction. In this section, we describe each of these models. The models' architectures are presented in Figure 2. The HCEPMODEL is based on multi-head self-attention mechanism for OPV properties prediction. In our work, we define our own multi-head self-attention layer, that is operating on the HCEP dataset to inspect donors' structures, more details are provided in the next paragraph and in Figure 1. The HCEP dataset is submitted to a Depthwise separable convolutional layer, to extract the feature maps of molecular structures that are scrutinized by the multi-head attention layer to detect and focus on interesting areas in each molecular sequence, and therefore identify how these features are related to OPV properties. Opting specifically for Separable CNN is justified by two reasons: They are faster in training, besides, they are less prone to overfitting. The layer number and filters size are among the hyperparameters that have an impact on the model performance [35]. We used a layer with 256 filters having size 4 as it is usually recognized that filters of this size perform well. The HOPV dataset is fine tuned for the downstream task, to refine the acquired prior knowledge. In this way, we make use of both datasets in an appropriate way.

3.3.1. Multi-Head attention for OPV

Attention is a very powerful concept that has shown its success in the recent deep learning approaches in several domains. The Attention mechanism is based on the idea of guiding the focus of a learning model to the most interesting part of data. It has been shown to be very useful in many machine learning tasks. Several forms of attention have been defined. Multi-head Attention is a powerful technique that runs through an attention mechanism several times in parallel. It is a combination of multiple self-attention structures, in which each head learns features in different representation subspaces, leading to multi-representations that are, afterwards, merged to enhance the overall model performance. The independent attention outputs are usually concatenated and

linearly transformed into the expected dimension. To produce the Query, Key and Value of our sequences representation, we don't use linear projections as it is usually implemented, instead, we make use of Depthwise separable convolutional layers to extract various maps. Let X be the features of the SMILES sequence obtained from the previous block, Figure 1, presents the equations for self-attention weights calculations together with the attention feature maps, obtained for one head, d is the projection dimension.

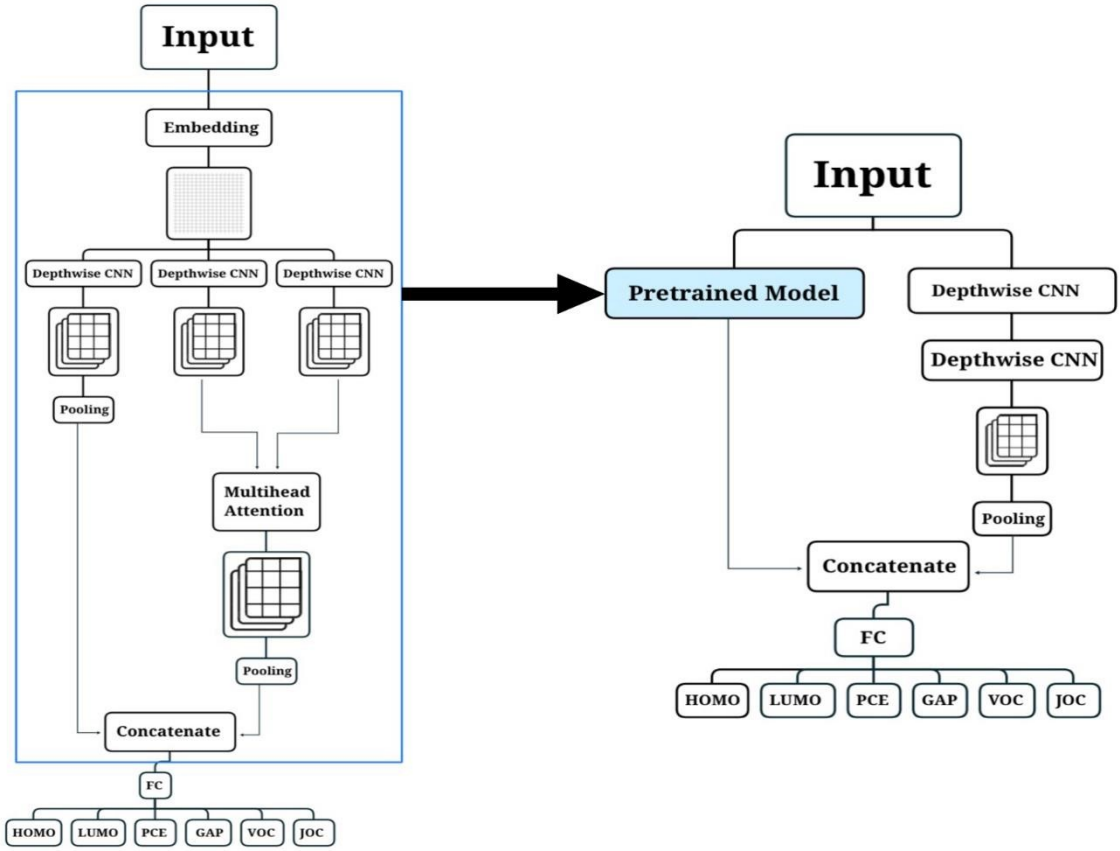
$$Q = Depthwise\text{-}Convolutional\text{-}1\text{-}1\text{-}Projection(X)$$
$$K = Depthwise\text{-}Convolutional\text{-}1\text{-}1\text{-}Projection(X)$$
$$V = Depthwise\text{-}Convolutional\text{-}1\text{-}1\text{-}Projection(X)$$
$$A = Align\ (Q, K) = Q * K^{t}$$
$$W = = \frac{Softmax(A)}{\sqrt{d}}$$
$$AttentionMap = W*V$$

Figure 1. Self-Attention Equations

After a pooling of both the obtained attention map and the hidden convolution feature maps, they are concatenated and submitted to a Fully Connected block (FC) for the prediction task. FC is constituted of three Dense layers, having respectively the output dimensions: 1024, 512 and 128. The activation functions used for the fully connected layers are all Rectified Linear Units (ReLU) that have been widely used in deep learning studies [36]. To prevent overfitting, we make extensive use of Dropout with various proportions. Finally, the prediction layer contains, a unique node and is activated with the 'Linear' activation function. For regularization, we used Dropout and Batch Normalization techniques [37] to avoid overfitting. Given that we intend to compare our method with previous approach, we used Mean squared error MSE as a loss function, since this is the usually used one. We used a mini-batch size of 128 to update the weights of the network, and Adam optimizer [38]. We opted for a dynamic tuning of the learning rate by using a Callback that starts with an initial value, and reduces it dynamically during the training when the monitored loss value is stationary, indicating that the model is not improving. the initial learning rate value was set to 1e-3, and the minimum value was set to 1e-6. The learning was completed within 150 epochs.

3.3.2. Transfer Learning

Once the training, of HCEPMODEL, is completed, we use it as a feature extractor for the HOPVMODEL. The weights of HCEPMODEL are frozen and the fully connected block, FC, is removed. The input layer in HOPVMODEL is followed by a block of two separable convolutional layers terminated by a pooling layer. The obtained feature maps are subsequently concatenated with the attention maps obtained from the pretrained sub-model. Finally, a fully connected block terminate the model. The idea behind the HOPVMODEL architecture is based on the same form of attention models, where an attentive sequence is concatenated to the Value sequence. The models' architectures are shown in figure 2. The first round of training was performed within 150 epochs, using Adam optimizer with a learning rate 1e-5. After the first round of training is completed, the HCEPMODEL weights are unfrozen and a fine tuning phase, using a very small learning rate lr=1e-8, is initiated.



a - HCEPMODEL **b- HOPVMODEL**

Figure 2. Model Architecture

4 Experiments and Results

In this section, we describe the experiments we conducted for our proposed model. We introduce succinctly the hyperparameters tuning, and the used performance metrics. We analyze the effectiveness of our approach, and we compare it against baseline methods.

4.1 Experimental Settings

Deep learning models' performance depends highly on various hyperparameters. For our study, we adopted a trial-error method in hyperparameter tuning. Hence, many combinations of the parameters: number of layers, number of filters by layer, size of filters, were assessed, and updated using the obtained validation results, before deciding about their values. The hyperparameters combination having the best validation results was selected. It is well-known that the learning rate is one of the most crucial hyperparameters. We opted for a dynamic tuning of the learning rate by using a Callback that starts with an initial value, and reduces it dynamically during the training when the monitored loss value is stagnant, meaning that the model is not improving. the initial learning rate value was set to 1e-3, and the minimum value was set to 1e-6. Overfitting

is a common problematic behavior of deep learning models that have been targeted by many researchers. This has led to the apparition of various techniques, in our model, we used dropout in various locations, with different percentages. We included a Batch-normalization in many positions of our model since it is usually recommended with convolutional layers. We also penalized the loss function with L2-norm regularization [39]. Finally, we updated the weights using the Adam optimizer with a penalized loss to give a generalized prediction for the model. We used a mini-batch size of 128 to update the weights of the network. The model was created using Keras [40] and Tensorflow [41]. For our experiments, we split each dataset into 70-10-20 ratio for training, validation, and test sets respectively; we used the same split for each dataset. A shuffling was performed to ensure that the values distribution for all the 3 subsets was similar.

4.2 Performance Metrics

To evaluate our model, we used the mean squared error (MSE) as loss function, and, root mean squared error (RMSE), mean absolute error (MAE) and the R-squared statistic, as performance metrics. In the subsequent, we introduce each of them.

Mean Squared Error: MSE is commonly used in regression tasks to measure how close the line obtained by connecting the predicted values to the actual data points. The formula below defines the MSE, where P denotes the vector of predicted values, Y denotes the vector of the actual outputs, and n is the number of samples. The smaller the MSE, the better the performance of the regressor.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(P_i - Y_i)^2$$

The Root Mean Squared Error: RMSE is the square root of MSE.

Mean Absolute Error: MAE is the average of all absolute errors. The formula is:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\text{xi} - \text{x}|$$

R-Squared: R2 statistic represents the proportion of variation in the outcome that is explained by the predictor variables. It corresponds to the squared correlation between the actual values and the predicted values in multiple regression models. The higher the R-squared, the better the model. R-squared is computed by the formulas below, where Yi is the actual values, the mean of the actual values, and Pi is the predicted ones. RSS is the sum of squares of the residuals also known as Residual sum of squares or RSS. while TSS is the Total variation in target variable it is the sum of squares of the difference between the actual values and their mean.

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$\text{where : } \text{RSS} = \sum_{i=1}^{n}(Y_i - P_i)^2 \quad \text{and} \quad \text{TSS} = \sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2$$

4.3 Baseline Methods

To assess our method, we compared it with four state of the art methods. Except the first, BRANNLP that is predicting all OPV properties; all the other methods predict only PCE value. In this section, we present succinctly each of the baseline methods.

•**BRANNLP [26]:** The authors used the HOPV Dataset to predict all the OPV properties. The aim was to demonstrate that simple molecular descriptors and ML methods can model and predict important OPV properties [26]. They generated several models of all the properties, for the 344 compounds in the dataset. The best results they reported was related to the Bayesian Regularized Artificial Neural Network with Laplacian Prior model: BRANNLP. Their results showed that the PCE models were the most robust and predictive, with R2 > 0.64 for training set and >0.58 for test set prediction. The BRANNLP nonlinear modelling and variable selection method was used to generate the QSPR models. We compare our method with the BRANNLP model as it showed the best results among the other methods they developed and presented in the same paper.

•**g-FSI/BiLSTM [28]:** A deep-learning model using a bi-directional long short-term memory network with an attention mechanism and multilayer perceptron is implemented. As feature extraction process: Atom types and their aromaticity are used to encode the node features whereas bond types (single, double, triple and aromatic bonds) are used to encode the edge features. The WeisfeilerLehman algorithm [42]is used to circulate around the nodes of molecular graphs. The algorithm consists of two stages: (i) an initial assignment stage, in which each node and edge are assigned the unique integer identifiers that represent their starting features set, and (ii) an iterative updating stage, in which each node identifier is updated to reflect the identifiers of its neighbors and each edge identifier is updated to reflect the identifiers of nodes it connects. The final node identifiers are assembled into the final molecular fingerprint vector. The g-FSI vector encodes two important structural aspects: fragment (fingerprint) types and their order in the molecule. An embedding layer is first used, the result is processed by the forward and backward BiLSTM cells resulting in a sequence of hidden forward and backward state vectors. The two BiLSTM hidden states are concatenated and weighted-summed in the attention layer. This is followed by propagating the attention layer output through a multilayer perceptron, in order to obtain the final PCE value.

•**SGNN [30]:** SGNN: Simple Graph Neural Network, belongs to the class of graph convolutional networks. All distinguishable tuples of the form (atom type, atom aromaticity) are identified. Then, each node type is mapped to its own vector in the embedding dimension space. While the node types for Simple GNN are different from the fragment types for the g-FSI/BiLSTM model, their embedding into a continuous space is analogous. A graph neural network processes the resulting embeddings. At the end, a multilayer perceptron with one output neuron is used to predict the PCE value.

•**AttentiveFP [29]:** AttentiveFP (fingerprint) is a graph neural network that was introduced by Xiong et al. [29] and achieves state of-the-art performance on several data sets that are widely used in ML research according to a comprehensive assessment of several neural networks and ML methods. AttentiveFP makes use both of Gated Recurrent Units (GRU, [43]). It applies the attention mechanism both on the atomic node level and on the molecular graph level to focus on substructures that are relevant for predicting target variables. Nine node features (including atom types and atom aromaticity) and four edge features (including bond type) are used as input features, with most of them being one-hot encoded. For each target node v in the molecular graph, its node features are fed into a fully connected layer with FS units where FS denotes the

Fingerprint Size. the node features of each direct neighbor u ∈ N(v) and the features of the edge between u and v are concatenated and fed into another fully connected layer having the same number FS of units. Attention weights are calculated between the states of v and u, then, their sum is fed as input into a GRU. A new node is added to the molecular graph and connected to all of its atom nodes. Its state represents the state of the entire graph and is initialized with the sum of all atom node states updated at the end of the last iteration. Finally, the new node state is fed into a fully connected layer which has one output predicting the PCE target variable.

4.4 Results and Comparison

This section presents our experimental results. We have assessed our models for the considered dataset: HOPV, that is the target dataset of experimental measurements. In the experiments reported in this paper, each dataset has been first shuffled, then was split into two sets, train set and test set. 20% of the train sets were used for validation. Our method is based on deep transfer learning and multi-head attention to predict OPV properties represented by six continuous values. We used R-squared statistics, mean squared error, root mean squared error, and mean absolute error, to assess the predictive power of our method. Table I summarizes our testing results and those of the baseline method BRANNLP in terms of RMSE and R2 metrics (MAE is not used in BRANNLP) for all the OPV properties. The results show clearly the supremacy of our method for all the properties in almost all the metrics.

Model		RMSE	R^2
BRANNLP	HOMO	0.007	0.49
	LUMO	**0.008**	0.67
	Gap	0.01	0.65
	PCE	0.48	0.78
	Voc	0.16	0.58
	Jsc	22	0.60
Our Model	HOMO	**0.003**	**0.87**
	LUMO	0.02	**0.80**
	Gap	**0.007**	**0.76**
	PCE	**0.46**	**0.83**
	Voc	**0.02**	**0.79**
	Jsc	**0.02**	**0.72**

Table I –Testing Results of Our method and BRANNLP for RMSE and R^2 metrics - The best results are in Bold.

Table II presents the results of our method with all the baseline methods for PCE prediction. The results for the baseline method were taken from the experiments reported in the literature. The metrics used in all these methods are: MSE, MAE and R2, so we report our results for these metrics. Additionally, since in the literature corresponding to these methods, the results were reported for training, validation and test, sets, so, we

report our results for all these sets. The experiments show that our method is the most performant since the achieved results are the best results for all the considered metrics.

Model	SET	MSE	MAE	R^2
g-FSI/BiLSTM	Train	1.072	0.780	0.776
	Validation	3.273	1.425	0.363
	Test	3.486	1.480	0.299
SGNN	Train	1.494	0.918	0.711
	Validation	2.641	1.293	0.426
	Test	3.295	1.454	0.330
AttentiveFP	Train	2.936	1.377	0.420
	Validation	3.020	1.397	0.355
	Test	4.417	1.672	0.127
Our Model	Train	0.174	0.290	0.89
	Validation	**0.203**	**0.324**	**0.85**
	Test	**0.243**	**0.364**	**0.83**

Table II –PCE Prediction Results: MSE, MAE and R^2 metrics - The best results are in Bold.

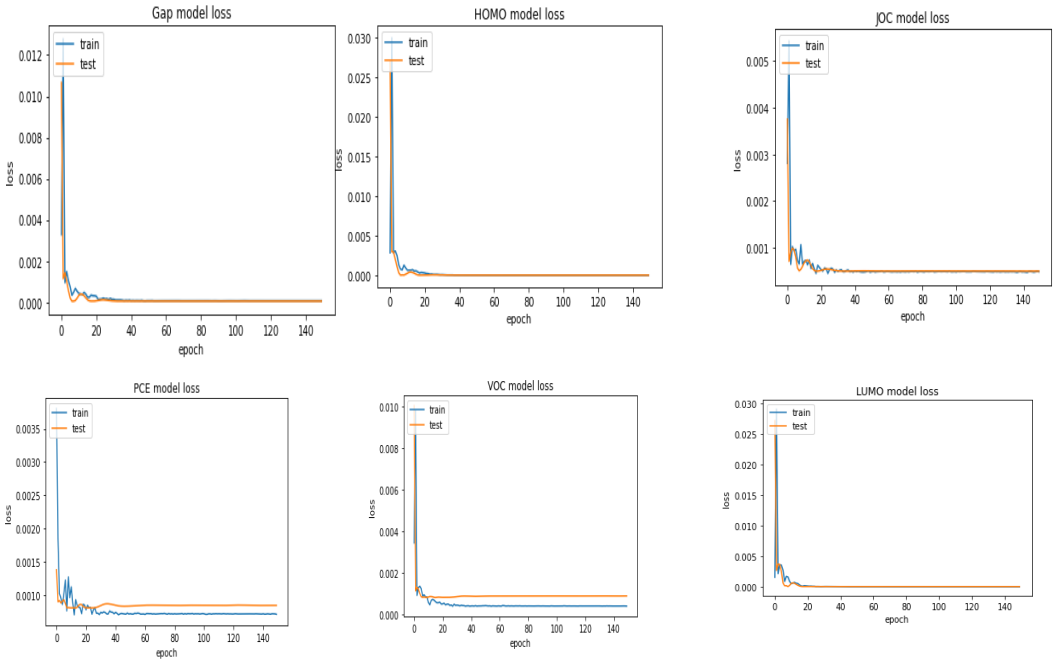


Figure 3. Training curves for all the properties prediction

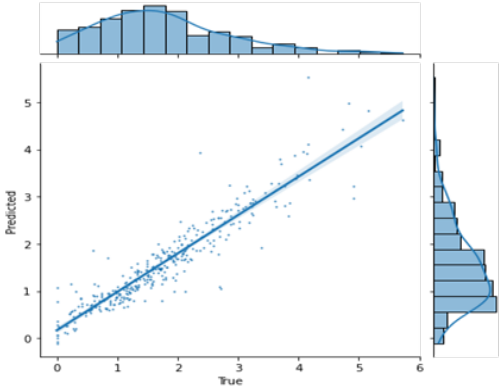


Figure 4. Predicted against Actual PCE values

The learning curves are presented in Figure 3, Figure 4 shows the graphs representing the predicted against Actual values of HOPV dataset. In an "ideal" model, predicted values equal the actual ones. Our graphs show that, points density is high around the line represented by the equation "y=x". This is clearly apparent in the graph

5 Conclusion

In this article, we have introduced a new approach for Organic Photovoltaic solar cells properties prediction. Our approach takes advantage of the recent successful deep learning techniques, that are: transfer learning and attention mechanism. It is well-established that DL-based techniques are data-driven, they depend highly on the quality, the size and the diversity of the underlying datasets. This motivated our choice to use two "complementary" datasets for our study. The first, the HCEP dataset is large enough for any DL study, but its data does not correlate with experimental results, the data of the second dataset, HOPV, is very accurate but its small size is not suitable for DL techniques. Thus, using the two datasets in an adequate way allowed to have all the necessary requirements for our ML method to succeed. We used the Harvard CEP, as a source dataset that contains millions of OPV candidates with their values calculated by DFT. HOPV is the target dataset, it contains both DFT-calculated and experimental data. Our results demonstrate that the implemented techniques are very efficient. Transfer learning from a larger dataset is noteworthy since it allowed a smooth and intense learning of donors hidden features. Additionally, attention mechanism is very efficient in making the learning focusing in interesting parts of donors sequences. Furthermore, using SMILES as the input representation makes it is easier for materials scientists to explore the addition or removal of subgroups to chemical compounds to explore the impact on energy efficiency at the core. We showed that simple molecular representations like SMILES combined with robust DL methods can model and predict important OPV properties. Our objective in this work was also to show that DL methods are well appropriate to model and predict a wide range of OPV properties. Even if it is obviously ideal to model and measure properties experimentally, there are various variables that can impact the performance of the OPV. Therefore, the measured OPV properties may vary between experiments and laboratories. Our study also shows that deep learning techniques might be very useful in various domains, such as: renewable energy and material design. This ability discloses an exciting field in the discovery of materials, and in particular for solar cells technology. In future work, we intend to extend both the scope and the DL-based techniques for OPV properties prediction. For the

scope, we plan to explore more calculated and experimental datasets. Regarding DL-approaches, a wide range of methods could be explored, as the field of machine and deep learning is very active and expanding.

References

[1] O. A. Abdulrazzaq, V. Saini, S. Bourdo, E. Dervishi, and A. S. Biris. (2013). Organic solar cells: a review of materials, limitations, and possibilities for improvement, Particulate science and technology, vol. 31, no. 5, pp. 427–442.

[2] Hossain, Eklas, and Slobodan Petrovic. (2021). Challenges of Renewable Sources of Energy. Renewable Energy Crash Course. Springer, 113-119.

[3] Moriarty P, Honnery D. (2020). Feasibility of a 100% Global Renewable Energy System. Energies. 13(21):5543. Https://doi.org/10.3390/en13215543.

[4] Kaldellis, J. K. (2020). Hybrid Wind Energy Solutions Including Energy Storage. The Age of Wind Energy. Springer, 103-129.

[5] Lopez, S. A. Et al. (2016). The Harvard organic photovoltaic dataset. Sci. Data 3, 160086.

[6] Hachmann, Johannes, Roberto Olivares-Amaya, Sule Atahan-Evrenk, Carlos Amador-Bedolla, Roel S. Sánchez-Carrera, Aryeh Gold-Parker, Leslie Vogt, Anna M. Brockway, and Alán Aspuru-Guzik. (2011). The Harvard Clean Energy Project: Large-Scale computational screening and design of organic photovoltaics on the world community grid. Journal of Physical Chemistry Letters 2(17): 241–2251.

[7] OPV"Cepdata.csv.zip," https://www.dropbox.com/s/3kqzt9u1ryflls0/cepdata.csv.zip?Dl=0

[8] Carvalho, Carlos Manuel Ferreira, and Nuno Filipe Silva Veríssimo Paulino. "Photovoltaic Cell Technologies." CMOS Indoor Light Energy Harvesting System for Wireless Sensing Applications. Springer, Cham, 2016. 43-71.

[9] Di Giacomo, Francesco, et al. (2016). Progress, challenges and perspectives in flexible perovskite solar cells. Energy & Environmental Science 9.10 3007-3035.

[10] Narayan, K. S. (2020). "Photovoltaics: Materials and Devices." Advances in the Chemistry and Physics of Materials: Overview of Selected Topics. 321-349.

[11] Seri, Mirko, et al. (2021). Toward Real Setting Applications of Organic and Perovskite Solar Cells: A Comparative Review. Energy Technology 9.5: 2000901

[12] S. R. Forrest. (2005). The limits to organic photovoltaic cell efficiency, MRS bulletin, vol. 30, no. 1, pp. 28–32.

[13] M. K. Riede, A. W. Liehr, M. Glatthaar, M. Niggemann, B. Zimmermann, T. Ziegler, A. Gombert, and G. Willeke. (2006). Datamining and analysis of the key parameters in organic solar cells. in Photonics Europe. International Society for Optics and Photonics, pp. 61970H– 61970H.

[14] Olivares-Amaya, C. Amador-Bedolla, J. Hachmann, S. Atahanevrenk, R. S. Sanchez-Carrera, L. Vogt, and A. Aspuru-Guzik. (2011). Accelerated computational discovery of high performance materials for organic photovoltaics by means of cheminformatics. Energy & Environmental Science, vol. 4, no. 12, pp. 4849–4861,

[15] Chemaxon Marvin Code. (2016). Software for chemistry and biology. https://www.chemaxon.com.

[16] Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, and R. Ramprasad. (2016). Machine learning strategy for accelerated design of polymer dielectrics. Scientific reports, vol. 6,

[17] I. Y. Kanal, S. G. Owens, J. S. Bechtel, and G. R. Hutchison. (2013). Efficient computational screening of organic polymer photovoltaics. The journal of physical chemistry letters, vol. 4, no. 10, pp. 1613–1623

[18] Nagasawa, S., Al-Naamani, E. & Saeki, A. (2018). Computer-aided screening of conjugated polymers for organic solar cell: classification by random forest. J. Phys. Chem. Lett. 9, 2639–2646.

[19] Sahu, H., Rao, W., Troisi, A. & Ma, H. (2018). Toward predicting efficiency of organic solar cells via machine learning and improved descriptors. Adv. Energy Mater. 8, 1801032.

[20] Padula, D., Simpson, J. D. & Troisi, A. (2019). Combining electronic and structural features in machine learning models to predict organic solar cells properties. Mater. Horiz. 6, 343–349.

[21] Pereira, F. Et al. (2016). Machine learning methods to predict density functional theory B3LYP energies of HOMO and LUMO orbitals. J. Chem. Inf. Model 57, 11–21.

[22] Zhao Z.W, Del Cuoto M. Geng Y. and Troisi A. (2020). Effect of increasing the descriptor set on machine learning prediction of small molecule-based organic solar cells Chemistry of Materials, 32(18): 7777-7787.

[23] E. O. Pyzer-Knapp, K. Li, and A. Aspuru-Guzik. (2015). Learning from the Harvard clean energy project: The use of neural networks to accelerate materials discovery," Advanced Functional Materials, vol. 25, no. 41, pp. 6495–6502,

[24] Sun, W. Et al. (2019). The use of deep learning to fast evaluate organic photovoltaic materials. Adv. Theory Simul. 2, 1800116.

[25] Kaya, M. & Hajimirza, S. (2018). Application of artificial neural network for accelerated optimization of ultra-thin organic solar cells. Sol. Energy 165,159–166.

[26] Meftahi, Nastaran & Klymenko, Mykhailo & Christofferson, Andrew & Bach, Udo & Winkler, David & Russo, Salvy. (2020). Machine learning property prediction for organic photovoltaic devices. Npj Computational Materials. 6. 10.1038/s41524-020-00429.

[27] D. Bahdanau, K. Cho, and Y. Bengio. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473

[28] J. Wu, S. Wang, L. Zhou, X. Ji, Y. Dai, Y. Dang, and M. Kraft.(2020). Deep-learning architecture in QSPR modeling for the prediction of energy conversion efficiency of solar cells. Industrial & Engineering Chemistry

[29] Z. Xiong, D. Wang, X. Liu, F. Zhong, X. Wan, X. Li, Z. Li, X. Luo, K. Chen, H. Jiang, et al. (2019). Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. Journal of Medicinal Chemistry,

[30] Eibeck, Andreas & Nurkowski, Daniel & Menon, Angiras & Bai, Jiaru & Wu, Jinkui & Zhou, Li & Mosbach, Sebastian & Akroyd, Jethro & Kraft, Markus. (2021).

Predicting Power Conversion Efficiency of Organic Photovoltaics: Models and Data Analysis. ACS Omega.10.1021/acsomega.1c02156.

[31] M. C. Scharber, D. Mühlbacher, M. Koppe, P. Denk, C. Waldauf, A. J. Heeger, and C. J. Brabec. (2006). Design rules for donors in bulk-heterojunction solar cells—towards 10% energy conversion efficiency. Advanced Materials, 18(6):789–794.

[32] Lopez, S.A., B. Sanchez-Lengeling, J. de Goes Soares, and A. Aspuru-Guzik. (2017). Design principles and top non-fullerene acceptor candidates for organic photovoltaics. Joule, 1(4):857–870.

[33] Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Model. 28, pp.31–36. Doi: 10.1021/ci00057a005.

[34] Goodfellow I., Bengio Y., Courville A.(2016). Deep Learning. MIT Press,

[35] Lecun, Y. Bengio, and G. Hinton. (2015). Deep learning. Nature, 521(7553):436-444.

[36] Sergey Ioffe, Christian Szegedy. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, Pages 448–456.

[37] Diederik P. Kingma and Jimmy Ba, Adam. (2017). A Method for Stochastic Optimization arxiv pre-print arxiv: 1412.6980.

[38] Jan Kukačka and Vladimir Golkov and Daniel Cremers. (2017). Regularization for Deep Learning: A Taxonomy, Arxiv preprint 1710.10686.

[39] Chollet, F. Et al. (2015) Keras. Https://github.com/fchollet/keras.

[40] Abadi, M. Et al. (2016). Tensorflow: a system for large- learning. In: OSDI scale machine, Vol. 16, pp. 265–283.

[41] B. Weisfeiler and A. Lehmann. (1968). A reduction of a graph to a canonical form and an algebra arising during this reduction. Nauchno-Technicheskaya Informatsia, 2(9): 12–16.

[42] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.