Noname manuscript No.
(will be inserted by the editor)

An Attention-driven LSTM Network for High Throughput Virtual Screening of Organic Photovoltaic Candidate Molecules

Ryan J. Richards[†*] and Arindam Paul[‡]

[†]University of Pennsylvania
[‡]Independent Researcher
[*]Corresponding author: ryry@seas.upenn.edu

Received: date / Accepted: date

Abstract — Organic Photovoltaic (OPV) Solar Cells are a rapidly developing technology with promising capabilities over leading renewable energy sources. Screening methods for determining promising donor and acceptor molecules to augment the efficiencies of such cells can be substantially accelerated through deep learning. Textual descriptors, specifically Simplified Molecular Input Line Entry System (SMILES), are utilized as network inputs, while quantum-chemical calculations based on density function theory (DFT) provide chemically-accurate targets for training and testing. We present a Long Short-Term Memory (LSTM) based network which uses a self-attention mechanism and a robust data augmentation routine to predict several OPV optoelectronic properties (e.g. highest occupied molecular orbital and lowest unoccupied molecular orbital). The LSTM cells, coupled with self-attention, learn the successive ordering and pairing of SMILES characters while attending to certain salient constituents of the molecule, which produce a robust understanding of the molecular graph. The Harvard Clean Energy Project (CEP) and National Renewable Energy Laboratory (NREL) OPV datasets are used for this study. The CEP dataset portion which we use contains $\sim 1.2E6$ candidate donor molecules with their respective DFT-computed properties, whereas the NREL OPV dataset possesses $\sim 9.1E4$ samples. Compared to contemporary graph-based model selections, our network reduces the MAE over all considered optoelectronic properties on the CEP and NREL OPV datasets by an average of 21.23% and 10.06% respectively. Furthermore, we demonstrate that our model generalizes well to the pharmaceutical drug discovery focused ZINC-250k dataset, reducing the MAE across all properties by an average of 28.2% from the current state-of-the-art model.

Keywords Machine Learning · Organic Photovoltaics · Recurrent Neural Network · High Throughput Virtual Screening · Attention · Drug Discovery

1 Introduction

The global warming crisis has induced a heavy demand on clean alternative energy sources, namely solar cells, whose technological development and manufacturing efficiency has been immensely improved over the past few decades. Although inorganic cells (e.g. conventional silicon-based) possess superior performance in terms of power conversion efficiency (PCE), they suffer from complicated, expensive fabrication processes and structural rigidity [1]. Organic Photovolatic (OPV) cells, based on thin film polymers or small molecules, offer simple and cost-efficient fabrication processes, novel applications, but lack a high enough PCE for commercialization [1].

Contemporary methods for locating new candidate compounds for OPV cells, which involve synthesis and evaluation, are exhaustive and laborious, and remain a heavy bottleneck in the screening process [2] [3]. OPV cell design seeks to maximize the PCE, or the percentage of electricity which is generated from absorption of photons. The PCE depends on specific optoelectronic properties of donor and accpetor molecules in the cell, namely the highest occupied molecular orbital (HOMO) energy of the donor and the lowest unoccupied molecular orbital (LUMO) energy of the acceptor [4]. The ability to rapidly and accurately predict these important properties and hence avoid a costly and time-

intensive screening process has been the objective of high throughput computational material design efforts.

OPV materials discovery has been substantially accelerated through High Throughput Virtual Screening (HTVS), whereby large quantities of thermodynamic or optoelectronic properties are generated through simulations or experiments and subsequently used for materials discovery, specifically targeting desirable properties [5, 6]. The most prevalent and accurate simulation methodology is density functional theory (DFT), a computational quantum mechanics modeling routine which determines molecular properties using functionals of the electron density [7]. DFT possesses chemically accurate computations but is severely bottlenecked by its processing time, especially when coupled with the demanding requirements of HTVS [8].

Machine Learning (ML), a field of statistical learning, has directed materials research into a new data-driven science paradigm [5, 9]. ML has the potential to match the chemical accuracy of DFT while significantly decreasing the processing (inference) time [10–12]; ML algorithm prediction times (operating at $\mathcal{O}(10^{-3}s)$) are nearly six orders of magnitude faster than DFT calculations ($\mathcal{O}(10^{3}s)$ on 30 heavy atom molecules) [8].

Demonstrating initial success in pharmaceutical chemistry, ML models have been leveraged to predict more challenging properties such as chemical reactivity, melting point, solubility, and electronic properties [13, 14]. Several descriptors have been utilized in ML frameworks for electronic properties prediction [5], including Coulomb matrices [15, 16], molecular strings or graphs [8, 17–19], and molecular fingerprinting [13, 20].

Among these approaches are natural language processing (NLP) derived techniques which depend on textual representations [21, 22] of molecular structures rather than relying on 2D or 3D-defined structures (i.e. spatial coordinates) [3, 20, 23]. Novel components used extensively in NLP, such as attention mechanisms [24], have shown great promise in the analysis of molecular structures [25–27]. Line notations also permit the usage of augmentation techniques that are easily realized and computationally efficient [26, 28], which greatly improve network performance especially when coupled with attention-mechanisms. Furthermore, NLP techniques allow for deeper analysis of the molecular structure. Multi-dimensional embeddings enable practitioners to generate reduced-space clusters of molecular tokens (predetermined individual or grouped SMILES characters) to understand the learned relationships between certain molecular components [26]. Attention enhances these analytical capabilities by narrowing down specific components of a molecule which most heavily resonate with target properties [26], enabling practitioners to create activation maps of complex molecules. A richer understanding of the encoded structure is especially useful in the automated creation of new molecules, whereby generative networks are required to learn the textual descriptor syntax and the respective semantics [29, 30].

In this work, we create an attention-driven LSTM network with 1D convolutions to predict optoelectronic properties of OPV candidate molecules from the Harvard Clean Energy Project (CEP) [31] and NREL OPV [8] datasets. Such properties include the HOMO and LUMO energies. We enhance our network training by employing a robust data augmentation scheme, which is also exploited during testing. We demonstrate that textual representations are effective descriptors which achieve better results than graph-based models for the considered OPV datasets.

Furthermore, although the intent of this study is to accelerate HTVS for organic solar cells, we also demonstrate the efficacy of our ML framework in the field of drug discovery. Analogous to the challenges in the OPV field, pharmaceutical research involves HTVS of organic molecules to identify suitable drug candidate compounds [32] [33]. We use the the ZINC-250k dataset, which contains 250k drug-like molecules extracted from the ZINC database [34], to predict the log octanol-water partition coefficient (logP) and the quantitative estimate of drug-likeness (QED) [35]. We show that our attention-LSTM model provides better results than leading state-of-the-art variational autoencoder (VAE) based models [36].

2 Related Works

Machine learning models have been successfully applied in materials and molecular design [11, 12, 37–43] by utilizing datasets created by experimental observations and theoretical simulations.

Among these ML works, Decision Tree-based methods such as Random Forests and Extremely Randomized Trees [44] were developed for screening organic monomers used for photovoltaic applications and predicting organic solar cell efficiency [45, 46].

Jorgensen et al. [47] used a VAE with predefined SMILES syntax (grammatical) rules for predicting molecular properties and generating new molecules with desirable properties. The All SMILES VAE [36] significantly improved the results from [47] by introducing a more efficient message passing system, which encodes multiple SMILES strings of the same molecule with stacked recurrent networks, pooling SMILES representations between the multiple inputs, and using at-

tentional pooling to construct the final latent representation; the decoder is then capable of mapping this latent space into a disconnected set of SMILES strings. The All SMILES VAE is capable of efficiently exploring the chemical space, searching for molecules with desirable properties, and can also be leveraged for property prediction (used on the ZINC-250k and Tox21 datasets) [36].

Paul et al. [3] explored the use of multiple line notations (SMILES and InChI) as inputs for a convolution-LSTM network (SINet). SINet aimed to learn unique representations of molecules captured in syntactically different encodings to predict the HOMO energies of the Harvard CEP dataset, while employing transfer learning to predict the HOMO energies of the HOPV-15 dataset [48].

3 Modeling and Assumptions

The common equation quantifying the PCE (η) of a solar cell is provided in 1; given an open-circuit voltage (V_{oc}), short-circuit current density (J_{sc}), electrical fill factor (FF), and incident light intensity (P_{in}).

$$\eta = \frac{V_{oc} J_{sc} FF}{P_{in}} \tag{1}$$

The Scharber model [49] was used to focus on salient optoelectronic properties that most heavily influence η. We make the same initial assumptions as *Scharber et al* [49] regarding FF and J_{sc}. Assuming a practical PCE, the external quantum efficiency (EQE) and FF is set to *65%*. The induced J_{sc} then reduces to an EQE-scaled maximal photo-generated current J_{ph} associated with the Air Mass 1.5 (AM1.5) spectrum, given in 2; where $\widetilde{J}_{sc,Sch}$ is the Scharber-assumed short-circuit current density and $\phi_{ph}(E)$ is the solar photon flux density. Following these assumptions, the FF and J_{sc} reduce to constants which render them negligible for this study.

$$\widetilde{J}_{sc,Sch} = 0.65 J_{ph} = 0.65q \int_{E_g}^{\infty} \phi_{ph}(E) dE \tag{2}$$

The remaining component of η to optimize is V_{oc}, which has been previously identified as a major deficiency for commercialization of bulk-heterojunction solar cells [50]. This limiting factor was investigated [49] by empirically deriving a relationship between V_{oc} and the HOMO energy level of the donor polymer, using [6,6]-phenyl-C61-butyric acid methyl ester (PCBM) as a fixed acceptor. It was deduced that V_{oc} is approximated by the equation given in 3, which ultimately suggests that the predominate factor in attaining a higher V_{oc} is maximizing the difference between the donor HOMO and acceptor LUMO. Fixing the acceptor LUMO suggests that the donor HOMO is the more important component of the equation.

$$V_{oc} = (1/e)(|E^{Donor}_{HOMO}| - |E^{PCBM}_{LUMO}|) - 0.3V \tag{3}$$

However, the PCE is more sensitive to changes in donor LUMO energy rather than strictly its bandgap [49]. For example, a variation of the donor bandgap by 0.65 eV induces a PCE change of 1%, whereas a variation of 0.65 eV of the donor LUMO energy induces PCE changes between 3.5% and 8% (depending on the donor bandgap) [49]. Therefore, it is imperative to optimize the donor LUMO energy when designing solar cells with target efficiencies exceeding 10%. In this work, we aim to construct regression models which accurately predict these essential energies which primarily govern the PCE of an OPV cell.

4 Methodology

4.1 Datasets

Two primary datasets were used in this study: the National Renewable Energy Laboratory (NREL) OPV [8] and the Harvard Clean Energy Project (CEP) [31] [13].

Developed in 2019, the NREL OPV dataset contains $9.1E4$ molecules with DFT-computed optoelectronic calculations specifically for OPV applications. NREL populated the dataset with relatively larger molecules ($\leq$ 201 atoms) when compared to other similar datasets such as QM9 ($\leq$ 29 atoms). The NREL OPV dataset hence stands as a more representative benchmark for electronic structure predictions. NREL utilized the B3LYP/6-31g(d) DFT functional/basis-set combination. The specific optoelectronic properties included in the dataset are: HOMO and LUMO energy levels of the monomer, first excitation energy of the monomer (Gap), and spectral overlap (optical absorption spectrum overlap area between a dimer and AM1.5). Additionally, properties extrapolated to the polymer limit were generated: polymer HOMO and LUMO, polymer Gap and polymer optical LUMO (sum of polymer HOMO and polymer Gap).

The Harvard CEP, created in 2011, featured an automated *in silico*, high-throughput system for screening millions of OPV candidates at first-principles electronic structure level [31]. The CEP sought to advance beyond a sophisticated screening method by also developing a systematic understanding of structure-property relationships, which aids in engineering novel organic

electronics [31]. The dataset portion employed in this work contains ~1.2E6 candidate donor molecules. The optoelectronic properties of which were computed using the BP86/def2-SVP DFT fucntional/basis-set combination; we focus on HOMO, LUMO and Gap for this study.

Finally, we further validate our model and demonstrate its versatility by predicting molecular properties on the ZINC-250k dataset [34]. ZINC-250k contains 2.5E5 drug-like commercially available organic molecules with $\leq$ 38 heavy atoms. In accordance with related works [36, 51], we focus on predicting the log octanol-water partition coefficient (logP) and quantitative estimate of drug-likeness (QED) [35].

4.2 Pre-processing and SMILES Encoding

All molecules in the considered datasets are given in Simplified Molecular Input Line Entry System (SMILES) [52] format. SMILES provides a textual representation of molecules that compresses the atomic connectivity and topological information into a single ASCII string. For example, 2-ethyl-1-butanol is encoded as "CCC(CC)CO". SMILES does not explicitly define protonation of molecules as it can be inferred through predetermined rules.

We consider each SMILES character to be a uniquely trainable component or "token" of each molecule, which is learned through an embedding layer [53]. A dictionary was created from each dataset that maps a set of tokens to an initial set of continuous values of shape $L_{max} \times 1$ where L_{max} is the maximum SMILES length across the entire dataset. The dictionary was used to convert all SMILES strings to their equivalent continuous vectors, $\mathbf{x}_i$, shown in Figure 1 (a)–(b).

A character embedding layer [53] was used to learn a mapping between the initial continuous SMILES vectors, Figure 1 (b), to a 32-dimensional vector space, shown in Figure 1 (c); more information is provided on this specific implementation in 4.4. Word and character embeddings have been used extensively for Natural Language Processing (NLP) tasks and have shown significant improvements over sparse encoding techniques (namely one-hot encoding). These embedding vectors represent projections of the original SMILES characters and are responsible for capturing the semantics of tokens and their relation in the SMILES string [26].

The regression problem is then reduced to minimizing the loss of the network output $f(\mathbf{x}_i)$ given a set of SMILES vectors and their respective ground truth targets $\mathbf{y}_i$ by tuning a parameter set θ, shown in 4.

$$\underset{\theta}{argmin} \sum_i L(f(\mathbf{x}_i : \theta), \mathbf{y}_i) \tag{4}$$

4.3 Augmentation Methods

Data augmentation techniques were used to better train the network on the NREL dataset. Bjerrum [28] first introduced that randomly changing the atomic order of a molecule can yield different SMILES representations for the same molecule, which can be used to generate more input-target pairs when training a neural network. For example, whereas the canonical form of 2-ethyl-1-butanol is CCC(CC)CO, we observe five non-canonical forms which can be used for the same original target:

C(CC)(CO)CC
C(C(CC)CO)C
C(C)C(CC)CO
C(O)C(CC)CC
C(CO)(CC)CC

Bjerrum demonstrated that using this augmentation technique yielded better results for an LSTM-based network; and has since been used in contemporary designs [26]. We designate the augmented samples as $\hat{\mathbf{x}}_i$.

Conformational isomers, molecules with identical connectivity but different atomic positioning, have slightly different optoelectronic properties when computed by DFT - discussed more in 5.2. Although our proposed network utilizes a textual descriptor, this uncertainty is captured by creating noisy targets ($\hat{\mathbf{y}}_i$). Zero-mean Gaussian noise ($\mathcal{N}$) is added to each augmented training sample's target; a mathematical formulation is provided in 5. Hence, the new input-target pairs are given as $(\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i)$

$$\hat{\mathbf{y}}_i = \mathbf{y}_i + \mathcal{N}(\mu, \sigma^2) \tag{5}$$

Adding noise to the augmented samples targets allows the model to generalize better, and is also leveraged during testing. The standard deviation of the Gaussian window was dependent on the acceptable error range provided by the DFT-deduced values for each property (given in Table 1, "Conf" column). Finally, the targets were scaled to have zero median and unit inner quartile range [8]. Hence, the regression problem is simplified to equation 6, where the loss is minimized between the network outputs given the augmented samples ($f(\hat{\mathbf{x}}_i)$) and the noisy targets ($\hat{\mathbf{y}}_i$).

$$\underset{\theta}{argmin} \sum_{i} L(f(\hat{\mathbf{x}}_i : \theta), \hat{\mathbf{y}}_i) \tag{6}$$

This augmentation technique is also exploited for network evaluation, a method of testing that has gained traction in the imaging community [54] referred to as Test-Time Augmentation (TTA). TTA involves executing model inference on augmented test samples; the outputs of which are averaged and used as the final predictions. We deduce that since the network is trained to recognize multiple SMILES permutations, the evaluation results will improve with such augmented SMILES. All results provided from our models for the NREL OPV dataset are the TTA outputs.

4.4 BiLSTM and the Self-Attention mechanism

As discussed in Section 4.2, a character embedding layer is used to understand the semantics of the molecule in terms of its constituent characters, or tokens. A molecule given by n tokens is represented by an embedding matrix E, given in 7. Each vector τ_i is a 32-dimensional token embedding for the ith token in the molecule. The full embedding matrix E has shape: $L_{max} \times 32$.

$$E = (\tau_0, \tau_1, ..., \tau_{n-1}) \tag{7}$$

We utilize an LSTM layer [24] to introduce a dependence between neighbor tokens; and since the encoded SMILES has no inherent direction or time-dependence, we apply the LSTM cells bidirectionally [55] to fully capture contextual details. Bidirectional LSTM (BiLSTM) based models involving character embeddings have demonstrated superb performance in works involving SMILES analysis [56] [26] [25].

For each time-step t, provided a past hidden state $\overrightarrow{h}_{t-1}$, or future state $\overrightarrow{h}_{t+1}$, the LSTM outputs are given as:

$$\overrightarrow{h}_t = \overrightarrow{LSTM}(\tau_i, \overrightarrow{h}_{t-1}) \tag{8}$$

$$\overleftarrow{h}_t = \overleftarrow{LSTM}(\tau_i, \overleftarrow{h}_{t+1}) \tag{9}$$

We then concatenate (σ) these hidden states for each time-step. Hence the final output (h_i) of the BiLSTM for each ith token of E is given in 10. This is further consolidated into a matrix y_i across all tokens of a given molecule.

$$h_i = \sigma(\overrightarrow{h}_t; \overleftarrow{h}_t) \tag{10}$$

$$y_i = (h_0, h_1, ..., h_{n-1}) \tag{11}$$

Network performance is enhanced by appending an Attention layer to the BiLSTM. The Attention mechanism has several variants and have been used extensively in machine learning models; primarily for NLP applications like *neural machine translation* (NMT) to resolve the short-term memory bottleneck of Recurrent Neural Networks (RNNs), which employ LSTM or GRU cells. *Cho et al* [57] showed that the performance of encoder-decoder networks for NMT suffered as the input vectors increased in size; LSTM-based networks would discard learned representations of early words in the sentence and utilize the last state for translation. *Bahdanau et al* [58] created the initial attention mechanism which learns to appropriately weigh all input states in the sentence rather than being limited to its last state. During the decoding phase, the network essentially "attends" to different contextual patterns across the entire input, hence it can make more informed predictions. *Cheng et al* [24] expanded this idea and created the self-attention (or intra-attention) mechanism which relates different positions of a single sequence to learn lexical relations between tokens [59].

Similarly, in the analysis of lengthy SMILES vectors, critical relations between tokens are highly susceptible to being neglected by a simple LSTM/GRU layer. We utilize a self-attention mechanism to exploit all interconnected relationships between tokens [26] [25], enabling the network to more heavily concentrate on salient constituents of the entire molecule which possess a heavier influence on the target value.

The intermediate self-attention matrix (e_i) is provided in 12, provided the concatenated LSTM output (y_i) for a given molecule. We employ multiplicative self-attention which introduces new weight and bias terms (W_a and b_a) and uses ReLU activation (ζ).

$$e_i = \zeta(y_i^T W_a y_i + b_a) \tag{12}$$

The *softmax* function is applied to e_i to generate the final attention matrix (a_i), given in 13.

$$\alpha_i = softmax(e_i) \tag{13}$$

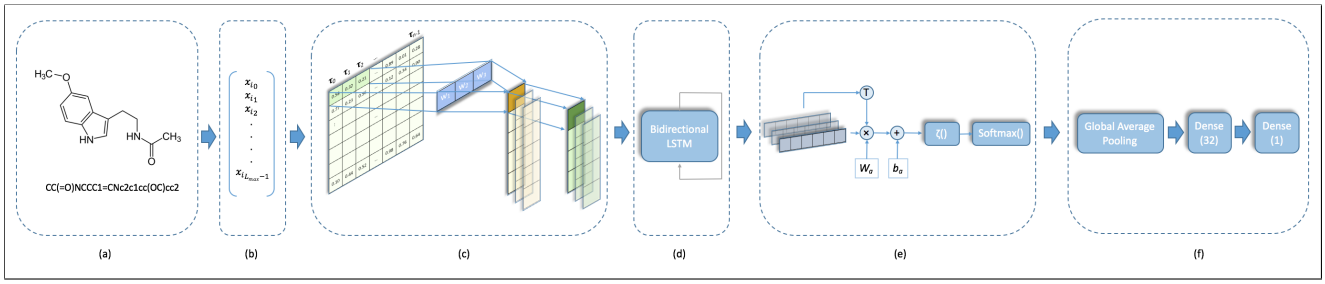

Fig. 1: Detailed architecture diagram of the proposed attention-driven LSTM model. The model has the following processing order: (a) sample molecule with its respective SMILES representation of an arbitrary size, (b) initial encoding to form the SMILES continuous-valued vector x_i, (c) the 32-dimensional character embedding layer producing matrix E, 1D convolutional layer, and 1D max pooling layer, (d) bidirectional LSTM layer, (e) self-attention layer with trainable weights W_a and b_a (f) global average pooling layer, dense layer with 32 nodes and leaky ReLU activation, and dense layer with single node and linear activation.

4.5 Model architecture

Our model architecture is shown in Figure 1. It's layer decomposition and respective hyperparameters are given in Appendix A. 1D-convolutional filters are applied on the embedding matrices to extract meaningful features, a technique also employed by *Paul et al's* SINet [3] and CheMixNet [20], followed by a Max Pooling layer to only retain relevant information extracted from the filters while simultaneously reducing the shape of the matrix read by the LSTM layer. A Bidirectional LSTM layer is subsequently used, followed by the self-attention layer. Afterwards, a global average pooling layer is used as a dimensionality reduction technique and reducing the number of trainable parameters (rather than flattening the previous tensor). Two dense layers follow the global average pooling layer, with leaky rectified linear unit (ReLU) and linear activation functions respectively.

4.6 Software

The presented network was implemented using Keras [60] and TensorFlow [61], while pre-processing steps were completed with Sci-Kit Learn. We note here that the TensorFlow CuDNN LSTM layer [62], a GPU-specific LSTM implementation to achieve maximum computational throughput, was used in our model to accelerate the training process. The NREL OPV dataset used in this study can be found in the original work [8].

5 Results & Discussion

5.1 Experimental Configuration

For the NREL OPV dataset, we use the train, validation, test sets provided by *St. John et al* [8], which contain ~$8.1E4/5E3/5E3$ molecules respectively. We performed a 90/5/5 stratified split of the Harvard CEP dataset (~$1.1E6/5.1E4/5.1E4$ molecules) to form the individual training, validation, and test sets respectively. In accordance with [36], a 80/10/10 stratified split was used for the ZINC-250k dataset.

We use mean squared error (MSE) as the loss function for our proposed attention model. And, in accordance with related works, we use the mean absolute error (MAE) as our evaluation metric. Models were trained using the Adam [63] optimizer with a starting learning rate of $1E-4$, using $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Callback functions were used during training to reduce the learning rate by a factor of 0.8 when the MAE of the validation set plateaued. All training was done on a NVIDIA GeForce RTX 2070 GPU, with 8GB of memory.

5.2 NREL OPV dataset prediction results

The NREL OPV test-set results on the B3LYP/6-31g(d) DFT computed molecules are shown in Table 1, showing the resultant MAEs of each property. Our networks' results are compared to leading (graph-based) message-passing neural networks (MPNN). The best results obtained from [8] between the single-input single-ouput (SISO) and single-input multiple-output (SIMO) MPNNs are included in the table. Furthermore, results from a MPNN adapted from *Jørgensen et al's* SchNet with edge updates [19], trained on DFT-optimized 3D

coordinates, are also included in the table. The results of our proposed attention-LSTM network are compared to each of these models.

DFT-computed properties on conformational isomers were used to determine an optimal or objective error rate. The size of the considered molecules induces different energy minimization routine convergences of lowest-energy states, which generates slightly different optoelectronic properties [8]. These convergence inconsistencies provided an acceptable range of values for each considered property, from which the conformer MAEs were computed. Since our model does not consider atomic spatial positioning, these conformer MAEs effectively served as the target error rate for our models and are included in the "Conf." column of Table 1.

B3LYP/6-31g(d)	Conf.	MPNN	SchNet	Our
Gap	*28.0*	35.4	32.7	**25.78**
HOMO	*22.0*	29.4	27.0	**22.97**
LUMO	*25.5*	27.9	24.8	**21.25**
Spectral Overlap	*81.3*	149.2	96.6	**96.42**
Polymer HOMO	*37.4*	47.4	56.9	**43.42**
Polymer LUMO	*45.0*	46.8	56.8	**42.91**
Polymer Gap	*46.3*	56.3	69.8	**51.66**
Pol. Optical LUMO	*42.6*	43.9	57.2	**41.72**

Table 1: **Results on NREL Dataset.** This table contains the MAEs for each property. The spectral overlap MAEs are provided in *W/mol*, whereas the other properties' MAEs are given in *meV*. The best scores between SISO and SIMO models [8] are shown in the "MPNN" column.

Separate attention-LSTM networks were trained on each property. Each attention-LSTM model was trained for approximately 50 epochs. For models trained on monomer properties, a maximum of 20 augmented molecules were generated for each original training sample, which constituted the training dataset. However, since there were far fewer training samples that contained polymer properties (around half of the original training set), a maximum of 50 augmented molecules were generated for each original training sample. During test-set evaluation, utilizing TTA, a maximum of 35 augmented molecules were used.

5.3 Harvard CEP dataset prediction results

Both our attention-LSTM model and the MPNN [8] were trained on the Harvard CEP dataset for 75 epochs. We employed a SIMO framework for the MPNN since it attained better results over individually trained SISO models [8]. We note here that our attention-LSTM model was *not* trained with augmented SMILES samples, nor did we employ TTA during evaluation. The Harvard CEP test-set results are shown in Table 2, which displays the MAEs for each property. Unlike the NREL OPV dataset, the Harvard CEP dataset did not provide any information on conformational isomers, hence a target or optimal error rate could not be established.

BP86/def2-SVP	MPNN	Our
Gap (meV)	12.52	**10.52**
HOMO (meV)	8.83	**6.71**
LUMO (meV)	9.32	**7.11**

Table 2: **Results on CEP Dataset**. This table contains the MAEs of each property.

5.4 ZINC-250k dataset prediction results

Our data augmentation technique was used on the training samples while also employing TTA. A maximum of 50 augmented samples were used for both the training and testing data. Our model was trained for approximately 20 epochs. Separate models were trained for each individual property. Our testset results on ZINC-250k are shown in Table 3. We compare our results to other contemporary models.

Model	logP	QED
ECFP [64]	0.38	0.045
CVAE [51]	0.15	0.054
CVAE ENC [51]	0.13	0.037
GraphConv [17]	0.05	0.017
All SMILES VAE [36]	0.005	0.0052
Our	**0.0042**	**0.0031**

Table 3: **Results on ZINC-250k Dataset.** This table contains the MAEs for each property.

5.5 Discussion

The attention-LSTM network showed immense improvement on the NREL OPV dataset compared to the graph networks. The attention-LSTM network not only significantly reduced the MAE for every property compared to contemporary models, but also scored within the optimal error range for the monomer Gap and LUMO as well as the polymer LUMO and Optical LUMO. Our model achieved a monomer Gap MAE of 25.78 meV and a monomer LUMO MAE of 21.25

meV, a percent decrease from the leading SchNet of 21.16% and 14.31% respectively; while achieving a polymer LUMO MAE of 42.91 meV and polymer Optical LUMO MAE of 41.72, a percent decrease from the leading MPNN of 8.31% and 4.97% respectively. The success of the attention network can not only be attributed to the inclusion of the attention-mechanism itself, but also the training augmentation technique used, as well as employing TTA during evaluation. Similar to its performance on the NREL OPV dataset, the attention network outperforms the MPNN on every property of the Harvard CEP dataset; achieving an average reduction among all properties of 21.23%. The results on the CEP dataset demonstrate that although network performance benefits from augmented training samples and TTA, it is not dependent on these methods. It is also noted here that the attention models were not pre-trained on any data beforehand, hence no transfer learning techniques were used to enhance results.

We further evaluated our regressor on the ZINC-250k dataset. Our model reduced the current state-of-the-art [36] logP and QED MAE by 16% and 40.39% respectively, hence making the attention-LSTM with TTA an auspicious model and augmentation routine for drug evaluation.

Although graph-based models have dominated recent studies on quantum mechanical and OPV predictive modeling, the feature generation can be impractical. The spatial information on which graph networks depend are not always available when searching for new materials [65], whereas textual descriptors ubiquitous and benefit from their simplicity and ease of generation.

However, generating the necessary 2D or 3D data for graph networks from textual data, using tools such as RDKit [66], is also more time consuming than using the textual features themselves. HTVS methods are time-sensitive operations which seek to minimize computation time for inferring molecular properties, since such methods operate on a large order of candidate compounds. Additional, time-intensive pre-processing steps, such as text to spatial coordinate calculations for graph network inputs, only hinder HTVS performance. Mitigation of such timely additional processing steps is ideal.

Using textual descriptors also allowed us to augment our data from a limited training set. The augmentation routine used was simple and computationally efficient. This is a useful data generation tactic for other size-limited datasets (such as the publicly available Quantum Machine datasets: QM7, QM8 and QM9), while not inducing severe overfitting.

The interpretability of the attention network is also more transparent compared to other proposed deep learning models. The attention layer enables the network to pinpoint constituents of the molecule which directly influence the prediction [26]; while the embedding layer displays learned relationships between SMILES tokens in a reduced (2D or 3D) vector space [65].

6 Conclusion

In this work, we focused on predicting properties of organic photovoltaic molecules, namely the HOMO, LUMO and Gap, which most directly impact the PCE. A novel attention-driven LSTM network was presented that is capable of predicting such optoelectronic properties by learning strictly from the SMILES representation of the molecule. This network was coupled with an effective data augmentation routine, which was utilized not only for generating new training samples, but also during the testset evaluation. The network was tested on two contemporary OPV datasets (NREL OPV and Harvard CEP) and was compared against leading (graph-based) message passing neural networks. Our attention-driven LSTM obtained better results than the graph networks and, for some properties, were within the conformational isomer-derived optimal error range for the NREL OPV dataset. We further demonstrated that our model is capable of generalizing well to cross-disciplinary tasks, specifically pharmaceutical drug design. Our model greatly reduced the leading VAE-based model's MAE across all considered targets on the ZINC-250k dataset.

A Network layer decomposition

The summary shown below details the sequential model architecture. "He normal" kernel initialization was used for each layer that required an initializer. The character embedding layer had a 32-dimensional output. A kernel size of 3 and the linear activation function was used for the Conv1D layer. A pool size of 2 was used for the MaxPooling layer. The bidirectional LSTM consisted of 1024 units. The self-attention layer consisted of 1024 units and used ReLU activation. The penultimate dense layer used leaky ReLU activation with $\alpha = 0.1$ and the final dense layer used linear activation. The total number of trainable parameters for this model is: $\sim 4.3e6$.

```
_________________________________________________________
Layer (type)        Output Shape              Param #
=========================================================
Input               (None, 270)               0
_________________________________________________________
Embedding           (None, 270, 32)           1184
_________________________________________________________
Conv-1d             (None, 270, 256)          24832
_________________________________________________________
```

```
Max-pool-1d          (None, 135, 256)          0
_________________________________________________________________
Bi-LSTM              (None, 135, 1024)         3153920
_________________________________________________________________
Self-attention       (None, 135, 1024)         1048577
_________________________________________________________________
Global-avg-pool-1d (None, 1024)                0
_________________________________________________________________
Dense                (None, 32)                32800
_________________________________________________________________
Dense                (None, 1)                 33
=================================================================
Total params: 4,261,346
Trainable params: 4,261,346
Non-trainable params: 0
```

References

1. Omar Abdulrazzaq, Viney Saini, Shawn Bourdo, Enkeleda Dervishi, and Alexandru Biris. Organic solar cells: A review of materials, limitations, and possibilities for improvement. *Particulate Science and Technology*, 31, 09 2013.
2. Stephen R. Forrest. The limits to organic photovoltaic cell efficiency. *MRS Bulletin*, 30(1):28–32, 2005.
3. Arindam Paul, Dipendra Jha, Reda Al-Bahrani, Wei keng Liao, Alok Choudhary, and Ankit Agrawal. Transfer learning using ensemble neural networks for organic solar cell screening, 2019.
4. A. Smets, K. Jäger, O. Isabella, R. van Swaaij, and M. Zeman. *Solar Energy: The Physics and Engineering of Photovoltaic Conversion, Technologies and Systems*. UIT Cambridge, 2016.
5. Gabriel Ravanhani Schleder, Antonio Claudio Padilha, Carlos Acosta, Marcio Costa, and Adalberto Fazzio. From dft to machine learning: recent approaches to materials science – a review. *Journal of Physics: Materials*, 02 2019.
6. Zhijian Liu, Hao Li, Kejun Liu, Hancheng Yu, and Kewei Cheng. Design of high-performance water-in-glass evacuated tube solar water heaters by a high-throughput screening based on machine learning: A combined modeling and experimental study. *Solar Energy*, 142:61 – 67, 2017.
7. Klaus Capelle. A bird's-eye view of density-functional theory, 2002.
8. Peter C. St. John, Caleb Phillips, Travis W. Kemper, A. Nolan Wilson, Yanfei Guan, Michael F. Crowley, Mark R. Nimlos, and Ross E. Larsen. Message-passing neural networks for high-throughput polymer screening. *The Journal of Chemical Physics*, 150(23):234111, Jun 2019.
9. Mine Kaya and Shima Hajimirza. Application of artificial neural network for accelerated optimization of ultra thin organic solar cells. *Solar Energy*, 165:159 – 166, 2018.
10. Felix A. Faber, Luke Hutchison, Bing Huang, Justin Gilmer, Samuel S. Schoenholz, George E. Dahl, Oriol Vinyals, Steven Kearnes, Patrick F. Riley, and O. Anatole von Lilienfeld. Machine learning prediction errors better than dft accuracy, 2017.
11. Dipendra Jha, Logan Ward, Arindam Paul, Wei-keng Liao, Alok Choudhary, Chris Wolverton, and Ankit Agrawal. Elemnet: Deep learning the chemistry of materials from only elemental composition. *Scientific Reports*, 8, 12 2018.
12. Dipendra Jha, Kamal Choudhary, Francesca Tavazza, Wei-keng Liao, Alok Choudhary, Carelyn Campbell, and Ankit Agrawal. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nature Communications*, 10, 12 2019.
13. Edward Pyzer-Knapp, Kewei Li, and Alán Aspuru-Guzik. Learning from the harvard clean energy project: The use of neural networks to accelerate materials discovery. *Advanced Functional Materials*, 25, 09 2015.
14. Iman Sajedian, Heon Lee, and Junsuk Rho. Design of high transmission color filters for solar cells directed by deep q-learning. *Solar Energy*, 195:670 – 676, 2020.
15. Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics*, 15(9):095003, Sep 2013.
16. Stéphanie Valleau, Florian Häse, Edward Pyzer-Knapp, and Alán Aspuru-Guzik. Machine learning exciton dynamics. *Chemical Science*, 7, 04 2016.
17. David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints, 2015.
18. Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry, 2017.
19. Peter Bjørn Jørgensen, Karsten Wedel Jacobsen, and Mikkel N. Schmidt. Neural message passing with edge updates for predicting properties of molecules and materials, 2018.
20. Arindam Paul, Dipendra Jha, Reda Al-Bahrani, Wei keng Liao, Alok Choudhary, and Ankit Agrawal. Chemixnet: Mixed dnn architectures for predicting chemical properties using multiple molecular representations, 2018.
21. David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28:31–36, 02 1988.
22. Stephen Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi, and Igor Pletnev. Inchi - the worldwide chemical structure identifier standard. *Journal of cheminformatics*, 5:7, 01 2013.
23. Garrett B. Goh, Nathan Hodas, Charles Siegel, and Abhinav Vishnu. Smiles2vec: Predicting chemical properties from text representations, 2018.
24. Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading, 2016.
25. Shuangjia Zheng, Xin Yan, Yuedong Yang, and Jun Xu. Identifying structure-property relationships through smiles syntax analysis with self-attention mechanism, 11 2018.
26. Guillaume Lambard and Ekaterina Gracheva. Smiles-x: autonomous molecular compounds characterization for small datasets without descriptors, 2019.
27. Bonggun Shin, Sungsoo Park, Keunsoo Kang, and Joyce C. Ho. Self-attention based molecule representation for predicting drug-target interaction, 2019.
28. Esben Jannik Bjerrum. Smiles enumeration as data augmentation for neural network modeling of molecules, 2017.

29. Gabriel Lima Guimaraes, Benjamin Sanchez-Lengeling, Carlos Outeiral, Pedro Luis Cunha Farias, and Alán Aspuru-Guzik. Objective-reinforced generative adversarial networks (organ) for sequence generation models, 2017.
30. Benjamin Sanchez-Lengeling, Carlos Outeiral, Gabriel L. Guimaraes, and Alan Aspuru-Guzik. Optimizing distributions over molecular space. an objective-reinforced generative adversarial network for inverse-design chemistry (organic), Aug 2017.
31. Johannes Hachmann, Roberto Olivares-Amaya, Sule Atahan-Evrenk, Carlos Amador-Bedolla, Roel Sanchez-Carrera, Aryeh Gold-Parker, Leslie Vogt, Anna Brockway, and Alán Aspuru-Guzik. The harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid. *The Journal of Physical Chemistry Letters*, 2, 08 2011.
32. Emiliangelo Ratti and David Trist. Continuing evolution of the drug discovery process in the pharmaceutical industry. *Farmaco (Società chimica italiana : 1989)*, 56:13–9, 03 2001.
33. Łukasz Maziarka, Agnieszka Pocha, Jan Kaczmarczyk, Krzysztof Rataj, Tomasz Danel, and Michał Warchoł. Mol-cyclegan: a generative model for molecular optimization. *Journal of Cheminformatics*, 12(1), Jan 2020.
34. Teague Sterling and John Irwin. Zinc 15 - ligand discovery for everyone. *Journal of chemical information and modeling*, 55, 10 2015.
35. Richard Bickerton, Gaia Paolini, Jérémy Besnard, Sorel Muresan, and Andrew Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4:90–8, 02 2012.
36. Zaccary Alperstein, Artem Cherkasov, and Jason Tyler Rolfe. All smiles variational autoencoder, 2019.
37. Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547, 2018.
38. Adam C Mater and Michelle L Coote. Deep learning in chemistry. *Journal of chemical information and modeling*, 2019.
39. Harikrishna Sahu, Weining Rao, Alessandro Troisi, and Haibo Ma. Toward predicting efficiency of organic solar cells via machine learning and improved descriptors. *Advanced Energy Materials*, 8(24):1801032, 2018.
40. Bing Cao, Lawrence A Adutwum, Anton O Oliynyk, Erik J Luber, Brian C Olsen, Arthur Mar, and Jillian M Buriak. How to optimize materials and devices via design of experiments and machine learning: Demonstration using organic photovoltaics. *ACS nano*, 12(8):7434–7444, 2018.
41. Zijiang Yang, Yuksel C Yabansu, Dipendra Jha, Weikeng Liao, Alok N Choudhary, Surya R Kalidindi, and Ankit Agrawal. Establishing structure-property localization linkages for elastic deformation of three-dimensional high contrast composites using deep learning approaches. *Acta Materialia*, 166:335–345, 2019.
42. A. Paul, M. Mozaffar, Z. Yang, W. Liao, A. Choudhary, J. Cao, and A. Agrawal. A real-time iterative machine learning approach for temperature profile prediction in additive manufacturing processes. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 541–550, 2019.
43. Peter Bjørn Jørgensen, Murat Mesta, Suranjan Shil, Juan Maria García Lastra, Karsten Wedel Jacobsen, Kristian Sommer Thygesen, and Mikkel N Schmidt. Machine learning-based screening of complex molecules for polymer solar cells. *The Journal of chemical physics*, 148(24):241735, 2018.
44. Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.
45. Min-Hsuan Lee. Robust random forest based non-fullerene organic solar cells efficiency prediction. *Organic Electronics*, 76:105465, 2020.
46. Arindam Paul, Alona Furmanchuk, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. Property prediction of organic donor molecules for photovoltaic applications using extremely randomized trees. *Molecular informatics*, 38(11-12):1900038, 2019.
47. Peter B Jørgensen, Mikkel N Schmidt, and Ole Winther. Deep generative models for molecular science. *Molecular informatics*, 37(1-2):1700133, 2018.
48. Steven Lopez, Edward Pyzer-Knapp, Gregor Simm, Trevor Lutzow, Kewei Li, Laszlo Seress, Johannes Hachmann, and Alán Aspuru-Guzik. The harvard organic photovoltaic dataset. *Scientific Data*, 3, 09 2016.
49. Markus Scharber, D. Mühlbacher, M. Koppe, Patrick Denk, Ch Waldauf, A.J. Heeger, and Christoph Brabec. Design rules for donors in bulk-heterojunction solar cells—towards 10 *Advanced Materials*, 18:789 – 794, 02 2006.
50. Pavel Schilinsky, Christoph Waldauf, and Christoph J. Brabec. Recombination and loss analysis in polythiophene based bulk heterojunction photodetectors. *Applied Physics Letters*, 81(20):3885–3887, 2002.
51. Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, Jan 2018.
52. Craig A James. Opensmiles specification, May 2016.
53. Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks, 2015.
54. Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, Apr 2019.
55. Mike Schuster and Kuldip Paliwal. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45:2673 – 2681, 12 1997.
56. Garrett B. Goh, Nathan O. Hodas, Charles Siegel, and Abhinav Vishnu. Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties, 2017.
57. Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches, 2014.
58. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014.
59. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.
60. François Chollet et al. Keras. `https://keras.io`, 2015.
61. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geof-

frey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems, 2016.
62. Jeremy Appleyard, Tomas Kocisky, and Phil Blunsom. Optimizing performance of recurrent neural networks on gpus, 2016.
63. Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.
64. David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50:742–54, 05 2010.
65. Peter Jørgensen, Murat Mesta, Suranjan Shil, Juan María García-Lastra, Karsten Jacobsen, Kristian Thygesen, and Mikkel Schmidt. Machine learning-based screening of complex molecules for polymer solar cells. *The Journal of Chemical Physics*, 148:241735, 06 2018.
66. Greg Landrum. Rdkit: Open-source cheminformatics software. 2016.