Solar Energy 224 (2021) 43–50

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

An attention-driven long short-term memory network for high throughput virtual screening of organic photovoltaic candidate molecules

Ryan J. Richards [a,*], Arindam Paul [b]

[a] University of Pennsylvania, United States
[b] Independent Researcher

ARTICLE INFO

Keywords:
Machine learning
Organic photovoltaics
Recurrent neural network
High throughput virtual screening
Attention
Drug discovery

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 overall 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.

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 have been immensely improved over the past few decades (Arent et al., 2011). The development of photovoltaic multifunctional devices have increased the utility of solar cells (Han et al., 2016). Broadly, there are two kinds of solar cells: inorganic (such as conventional silicon-based) and organic (based on thin-film polymers or small molecules). Inorganic cells possess superior performance in terms of power conversion efficiency (PCE). Recent studies have demonstrated that design of optoelectronic devices can further improve device efficiency for inorganic solar cells (Lu et al., 2020; Cai et al., 2019; Chen et al., 2019). However, inorganic solar cells suffer from complicated, expensive fabrication processes and structural rigidity (Abdulrazzaq et al., 2013). Organic Photovoltaic (OPV) cells, based on thin-film polymers or small molecules, offer simple and cost-efficient fabrication processes and novel applications, but lack a high enough PCE for commercialization (Abdulrazzaq et al., 2013). Low PCE and poor ambient stability are the primary barriers to widespread implementation of OPVs.

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 (Forrest, 2005; Paul et al., 2019d). OPV cell design seeks to maximize the PCE or the percentage of electricity that is generated from the absorption of photons. The PCE depends on specific optoelectronic properties of donor and acceptor 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 (Smets et al., 2019). The ability to rapidly and accurately predict these important properties and hence avoid a costly and time-intensive

[*] Corresponding author.
E-mail address: ryry@seas.upenn.edu (R.J. Richards).

https://doi.org/10.1016/j.solener.2021.05.064
Received 25 May 2020; Received in revised form 8 May 2021; Accepted 17 May 2021

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 (Schleder et al., 2019; Liu et al., 2017). The most prevalent and accurate methodology is Density Functional Theory (DFT), a computational quantum mechanics modeling routine which determines molecular properties using functionals of the electron density (Capelle, 2002). DFT possesses chemically accurate computations but is severely bottlenecked by its processing time, especially when coupled with the demanding requirements of HTVS (Peter et al., 2019).

Machine Learning (ML), a field of statistical learning, has directed materials research into a new data-driven science paradigm (Schleder et al., 2019; Kaya and Hajimirza, 2018). ML has the potential to match the chemical accuracy of DFT while significantly decreasing the processing (inference) time (Faber et al., 2018; Jha et al., 2018; Jha et al., 2019); ML algorithm prediction times (operating at $\mathcal{O}(10^{-3}\ \mathrm{s})$) are nearly six orders of magnitude faster than DFT calculations ($\mathcal{O}(10^{3}\ \mathrm{s})$ on 30 heavy-atom molecules) (Peter et al., 2019).

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 (Pyzer-Knapp et al., 2015; Sajedian et al., 2020). Several descriptors have been utilized in ML frameworks for electronic properties prediction (Schleder et al., 2019), including Coulomb matrices (Montavon et al., 2013; Valleau et al., 2016), molecular strings or graphs (Duvenaud et al., 2015; Gilmer et al., 2017; Jørgensen et al., 2018a; Peter et al., 2019), and molecular fingerprinting (Paul et al., 2017; Pyzer-Knapp et al., 2015).

Among these approaches are natural language processing (NLP) derived techniques which depend on textual representations (Weininger, 1988; Heller et al., 2013) of molecular structures rather than relying on 2D or 3D-defined structures (i.e. spatial coordinates) (Paul et al., 2017; Goh et al., 2018; Paul et al., 2019d). Novel components used extensively in NLP, such as attention mechanisms (Cheng et al., 2016), have shown great promise in the analysis of molecular structures (Zheng et al., 2018; Lambard and Gracheva, 2019; Shin et al., 2019). Line notations also permit the usage of augmentation techniques that are easily realized and computationally efficient (Bjerrum, 2017; Lambard and Gracheva, 2019), which greatly improve network performance especially when coupled with attention-mechanisms. Furthermore, NLP techniques allow for deeper analysis of the molecular structure. Multidimensional 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 (Lambard and Gracheva, 2019). Attention enhances these analytical capabilities by narrowing down specific components of a molecule that most heavily resonate with target properties (Lambard and Gracheva, 2019), 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 (Guimaraes et al., 2017; Sanchez-Lengeling et al., 2017).

In this work, we create an attention-driven LSTM network with 1D convolutions to predict optoelectronic properties of OPV candidate molecules from the CEP (Hachmann et al., 2011) and NREL OPV (Peter et al., 2019) 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 that achieve better results than graph-based models for the considered OPV datasets.

Furthermore, although this study intends 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 (Ratti and Trist, 2001; Maziarka et al., 2020b). We use the ZINC-250k dataset, which contains 250k drug-like molecules extracted from the ZINC database (Sterling and Irwin, 2015), to predict the log octanol-water partition coefficient (logP) and the quantitative estimate of drug-likeness (QED) (Bickerton et al., 2012). We show that our attention-LSTM model provides better results than leading state-of-the-art Variational Autoencoder (VAE) based models (Alperstein et al., 2019).

2. Related works

Machine learning models have been successfully applied in materials and molecular design (Paul et al., 2019a; Butler et al., 2018; Mater and Coote, 2019; Paul et al., 2019b; Sahu et al., 2018; Cao et al., 2018; Yang et al., 2020; Playe and Stoven, 2020; Jørgensen et al., 2018b) 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 (Geurts et al., 2006) were developed for screening organic monomers used for photovoltaic applications and predicting organic solar cell efficiency (Lee, 2020; Paul et al., 2019c).

Jørgensen et al. (2018c) used a VAE with predefined SMILES syntax (grammatical) rules for predicting molecular properties and generating new molecules with desirable properties. The All SMILES VAE (Alperstein et al., 2019) significantly improved the results from (Jørgensen et al., 2018c) 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 attentional 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) (Alperstein et al., 2019).

Paul et al. (2019d) 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 (Lopez et al., 2016). Munshi et al. (2021) utilized transfer learning based LSTM on SMILES notations to generate novel designer chemistries for polymer devices for OPV applications. The generative model retrained on a small OPV target set predicts new polymer repeat units with potentially high PCE. Lee et al. (2020) harnessed transfer learning using graph neural networks for predicting opto-electronic properties of conjugated polymers on a very small training set.

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 (Scharber et al., 2006) was used to focus on salient optoelectronic properties that most heavily influence η. We make the same initial assumptions as Scharber et al. (2006) 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.

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

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 (Schilinsky et al., 2002). This limiting factor was investigated (Scharber et al., 2006) 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 predominant 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.

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

While it is important to optimize the donor HOMO energy, which dictates V_{oc}, the consequential change in bandgap energy must also be considered, which influences $\widetilde{J}_{sc,Sch}$. Fixing E_{HOMO}^{Donor} at an optimal level to maximize V_{oc} means that the donor LUMO level (E_{LUMO}^{Donor}) must be modified to achieve significant bandgap energy. Scharber et al. indeed note that the PCE is more sensitive to changes in donor LUMO energy rather than strictly its bandgap (Scharber et al., 2006). 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) (Scharber et al., 2006). 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 that 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 NREL OPV (Peter et al., 2019) and CEP (Hachmann et al., 2011; Pyzer-Knapp et al., 2015). Developed in 2019, the NREL OPV dataset contains 9.1*E*4 molecules with DFT-computed optoelectronic calculations specifically for OPV applications. NREL populated the dataset with relatively larger molecules ($\leqslant$201 atoms) when compared to other similar datasets such as QM9 ($\leqslant$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 (Hachmann et al., 2011) using a custom version of the Q-Chem 3.2 package (Shao et al., 2006). 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 (Hachmann et al., 2011). The dataset portion employed in this work contains $\sim 1.2E6$ candidate donor molecules. The optoelectronic properties were computed using the BP86/def2-SVP DFT functional/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 (Sterling and Irwin, 2015). ZINC-250k contains 2.5*E*5 drug-like commercially available organic molecules with $\leqslant$38 heavy atoms. Following related works (Alperstein et al., 2019; Gómez-Bombarelli et al., 2018), we focus on predicting the log octanol-water partition coefficient (logP) and quantitative estimate of drug-likeness (QED) (Bickerton et al., 2012).

4.2. Pre-processing and SMILES encoding

All molecules in the considered datasets are given in the Simplified Molecular Input Line Entry System (SMILES) (James, 2016) 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 the 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 (Gal and Ghahramani, 2015). 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 Fig. 1 (a)–(b).

A character embedding layer (Gal and Ghahramani, 2015) was used to learn a mapping between the initial continuous SMILES vectors, Fig. 1 (b), to a 32-dimensional vector space, shown in Fig. 1 (c); more information is provided on this specific implementation in Section 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 (Lambard and Gracheva, 2019).

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.

$$argmin_{\theta} \sum_i L\left(f\left(\mathbf{x}_i : \theta\right), \mathbf{y}_i\right) \tag{4}$$

4.3. Augmentation methods

Data augmentation techniques were used to better train the network on the NREL dataset. Bjerrum (2017) 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 that 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 (Lambard and Gracheva, 2019). We designate the

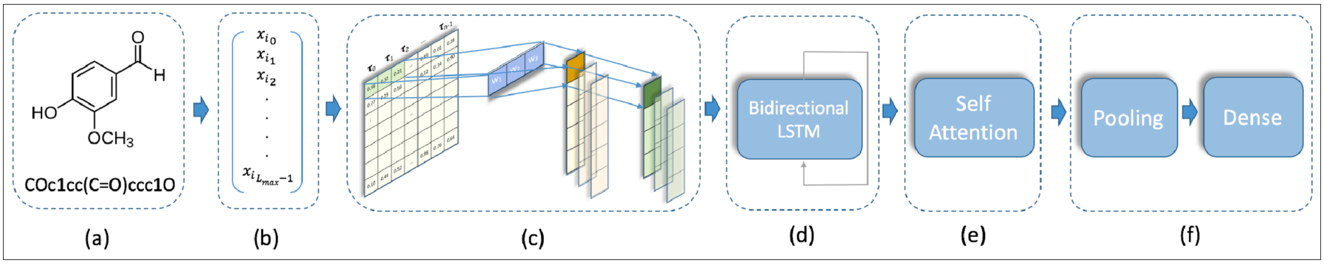

Fig. 1. Detailed architecture diagram of the proposed attention-driven LSTM model. The model has the following processing order: (a) sample molecule with its respective SMILES representation of an arbitrary size, (b) initial encoding to form the SMILES continuous-valued vector x_i, (c) the 32-dimensional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

augmented samples as $\widehat{\mathbf{x}}_i$.

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 ($\widehat{\mathbf{y}}_i$). 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 $(\widehat{\mathbf{x}}_i, \widehat{\mathbf{y}}_i)$

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

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 (Peter et al., 2019). Hence, the regression problem is simplified to Eq. 6, where the loss is minimized between the network outputs given the augmented samples ($f(\widehat{\mathbf{x}}_i)$) and the noisy targets ($\widehat{\mathbf{y}}_i$).

$$argmin_\theta \sum_i L\left(f\left(\widehat{\mathbf{x}}_i : \theta\right), \widehat{\mathbf{y}}_i\right) \tag{6}$$

This augmentation technique is also exploited for network evaluation, a method of testing that has gained traction in the imaging community (Wang et al., 2019) 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.

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 (Peter et al., 2019) are shown in the "MPNN" column.

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**

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: $L_{max} \times 32$.

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

We utilize an LSTM layer (Cheng et al., 2016) 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 (Schuster and Paliwal, 1997) to fully capture contextual details. Bidirectional LSTM (BiLSTM) based models involving character embeddings have demonstrated superb performance in works involving SMILES analysis (Goh et al., 2017) (Lambard and Gracheva, 2019) (Zheng et al., 2018).

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

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

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

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.

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

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

Network performance is enhanced by appending an Attention layer to the BiLSTM. The Attention mechanism has several variants and has 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. (2014) 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. Dzmitry et al. (2014) 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. (2016) 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 (Vaswani et al., 2014).

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 (Lambard and Gracheva, 2019; Zheng et al., 2018), 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).

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

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

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

4.5. Model architecture

Our model architecture is shown in Fig. 1. Its 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 (Paul et al., 2019d) and CheMixNet (Paul et al., 2017), 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. Afterward, 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 a leaky rectified linear unit (ReLU) and linear activation functions respectively.

4.6. Software

The presented network was implemented using Keras (Chollet et al., 2015) and TensorFlow (Abadi et al., 2016), while pre-processing steps were completed with Sci-Kit Learn. We note here that the TensorFlow CuDNN LSTM layer (Appleyard et al., 2016), 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 (Peter et al., 2019).

5. Results & discussion

5.1. Experimental configuration

For the NREL OPV dataset, we use the train, validation, test sets provided by Peter et al. (2019), which contain $\sim 8.1E4/5E3/5E3$ molecules respectively. We performed a 90/5/5 stratified split of the Harvard CEP dataset ($\sim 1.1E6/5.1E4/5.1E4$ molecules) to form the individual training, validation, and test sets respectively. In accordance with (Alperstein et al., 2019), 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, following related works, we use the mean absolute error (MAE) as our evaluation metric. Models were trained using the Adam (Kingma and Ba, 2014) 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 an NVIDIA GeForce RTX 2070 GPU, with 8 GB 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 (Peter et al., 2019) between the single-input single-output (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 (Jørgensen et al., 2018a), 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 (Peter et al., 2019). 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.

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 (Peter et al., 2019) 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 (Peter et al., 2019). 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.

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 property. Our test set results on ZINC-250k are shown in Table 3. We compare our results to other contemporary models.

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

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

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

Model	logP	QED
ECFP (Rogers and Hahn, 2010)	0.38	0.045
CVAE (Gómez-Bombarelli et al., 2018)	0.15	0.054
CVAE ENC (Gómez-Bombarelli et al., 2018)	0.13	0.037
GraphConv (Duvenaud et al., 2015)	0.05	0.017
All SMILES VAE (Alperstein et al., 2019)	0.005	0.0052
Our	**0.0042**	**0.0031**

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 (Alperstein et al., 2019) 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 (Jørgensen et al., 2018d), 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 (Landrum, 2016), is also more time consuming than using the textual features themselves. HTVS methods are time-sensitive operations that 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 (QM) 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 (Lambard and Gracheva, 2019); while the embedding layer displays learned relationships between SMILES tokens in a reduced (2D or 3D) vector space (Jørgensen et al., 2018d).

The potential of this research extends beyond accelerating HTVS and improving interpretability. Research has been widely conducted (Jin et al., 2019; Maziarka et al., 2020a; Zhou et al., 2019; Winter et al., 2019; Yan et al., 2020; Popova et al., 2019; Shi et al., 2020; Zang and Wang, 2020; Korovina et al., 2019) investigating the use of VAEs, Generative Adversarial Networks (GANs), Reinforcement Learning (RL), and other generative networks to produce molecules with desirable properties. These networks generally include some form of regressor or discriminator portion, which attempts to predict the target property from either the input structure or a lower-dimensional representation. Works including VAEs generally depend on some post-fitting optimization routine which navigates a learned latent space and is penalized on incorrectly formed geometries or SMILES syntax. Since our attention-LSTM achieved low MAEs for each target property across chemically diverse datasets, it can be a suitable replacement for a more simple fully connected regressor or Sparse Gaussian Process (SGP) (Yan et al., 2020; Jin et al., 2019), which can aid generative models in producing more suitable and accurate candidate molecules. Post-optimization evaluation of the learned and explored latent space can reveal salient features, geometries, and other characteristics of molecules. Applying such a routine to solar cell candidate molecules has the potential to rapidly advance exploration and optimization of the cells.

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 test set 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. Further, with its low MAE across several chemically diverse datasets, this network can be coupled with a generative network to produce more fit candidate molecules with desirable properties. Although we mostly focused on predicting orbital energies in this work because of data constraints, we encourage researchers with more diverse datasets to expand this work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix 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

Abadi, Martín, Agarwal, Ashish, Barham, Paul, Brevdo, Eugene, Chen, Zhifeng, Citro, Craig, Corrado, Greg S., Davis, Andy, Dean, Jeffrey, Devin, Matthieu, Ghemawat, Sanjay, Goodfellow, Ian, Harp, Andrew, Irving, Geoffrey, Isard, Michael, Jia, Yangqing, Jozefowicz, Rafal, Kaiser, Lukasz, Kudlur, Manjunath, Levenberg, Josh, Mane, Dan, Monga, Rajat, Moore, Sherry, Murray, Derek, Olah, Chris, Schuster, Mike, Shlens, Jonathon, Steiner, Benoit, Sutskever, Ilya, Talwar, Kunal, Tucker, Paul, Vanhoucke, Vincent, Vasudevan, Vijay, Viegas, Fernanda, Vinyals, Oriol, Warden, Pete, Wattenberg, Martin, Wicke, Martin, Yu, Yuan, Zheng, Xiaoqiang, 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems.

Abdulrazzaq, Omar, Saini, Viney, Bourdo, Shawn, Dervishi, Enkeleda, Biris, Alexandru, 2013. Organic solar cells: A review of materials, limitations, and possibilities for improvement. Part. Sci. Technol. 31, 09.

Alperstein, Zaccary, Cherkasov, Artem, Rolfe, Jason Tyler, 2019. All smiles variational autoencoder.

Appleyard, Jeremy, Kocisky, Tomas, Blunsom, Phil, 2016. Optimizing performance of recurrent neural networks on gpus.

Arent, Douglas J, Wise, Alison, Gelman, Rachel, 2011. The status and prospects of renewable energy for combating global warming. Energy Econ. 33 (4), 584–593.

Bickerton, Richard, Paolini, Gaia, Besnard, Jérémy, Muresan, Sorel, Hopkins, Andrew, 2012. Quantifying the chemical beauty of drugs. Nat. Chem. 4, 90–98.

Bjerrum, Esben Jannik, 2017. Smiles enumeration as data augmentation for neural network modeling of molecules.

Butler, Keith T., Davies, Daniel W., Cartwright, Hugh, Isayev, Olexandr, Walsh, Aron, 2018. Machine learning for molecular and materials science. Nature 559 (7715), 547.

Cai, Xinyong, Chen, Yuanzheng, Sun, Bai, Chen, Jiao, Wang, Hongyan, Ni, Yuxiang, Tao, Li, Wang, Hui, Zhu, Shouhui, Li, Xiumei, et al., 2019. Two-dimensional blue-asp monolayers with tunable direct band gap and ultrahigh carrier mobility show promising high-performance photovoltaic properties. Nanoscale 11 (17), 8260–8269.

Cao, Bing, Adutwum, Lawrence A., Oliynyk, Anton O., Luber, Erik J., Olsen, Brian C., Mar, Arthur, Buriak, Jillian M., 2018. How to optimize materials and devices via design of experiments and machine learning: Demonstration using organic photovoltaics. ACS Nano 12 (8), 7434–7444.

Capelle, Klaus, 2002. A bird's-eye view of density-functional theory.

Chen, Yuanzheng, Lao, Zebin, Sun, Bai, Feng, Xiaolei, Redfern, Simon A.T., Liu, Hanyu, Lv, Jian, Wang, Hongyan, Chen, Zhongfang, 2019. Identifying the ground-state np sheet through a global structure search in two-dimensional space and its promising high-efficiency photovoltaic properties. ACS Mater. Lett. 1 (3), 375–382.

Cheng, Jianpeng, Dong, Li, Lapata, Mirella, 2016. Long short-term memory-networks for machine reading.

Cho, Kyunghyun, van Merrienboer, Bart, Bahdanau, Dzmitry, Bengio, Yoshua, 2014. On the properties of neural machine translation: Encoder-decoder approaches.

Chollet, François et al., 2015. Keras. https://keras.io.

Duvenaud, David, Maclaurin, Dougal, Aguilera-Iparraguirre, Jorge, Gómez-Bombarelli, Rafael, Hirzel, Timothy, Aspuru-Guzik, Alán, Adams, Ryan P., 2015. Convolutional networks on graphs for learning molecular fingerprints.

Bahdanau, Dzmitry, Cho, Kyunghyun, Bengio, Yoshua, 2014. Neural machine translation by jointly learning to align and translate.

Faber, Felix A., Hutchison, Luke, Huang, Bing, Gilmer, Justin, Schoenholz, Samuel S., Dahl, George E., Vinyals, Oriol, Kearnes, Steven, Riley, Patrick F., Anatole von Lilienfeld, O., 2017. Machine learning prediction errors better than dft accuracy.

Forrest, Stephen R., 2005. The limits to organic photovoltaic cell efficiency. MRS Bull. 30 (1), 28–32.

Gal, Yarin, Ghahramani, Zoubin, 2015. A theoretically grounded application of dropout in recurrent neural networks.

Geurts, Pierre, Ernst, Damien, Wehenkel, Louis, 2006. Extremely randomized trees. Mach. Learn. 63 (1), 3–42.

Gilmer, Justin, Schoenholz, Samuel S., Riley, Patrick F., Vinyals, Oriol, Dahl, George E., 2017. Neural message passing for quantum chemistry.

Goh, Garrett B., Hodas, Nathan O., Siegel, Charles, Vishnu, Abhinav, 2017. Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties.

Goh, Garrett B., Hodas, Nathan, Siegel, Charles, Vishnu, Abhinav, 2018. Smiles2vec: Predicting chemical properties from text representations.

Gómez-Bombarelli, Rafael, Wei, Jennifer N., Duvenaud, David, Hernández-Lobato, José Miguel, Sánchez-Lengeling, Benjamín, Sheberla, Dennis, Aguilera-Iparraguirre, Jorge, Hirzel, Timothy D., Adams, Ryan P., Aspuru-Guzik, Alán, 2018. Automatic chemical design using a data-driven continuous representation of molecules. ACS Central Sci. 4 (2), 268–276.

Guimaraes, Gabriel Lima, Sanchez-Lengeling, Benjamin, Outeiral, Carlos, Cunha Farias, Pedro Luis, Aspuru-Guzik, Alán, 2017. Objective-reinforced generative adversarial networks (organ) for sequence generation models.

Hachmann, Johannes, Olivares-Amaya, Roberto, Atahan-Evrenk, Sule, Amador-Bedolla, Carlos, Sánchez-Carrera, Roel, Gold-Parker, Aryeh, Vogt, Leslie, Brockway, Anna, Aspuru-Guzik, Alán, 2011. The harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid. J. Phys. Chem. Lett. 2, 08.

Han, Pengde, Sun, Bai, Cheng, Sen, Fangli, Yu., Jiao, Baoxiang, Qisheng, Wu., 2016. An optoelectronic resistive switching memory behavior of ag/α-snwo4/fto device. J. Alloy. Compd. 681, 516–521.

Heller, Stephen, McNaught, Alan, Stein, Stephen, Tchekhovskoi, Dmitrii, Pletnev, Igor, 2013. Inchi - the worldwide chemical structure identifier standard. J. Cheminformatics 5 (7).

James, Craig A., 2016. Opensmiles specification.

Jha, Dipendra, Ward, Logan, Paul, Arindam, Liao, Wei-keng, Choudhary, Alok, Wolverton, Chris, Agrawal, Ankit, 2018. Elemnet: Deep learning the chemistry of materials from only elemental composition. Sci. Rep. 8, 12.

Jha, Dipendra, Choudhary, Kamal, Tavazza, Francesca, Liao, Wei-keng, Choudhary, Alok, Campbell, Carelyn, Agrawal, Ankit, 2019. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. Nat. Commun. 10, 12.

Jin, Wengong, Barzilay, Regina, Jaakkola, Tommi, 2019. Junction tree variational autoencoder for molecular graph generation.

Jørgensen, Peter Bjørn, Jacobsen, Karsten Wedel, Schmidt, Mikkel N., 2018. Neural message passing with edge updates for predicting properties of molecules and materials.

Jørgensen, Peter Bjørn, Mesta, Murat, Shil, Suranjan, Lastra, Juan Maria García, Jacobsen, Karsten Wedel, Thygesen, Kristian Sommer, Schmidt, Mikkel N., 2018b. Machine learning-based screening of complex molecules for polymer solar cells. J. Chem. Phys. 148 (24), 241735.

Jørgensen, Peter B., Schmidt, Mikkel N., Winther, Ole, 2018c. Deep generative models for molecular science. Mol. Informat. 37 (1–2), 1700133.

Jørgensen, Peter, Mesta, Murat, Shil, Suranjan, García-Lastra, Juan María, Jacobsen, Karsten, Thygesen, Kristian, Schmidt, Mikkel, 2018d. Machine learning-based screening of complex molecules for polymer solar cells. J. Chem. Phys. 148, 241735.

Kaya, Mine, Hajimirza, Shima, 2018. Application of artificial neural network for accelerated optimization of ultra thin organic solar cells. Sol. Energy 165, 159–166.

Kingma, Diederik P., Ba, Jimmy, 2014. Adam: A method for stochastic optimization.

Korovina, Ksenia, Xu, Sailun, Kandasamy, Kirthevasan, Neiswanger, Willie, Poczos, Barnabas, Schneider, Jeff, Xing, Eric P., 2019. ChemBO: Bayesian optimization of small organic molecules with synthesizable recommendations. version: 2.

Lambard, Guillaume, Gracheva, Ekaterina, 2019. Smiles-x: autonomous molecular compounds characterization for small datasets without descriptors.

Landrum, Greg, 2016. Rdkit: Open-source cheminformatics software.

Lee, Min-Hsuan, 2020. Robust random forest based non-fullerene organic solar cells efficiency prediction. Org. Electron. 76, 105465.

Lee, Chee Kong, Lu, Chengqiang, Yu, Yue, Sun, Qiming, Hsieh, Chang-Yu, Zhang, Shengyu, Liu, Qi, Shi, Liang, 2020. Transfer learning with graph neural networks for optoelectronic properties of conjugated oligomers.

Liu, Zhijian, Li, Hao, Liu, Kejun, Hancheng, Yu., Cheng, Kewei, 2017. 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. Sol. Energy 142, 61–67.

Lopez, Steven, Pyzer-Knapp, Edward, Simm, Gregor, Lutzow, Trevor, Li, Kewei, Seress, Laszlo, Hachmann, Johannes, Aspuru-Guzik, Alán, 2016. The harvard organic photovoltaic dataset. Sci. Data 3, 09.

Lu, Ziyang, Neupane, Guru Prakash, Jia, Guohua, Zhao, Haitao, Qi, Dongchen, Du, Yaping, Lu, Yuerui, Yin, Zongyou, 2020. 2d materials based on main group element compounds: phases, synthesis, characterization, and applications. Adv. Funct. Mater. 30 (40), 2001127.

Mater, Adam C., Coote, Michelle L., 2019. Deep learning in chemistry. J. Chem. Inform. Model.

Maziarka, Lukas, Pocha, Agnieszka, Kaczmarczyk, Jan, Rataj, Krzysztof, Danel, Tomasz, Warchol, Michal, 2020. Mol-CycleGAN: a generative model for molecular optimization 12(1), 2.

Maziarka, Lukasz, Pocha, Agnieszka, Kaczmarczyk, Jan, Rataj, Krzysztof, Danel, Tomasz, Warchol, Michal, 2020b. Mol-cyclegan: a generative model for molecular optimization. J. Cheminformatics 12 (1).

Montavon, Grégoire, Rupp, Matthias, Gobre, Vivekanand, Vazquez-Mayagoitia, Alvaro, Hansen, Katja, Tkatchenko, Alexandre, Müller, Klaus-Robert, Anatole von Lilienfeld, O., 2013. Machine learning of molecular electronic properties in chemical compound space. New J. Phys. 15 (9), 095003.

Munshi, Joydeep, Chen, Wei, Chien, TeYu, Balasubramanian, Ganesh, 2021. Transfer learned designer polymers for organic solar cells. J. Chem. Inform. Model.

Paul, Arindam, Jha, Dipendra, Al-Bahrani, Reda, Liao, Wei keng, Choudhary, Alok, Agrawal, Ankit, 2018. Chemixnet: Mixed dnn architectures for predicting chemical properties using multiple molecular representations.

Paul, Arindam, Mozaffar, Mojtaba, Yang, Zijiang, Liao, Wei-keng, Choudhary, Alok, Cao, Jian, Agrawal, Ankit, 2019. 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). IEEE, pp. 541–550.

Paul, Arindam, Acar, Pinar, Liao, Wei-keng, Choudhary, Alok, Sundararaghavan, Veera, Agrawal, Ankit, 2019b. Microstructure optimization with constrained design objectives using machine learning-based feedback-aware data-generation. Comput. Mater. Sci. 160, 334–351.

Paul, Arindam, Furmanchuk, Alona, Liao, Wei-keng, Choudhary, Alok, Agrawal, Ankit, 2019c. Property prediction of organic donor molecules for photovoltaic applications using extremely randomized trees. Mol. Informat. 38 (11–12), 1900038.

Paul, Arindam, Jha, Dipendra, Al-Bahrani, Reda, Liao, Wei keng, Choudhary, Alok, Agrawal, Ankit, 2019. Transfer learning using ensemble neural networks for organic solar cell screening.

Peter, C. St., John, Caleb Phillips, Kemper, Travis W., Nolan Wilson, A., Guan, Yanfei, Crowley, Michael F., Nimlos, Mark R., Larsen, Ross E., 2019. Message-passing neural networks for high-throughput polymer screening. Jun. J. Chem. Phys. 150 (23), 234111.

Playe, Benoit, Stoven, Veronique, 2020. Evaluation of deep and shallow learning methods in chemogenomics for the prediction of drugs specificity. J. Cheminformatics 12 (1), 11.

Popova, Mariya, Shvets, Mykhailo, Oliva, Junier, Isayev, Olexandr, 2019. MolecularRNN: Generating realistic molecular graphs with optimized properties.

Pyzer-Knapp, Edward, Li, Kewei, Aspuru-Guzik, Alán, 2015. Learning from the harvard clean energy project: The use of neural networks to accelerate materials discovery. Adv. Funct. Mater. 25, 09.

Ratti, Emiliangelo, Trist, David, 2001. Continuing evolution of the drug discovery process in the pharmaceutical industry. Farmaco (Società chimica italiana: 1989) 56, 13–19.

Rogers, David, Hahn, Mathew, 2010. Extended-connectivity fingerprints. J. Chem. Inform. Model. 50, 742–754.

Sahu, Harikrishna, Rao, Weining, Troisi, Alessandro, Ma, Haibo, 2018. Toward predicting efficiency of organic solar cells via machine learning and improved descriptors. Adv. Energy Mater. 8 (24), 1801032.

Sajedian, Iman, Lee, Heon, Rho, Junsuk, 2020. Design of high transmission color filters for solar cells directed by deep q-learning. Sol. Energy 195, 670–676.

Sanchez-Lengeling, Benjamin, Outeiral, Carlos, Guimaraes, Gabriel L., Aspuru-Guzik, Alan, 2017. Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (organic).

Scharber, Markus, Mühlbacher, D., Koppe, M., Denk, Patrick, Waldauf, Ch, Heeger, A.J., Brabec, Christoph, 2006. Design rules for donors in bulk-heterojunction solar cells–towards 10% energy-conversion efficiency. Adv. Mater. 18, 789–794.

Schilinsky, Pavel, Waldauf, Christoph, Brabec, Christoph J., 2002. Recombination and loss analysis in polythiophene based bulk heterojunction photodetectors. Appl. Phys. Lett. 81 (20), 3885–3887.

Schleder, Gabriel Ravanhani, Padilha, Antonio Claudio, Acosta, Carlos, Costa, Marcio, Fazzio, Adalberto, 2019. From dft to machine learning: recent approaches to materials science – a review. J. Phys. Mater. 02.

Schuster, Mike, Paliwal, Kuldip, 1997. Bidirectional recurrent neural networks. IEEE Trans. Signal Process. 45, 2673–2681.

Shao, Yihan, Molnar, Laszlo Fusti, Jung, Yousung, Kussmann, Jörg, Ochsenfeld, Christian, Brown, Shawn T., Gilbert, Andrew T.B., Slipchenko, Lyudmila V., Levchenko, Sergey V., O'Neill, Darragh P., DiStasio Jr., Robert A., Lochan, Rohini C., Wang, Tao, Beran, Gregory J.O., Besley, Nicholas A., Herbert, John M., Lin, Ching Yeh, Van Voorhis, Troy, Chien, Siu Hung, Sodt, Alex, Steele, Ryan P., Rassolov, Vitaly A., Maslen, Paul E., Korambath, Prakashan P., Adamson, Ross D., Austin, Brian, Baker, Jon, Byrd, Edward F.C., Dachsel, Holger, Doerksen, Robert J., Dreuw, Andreas, Dunietz, Barry D., Dutoi, Anthony D., Furlani, Thomas R., Gwaltney, Steven R., Heyden, Andreas, Hirata, So, Hsu, Chao-Ping, Kedziora, Gary, Khalliulin, Rustam Z., Klunzinger, Phil, Lee, Aaron M., Lee, Michael S., Liang, WanZhen, Lotan, Itay, Nair, Nikhil, Peters, Baron, Proynov, Emil I., Pieniazek, Piotr A., Rhee, Young Min, Ritchie, Jim, Rosta, Edina, David Sherrill, C., Simmonett, Andrew C., Subotnik, Joseph E., Lee Woodcock III, H. Zhang, Weimin, Bell, Alexis T., Chakraborty, Arup K., Chipman, Daniel M., Keil, Frerich J., Warshel, Arieh, Hehre, Warren J., Schaefer III, Henry F., Kong, Jing, Krylov, Anna I., Gill, Peter M.W., Head-Gordon, Martin, 2006. Advances in methods and algorithms in a modern quantum chemistry program package. Phys. Chem. Chem. Phys. 8, 3172–3191.

Shi, Chence, Xu, Minkai, Zhu, Zhaocheng, Zhang, Weinan, Zhang, Ming, Tang, Jian, 2020. GraphAF: a flow-based autoregressive model for molecular graph generation.

Shin, Bonggun, Park, Sungsoo, Kang, Keunsoo, Ho, Joyce C., 2019. Self-attention based molecule representation for predicting drug-target interaction.

Smets, A., Jäger, K., Isabella, O., van Swaaij, R., Zeman, M., 2019. Solar Energy: The Physics and Engineering of Photovoltaic Conversion, Technologies and Systems. UIT Cambridge.

Sterling, Teague, Irwin, John, 2015. Zinc 15 - ligand discovery for everyone. J. Chem. Inform. Model. 55, 10.

Valleau, Stéphanie, Häse, Florian, Pyzer-Knapp, Edward, Aspuru-Guzik, Alán, 2016. Machine learning exciton dynamics. Chem. Sci. 7, 04.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, Polosukhin, Illia, 2017. Attention is all you need.

Wang, Guotai, Li, Wenqi, Aertsen, Michael, Deprest, Jan, Ourselin, Sébastien, Vercauteren, Tom, 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing 338, 34–45.

Weininger, David, 1988. Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inform. Comput. Sci. 28, 31–36.

Winter, Robin, Montanari, Floriane, Steffen, Andreas, Briem, Hans, Noé, Frank, Clevert, Djork-Arné, 2019. Efficient multi-objective molecular optimization in a continuous latent space.

Yan, Chaochao, Wang, Sheng, Yang, Jinyu, Xu, Tingyang, Huang, Junzhou, 2020. Re-balancing variational autoencoder loss for molecule sequence generation.

Yang, Zijiang, Jha, Dipendra, Paul, Arindam, Liao, Wei keng, Choudhary, Alok, Agrawal, Ankit, 2020. Generative adversarial networks with mixture density networks for inverse modeling in materials microstructural design.

Zang, Chengxi, Wang, F., MoFlow: An invertible flow model for generating molecular graphs.

Zheng, Shuangjia, Yan, Xin, Yang, Yuedong, Xu, Jun, 2018. Identifying structure-property relationships through smiles syntax analysis with self-attention mechanism.

Zhou, Zhenpeng, Kearnes, Steven, Li, Li, Zare, Richard N., Riley, Patrick, 2019. Optimization of molecules via deep reinforcement learning. 9(1), 10752. Number: 1 Publisher: Nature Publishing Group.