

Transfer Learning Using Ensemble Neural Networks for Organic Solar Cell Screening

Arindam Paul*, Dipendra Jha*, Reda Al-Bahrani, Wei-keng Liao, Alok Choudhary and Ankit Agrawal
Electrical Engineering and Computer Science, Northwestern University
Email: arindam.paul, dkj755, rav650, wkliao, choudhar, ankitag@eecs.northwestern.edu

Abstract—Organic Solar Cells are a promising technology for solving the clean energy crisis in the world. However, generating candidate chemical compounds for solar cells is a time-consuming process requiring thousands of hours of laboratory analysis. For a solar cell, the most important property is the power conversion efficiency which is dependent on the highest occupied molecular orbitals (HOMO) values of the donor molecules. Recently, machine learning techniques have proved to be very useful in building predictive models for HOMO values of donor structures of Organic Photovoltaic Cells (OPVs). Since experimental datasets are limited in size, current machine learning models are trained on data derived from calculations based on density functional theory (DFT). Molecular line notations such as SMILES or InChI are popular input representations for describing the molecular structure of donor molecules. The two types of line representations encode different information, such as SMILES defines the bond types while InChI defines protonation. In this work, we present an ensemble deep neural network architecture, called SINet, which harnesses both the SMILES and InChI molecular representations to predict HOMO values and leverage the potential of transfer learning from a sizeable DFT-computed dataset- Harvard CEP to build more robust predictive models for relatively smaller HOPV datasets. Harvard CEP dataset contains molecular structures and properties for 2.3 million candidate donor structures for OPV while HOPV contains DFT-computed and experimental values of 350 and 243 molecules respectively. Our results demonstrate significant performance improvement from the use of transfer learning and leveraging both molecular representations.

I. INTRODUCTION

Based on current statistics, energy consumption had increased from 238 Exajoules (EJ) in 1972 to 464 EJ in 2004. A further 65% increase is projected by the year 2030 [1]. Sustained usage of fossil fuels leads to irreversible changes to the planet with sea level rising from 1.7 to 3.2 mm per year and ocean temperatures increasing [1], [2]. It is imperative to search for versatile and cost-efficient clean energy solutions to prevent further irreversible damage. One limitation with renewable energy is that it is difficult to generate the quantities of electricity that are as large as those produced by traditional fossil-fuel generators [3]. Wind and hydro-energy solutions require expensive installations [4] and maintenance, which requires large government grants, and are dependent on weather and climate conditions [5]. Moreover, most renewable energy technology is new and has enormous capital costs compared to traditional fossil fuels. Solar energy provides a more cost-effective solution with faster installation, and more

predictive energy outputs based on the Bureau of Meteorology and National Aeronautics and Space (NASA) reports [6]. Although inorganic silicon-based solar energy systems are currently more conventional, organic or plastic photovoltaic (OPV) [7] technology has become very popular because of its flexibility. Organic or Plastic Technology is very versatile as demonstrated by how plastics in consumer goods can be made very hard and durable, or very light or transparent as dictated by needs [8]. Further, manufacturing costs are lower for organic solar cells compared to silicon-based materials due to the ease of device manufacturing, and lower cost of organic components compared to silicon [9].

However, the main bottleneck in the deployment of organic solar cells is that the search for candidate chemical compounds for creating organic solar cells is very time consuming [10]; it can take up to thousands of hours of laboratory analysis. For a solar cell, the most important property is power conversion efficiency (PCE) or the percentage of electricity which can be generated due to the interaction of electron donors and acceptors after absorption of energy from the sun. The PCE is dependent on the highest occupied molecular orbital (HOMO) energy of the donor and the lowest unoccupied molecular orbital (LUMO) energy of the acceptor molecule [11]. However, as the LUMO values across known acceptors do not vary much, and only a few acceptor molecules exist, predicting HOMO values of donor molecules can give us estimates of PCE when those donors are used in solar cells.

There are two main issues with the current practice of building predictive models using machine learning (ML) techniques. First, these predictive models are built using a single representation of the molecular structure - line notations [12] such as SMILES or InChI, molecular fingerprints [13] or molecular graphs [14]. Line notations are increasingly becoming popular for use in ML models as molecular fingerprints are difficult to interpret and models trained on molecular graphs usually perform worse [15]. However, these approaches have restricted themselves to only one type of line notation - either SMILES or InChI as input representations for predictive models. This limits the information that can be harnessed from these representations as SMILES and InChI express the molecular structure in very different ways. SMILES defines the chemical bond types present in the molecular structure from which one can infer the protonation while InChI serves the opposite purpose- it defines the protonation from which one can infer the chemical bond types present in the molec-

*equal contribution

ular structure. Also, SMILES was designed to be read and written by humans whereas InChI was intended to ignore tautomeric form and be more consistent. Since the two line representations are distinct in their properties, a predictive model can benefit from the use of both of them. Second, most of the datasets, especially, experimental datasets are limited in size. Hence, current ML models are either built using publicly available large DFT-computed datasets such as the Harvard CEP dataset [13], [16] or other limited experimental or DFT-computed datasets which are relatively smaller in size and hence, the model cannot learn the required data representation for making robust predictions. In this work, our goal is to leverage together larger DFT-computed datasets with relatively smaller datasets such as experimental observations and combine both types of line representations- SMILES and InChI – to build more robust predictive models for predicting HOMO values for donor candidates for OPV.

We present an ensemble deep neural network architecture, called SINet, which leverages both the SMILES and InChI molecular representations to learn to predict the HOMO values, and leverage transfer learning from large datasets to build more robust predictive models for a relatively smaller dataset. SINet is composed of two identical branches for both types of inputs; each branch consists of 1-D CNN layers followed by LSTM layers. The features learned by the two branches from the two input representations are combined and fed into a fully connected network for predicting the regression output of HOMO value. The deep neural network architecture of SINet enables us to perform transfer learning from a large dataset to relatively smaller dataset in a similar domain. Transfer learning has already been adopted in the fields of computer vision, natural language processing and other application domains [17], [18].

Our source dataset for transfer learning is the Harvard CEP dataset [13], [16] which contains molecular structures and properties for 2.3 million candidate donor structures for OPV. For the target dataset, we leverage DFT-computed and experimental values of 350 and 243 molecules respectively, from the HOPV [19]. Our results demonstrate significant performance improvement from the use of both types of inputs- the MAPE drops from 0.972% and 0.457% using SMILES and InChI respectively, to 0.213% when using them together using the SINet architecture. We also find significant benefit from using transfer learning from Harvard CEP to the HOPV datasets. The MAPE for experimental and DFT-computed datasets from HOPV drops from 2.782% to 1.513% and 2.118% to 1.478%, respectively. Since the model is first trained on a large dataset, it learns the required set of features from the input data representation, and this helps in learning the similar features present in the smaller target dataset, on which the model is fine-tuned. Our results demonstrate significant benefit from the use of both types of input representations as well as from transfer learning from a larger dataset. It showcases that leveraging machine learning with computational and experimental chemistry can play an essential role in the expedition of a systematic design of high-efficiency OPV materials, and holds

significant promise as a potential solution to future energy needs.

II. BACKGROUND AND RELATED WORKS

In this section, we present a description of the OPV technology and Scharber model, and the two molecular representations we use in this work – SMILES and InChI, and discuss existing machine learning systems for predictive modeling for materials.

A. Organic Photovoltaic Cells

Among current solar cell design paradigms, organic photovoltaic cell technology is a promising technology for the inexpensive and versatile utilization of solar energy. The traditional development of new OPV materials is predominantly based on empirical intuition or experience of materials scientists. A new design idea is followed by a labor-intensive synthesis, characterization, and prototype device optimization. Hence, the problem space of OPV renewable energy research is notably complicated as the design of successful OPV materials is a multifaceted problem. The conversion of sunlight into electricity can be achieved using a solar cell and is one of the most attractive future sources of energy. Ever since the development of the first solar cells, there has been an accelerated and comprehensive exploration for cost-effective photovoltaics. OPVs are potential cost-effective and lightweight alternatives to silicon-based solar cells and could lead to the most substantial reduction of production cost. After being excited with light, firmly bound electron-hole pairs (excitons) are generated. As illustrated in Figure 1, an OPV works by absorbing a photon emitted by the sun. The photon carries energy that is used to excite an electron off a donor layer, often comprised of a semiconducting polymer.

However, for the solar cell to generate electricity, the electron and hole must be separated and subsequently collected at electrodes of opposite polarity. In order to accomplish this, the exciton bond must be broken. This happens at the donor-acceptor interface, where the exciton splits into separate free electron and hole. As the charges separate further, they can reach electrodes which upon becoming charged, generate electricity as the electrons move from the cathode to the anode.

OPVs have the advantage of combining the versatility and flexibility of plastics with photo-electronics. They can be made semi-transparent, and moldable into different forms and shapes. Researchers have even tried spray-coating OPVs on various surfaces [21]. Nonetheless, production scale small-area devices yield efficiency of only about 5% [22], with laboratory experiments yielding the highest efficiency of around 10%, and hence there is much room for improvement. Several models predict the efficiencies to reach 15% assuming the usage of state of the art materials and device architectures [23]. The conventional process for the generation of such devices is iterative and time-consuming. However, due to the labor-intensive process of generating candidates for OPVs, producing virtual screening techniques as elucidated by Pyzer-Knapp

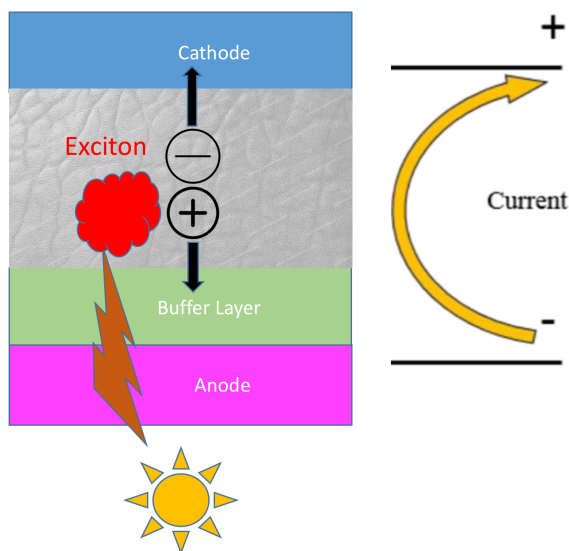


Fig. 1. Photo-electricity generation in a bulk heterojunction Organic photovoltaic cell [20]. When the photons from the sun hit the surface of the OPV device, an electron from the donor is excited and combines with the corresponding hole at the acceptor layer to form an exciton. Electricity is generated when the exciton splits at the interface, and electrons move from the cathode to the anode.

et al. [13] and our current study can potentially fasten the process considerably.

B. Scharber Model

For a solar cell, the most important property is power conversion efficiency or the amount of electricity which can be generated due to the interaction of electron donors and acceptors. The Scharber model [11] provides a relation between the voltage V_{oc} and the energies of the HOMO and the LUMO level of the donor and acceptor molecules, which in turn can be related to the power conversion efficiency (PCE), the maximum efficiency of solar cells. In the following equation, J_{sc} is the short-circuit current density, FF is electrical fill factor and P_{in} is incident-light intensity. $E^{Donor\ HOMO}$ and $E^{Acceptor\ LUMO}$ indicate the HOMO and LUMO energy levels of the donor and acceptor molecules respectively.

$$V_{oc} = (1/e)(E^{Donor\ HOMO} - E^{Acceptor\ LUMO}) - 0.3V$$

$$PCE = 100 * (V_{oc} * FF * J_{sc}) / P_{in}$$

However, it is necessary to assert that the predictions from the Scharber analysis are restrained by various assumptions and the quality of the input data. The assumptions include requirements related to complicated bulk and interface behavior, and exciton and charge carrier dynamics. Hence, the resultant values should be understood as the potential performance that can be achieved if the assumptions are met.

C. SMILES

Line notations are linear representations of chemical structures which encode the connection table and the stereochemistry of a molecule as a line of text [12]. Simplified

Molecular Input Line Entry System (SMILES) [24], [25] is the most popular specification in the form of a line notation for describing the structure of chemical species using short ASCII strings encoding molecular structures and specific instances. One or more organic molecules attach to form long continuous chains known as branches. In the SMILES notation, branches are described by parentheses.

Molecules and reactions can be specified using ASCII characters representing atom and bond symbols. Most molecular editors and libraries can import SMILES strings. The other advantages of the SMILES format are that it is both human readable and writable. Further, it encodes the stereochemistry of the molecule intuitively. In this work, we limit ourselves to character level representation and do not explicitly encode the grammar. Based on the canonical ordering of atoms, unique SMILES is generated by depth-first search.

D. InChI


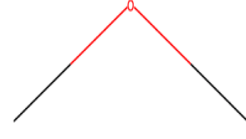
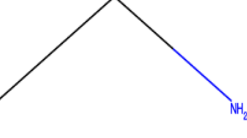
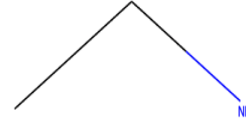
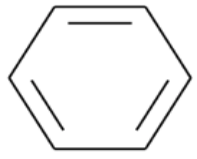
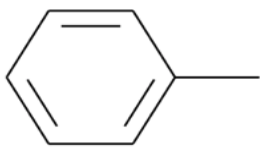
The IUPAC International Chemical Identifier (InChI) [26] was developed by IUPAC and NIST (National Institute of Standards and Technology). It is a textual identifier for chemical substances which provides a standard way to encode molecular information. Every InChI string starts with the string "InChI=" followed by the version number and letter S in the case of standardized InChIs. Rest of the InChI string is structured as a sequence of layers and sub-layers. Each layer provides a specific type of information, and are separated by "/". The InChI algorithm transforms the structural information of the molecule into a unique InChI identifier in a three-step process. The first step is normalization which removes redundant information. This is followed by canonicalization that generates a unique number label for each atom. The last step is serialization that produces a string of characters.

SMILES and InChI are distinct notations. SMILES defines the bond types from which one can infer protonation, while InChI defines protonation from which one can infer the bond types. SMILES was designed to be read and written by humans and is therefore relatively straightforward to read, provided the user knows a few basic principles of the format. InChI, is comparatively less readable, is intended to ignore tautomeric form and is consistent. Table I illustrates the distinction between SMILES and InChI representations.

E. Related Works

There have been many efforts in the field of ML-assisted materials prediction and discovery [13], [14], [27]–[35]. For instance, Sajeev et al. [36] developed a virtual library of semi-conductors and non semi-conductors of small organic compounds and applied various ML classifiers to predict the conductivity of a given compound. They used block descriptors as attributes which were generated based on E-Dragon Software [37]. Riede et al. [38] used principal component analysis on a small data set of 62 organic solar cells which consisted of manufacturing parameters and measurement results on seven different substrates. Olivares-Amaya et al. [39] employed Marvin code by ChemAxon [40] on a set

TABLE I
 EXAMPLES OF SET OF SIMILAR CHEMICAL COMPOUNDS WITH THEIR CORRESPONDING SMILES AND InChI NOTATIONS WITH EXPLANATION

Compound 1	Compound 2	Line Notations (with explanation)
 Ethanol	 Dimethyl Ether	<ul style="list-style-type: none"> • SMILES: CCO and COC • InChI: InChI=1S/C2H6O/c1-2-3/h3H,2H2,1H3 and InChI=1S/C2H6O/c1-3-2/h1-2H3 • "C2H6O" means that the first and second atoms (1 and 2) are C atoms and the third (3) is an O atom. The connectivity is 1-2-3 for ethanol and 1-3-2 for dimethyl ether. For ethanol atom 3 has 1 H atom, atom 2 has 2 H atoms, and atom 1 has 3 H atoms. For dimethyl ether atom 1-2 have 3 H atoms, while atom 3 has none.
 Ethylamine	 Ethylammonium	<ul style="list-style-type: none"> • SMILES: CCN and CC[NH3+] • InChI=1S/C2H7N/c1-2-3/h2-3H2,1H3 and InChI=1S/C2H7N/c1-2-3/h2-3H2,1H3/p+1 • For the InChI notation, the protonation (h2-3H2,1H3) is identical in both cases and corresponds to ethylamine. For ethylammonium "p+1" indicates that an extra proton is added.
 Benzene	 Toluene	<ul style="list-style-type: none"> • SMILES: c1ccccc1 and Cc1ccccc1 • InChI=1S/C6H6/c1-2-4-6-5-3-1/h1-6H and InChI=1S/C7H8/c1-7-5-3-2-4-6-7/h2-6H,1H3 • For the SMILES notation, in the case of benzene atom 1 is connected to both atom 2 and atom 6, i.e. a ring is formed. "1" is a label and does not refer to atom number 1 (see toluene). A lower case "c" is used to indicate aromatic carbons, meaning they should be singly protonated. For toluene, the methyl group is bonded to atom number 2, which is also bonded to atom number 7. For the InChI notation, in the case of benzene aromaticity is inferred from the fact that all 6 carbon has 1 H atom (h1-6H). For toluene, the methyl group is bonded to atom number 7, which is also bonded to atom number 6.

of 50 training molecules compiled from the literature which provides a set of over 200 descriptors relevant for applications based on organic materials. Mannodi-Kanakithodi et al. [28] used kernel ridge regression (KRR) to transform the input fingerprint of a chemical compound into higher dimensional space to establish a linear relationship between the transformed fingerprint and the property of interest. They addressed the issue of accelerating polymer dielectrics design by extracting learning models from data generated by accurate state-of-the-art first-principles computations for polymers occupying a vital part of the chemical subspace. In their 2013 work, Kanal et al. [20] focused on a screening pipeline which used a genetic algorithm for initial screening and multiple filtering stages for further refinement of the HOMO and LUMO properties of acceptors and donors. In the Harvard Energy Clean Project, Pyzer-Knapp et al. [13] used a multi-layered perceptron (MLP) for predicting power conversion efficiency of organic photovoltaic materials from a sizable database of DFT-derived properties.

Molecular graph convolutions [14] is an ML architecture for learning from undirected graphs, specifically from small molecules, and have become popular representations for molecular structures. It uses a simple encoding of the molecular graph atoms, bonds, the distance between atoms that allow the model to take more advantage of information in the graph structure. More recently, Goh et al. [34] developed an RNN neural network architecture SMILES2vec trained on SMILES for predicting chemical properties across different datasets. They did not explicitly encode information about

SMILES grammar and rather allowed the RNN units to understand the intermediate features useful for property prediction implicitly. Their work also demonstrated that networks trained on line notations can perform better than other representations such as molecular graphs. Paul et al. [35] used a combination of fingerprints and SMILES to develop a deep neural network architecture called CheMixNet, and it outperformed SMILES2vec trained only on SMILES strings on the same datasets.

III. METHOD

In this section, we present the source and target organic photovoltaic datasets used in our experiments, discuss the preprocessing of the SMILES and InChI strings and propound our methodology.

A. Datasets

The source dataset for transfer learning is the Harvard CEP Dataset [13], [16] which contains molecular structures and properties for 2.3 million candidate donor structures for organic photovoltaic cells. For a solar cell, the most important property is power conversion efficiency or the amount of electricity which can be generated due to the interaction of electron donors and acceptors, which are dependent on the HOMO values of the donor molecules. In this work, we considered the highest occupied molecular orbitals (HOMO) as the target property as it determines the power conversion efficiency of a solar cell according to the Scharber model [11].

The target dataset was the Harvard Organic Photovoltaic (HOPV) dataset [19] which is a collection of photovoltaic

measurements for a diverse set of 350 organic molecules generated by extensively searching the literature. Of these, experimental values were available for 243 molecules and calculated values using density functional theory (DFT) were available for 344 molecules. In our experiments, the DFT-computed values in the HOPV dataset were reduced to 344 molecules after removing redundant isomeric samples [41]. We used both the experimental and calculated datasets as target datasets for transfer learning. The HOPV dataset contains density functional theory (DFT) calculations for four functionals B3LYP, BP86, PBE and M06 using the basis set def2-SVP [42]. We used B3LYP functional values as it is the most popular functional for HOMO value calculations. Further, HOMO values across all conformers were Boltzmann-weight averaged [43].

B. Preprocessing

For both the SMILES and InChI sequences, one-hot encoding was performed separately to convert them into fixed length representations. The sequence lengths were calculated using the length of the longest SMILES and InChI sequences in the dataset respectively. To maintain a uniform sequence size, shorter strings were padded with zeros. Similar to SMILES2vec, vocabulary size was equal to the number of unique characters. The sequence lengths are 82 and 162 respectively for SMILES and InChI input representations.

C. SINet

Figure 2 illustrates the proposed approach for performing transfer learning from the Harvard CEP dataset to the two smaller target datasets in the HOPV dataset. The deep neural network architecture used for the task consists of two branches for the two types of the input representations of SMILES and InChI. The SMILES input vector has a length of 82 while the InChI input has 162 values. Both branches have the same network configuration. Each branch is composed of a 1-D CNN followed by an LSTM network. The 1-D CNN is composed of two layers with 32 filters each; the filter size used in each layer is 3 and same padding for the inputs and output. The convolutional layers are followed by max pooling with a pool size of 2. There was no significant difference with other types of pooling and other pooling sizes. The output of the 1-D CNN is fed into the LSTM network which is composed of 2 layers having 64 units each. Finally, the outputs from both branches are concatenated into the merge layer and fed into a fully connected network which is composed of a penultimate dense layer with 64 units and the final layer that gives the HOMO value as the regression output. Since the network architecture leverages both SMILES and InChI molecular representations, we refer to it as SINet.

For transfer learning, first, we train a model on the source dataset of Harvard CEP from scratch (by initializing the model parameters from scratch before training). While being trained on the large dataset, the model learns a rich set of feature representations present in the large training data which is useful for making predictions in the source domain. Next,

for using transfer learning, we can follow one of the two techniques. Either, the same trained model can be fine-tuned by training on the target dataset, or we can initialize a new model using the model parameters from the model trained on the source data and then fine-tune it on the target data. In this work, we use the first approach as the target dataset is very small, and we wanted to harness the source dataset as much as possible. In the case of transfer learning, rather than learning all the feature representations present in the input data from scratch, the model already knows the input data distribution from the source dataset and only fine-tunes its parameters to adapt to the target dataset.

IV. EXPERIMENTAL RESULTS

In this section, we present the experimental settings and results of the ensemble SINet architecture including the impact of transfer learning for performance gain on the smaller HOPV datasets.

A. Experimental Settings

The models were implemented using Python and Keras [44] with TensorFlow [45] as the backend. We used Adam as the optimization algorithm with a mini-batch size of 32. For generating the InChI fingerprints for the CEP dataset, we used RDKit [46] library to generate InChI from the molecules. Scikit-Learn [47] was used for data preprocessing and for evaluating the test set errors. All experiments are carried out using NVIDIA DIGITS DevBox with a Core i7-5930K 6 Core 3.5GHz desktop processor, 64GB DDR4 RAM, and 4 TITAN X GPUs with 12GB of memory per GPU. We performed extensive hyperparameter search as well as architecture search for SINet. For our experiments, we used a learning rate of 0.001. We used the mean squared error (MSE) as the loss function and used the mean absolute % error (MAPE) as the performance metric. Early stopping was used during training to avoid over-fitting. For our experiments, we split each dataset into 70-20-10 ratio for training, test and validation sets; we used the same split for all experiments of each dataset. Stratified shuffling was used to ensure that the distribution of HOMO values for all the 3 subsets was similar.

B. Impact of Leveraging SMILES & InChi

First, we explored the performance of using different types of input representations and their combinations on the source and the target datasets. On the source dataset of Harvard CEP, we observe that the MAPE decreased to 0.213% while using the SINet model as shown in Figure 3. In contrast, while using the individual input representations with the individual branch of SINet, the MAPE values were 0.972% and 0.457% using SMILES and InChi. This was also true for the target datasets of experimental and DFT-computed values from HOPV. We conjecture the prediction mainly improves while leveraging multiple molecular representations because the two line notations- SMILES and InChI differ in the representation and detail; hence, the model can learn different feature representation from the two input representation, leading to better performance.

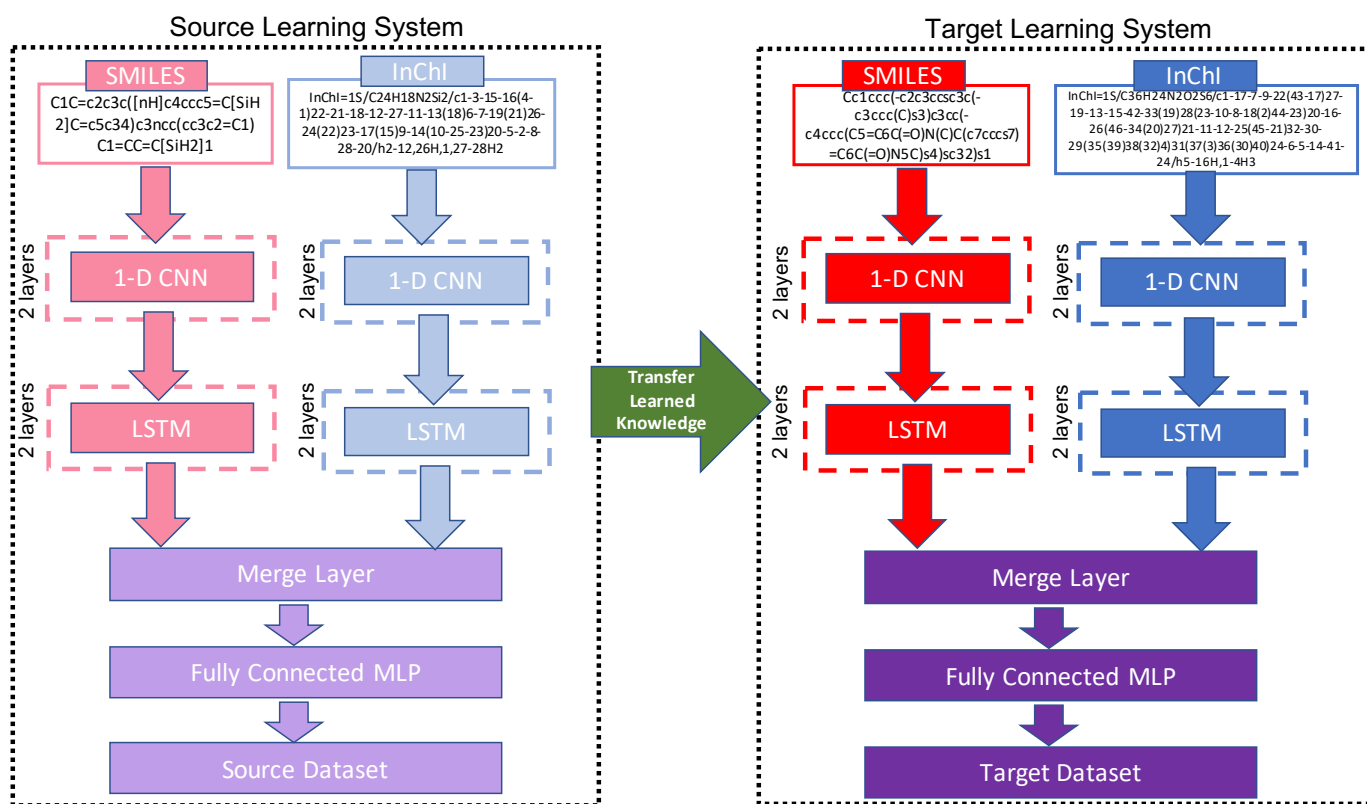


Fig. 2. The proposed SINet architecture for learning from the two text-based molecular representations - SMILES and InChI. The left side (represented by faded colors) represents the learning from the source dataset while the right side (represented by darker colors) represents the learning for the target dataset. For both the learning systems, the red branch represents the network for sequence modeling from SMILES while the blue branch represents the network for sequence modeling from InChI. The purple part represents the fully connected layers that learn the final output from the combination of features learned by the two network branches. We exemplify the SMILES and InChI with one representative example in this illustration from both the source as well as target datasets.

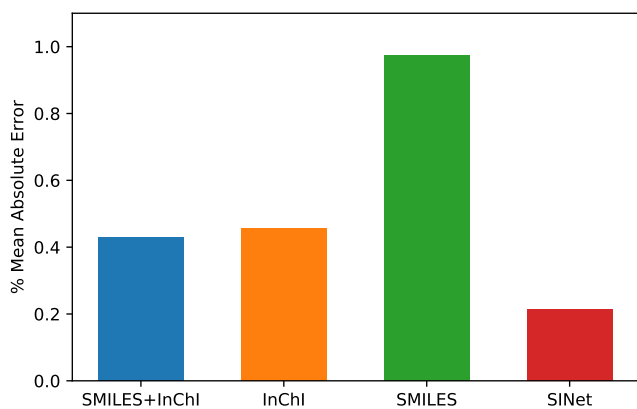


Fig. 3. Mean Absolute Error Percentage for the CEP Dataset (source dataset)

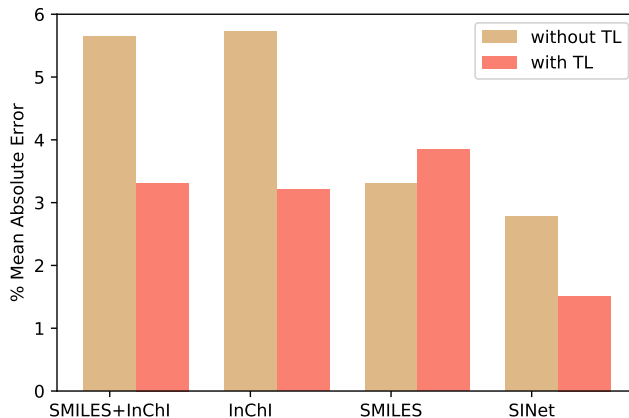


Fig. 4. Mean Absolute Error Percentage for the OPV Experimental Dataset without and with Transfer Learning (TL)

Furthermore, we experimented with simply combining the two input representations- SMILES and InChI into a single input vector (represented as SMILES + InChI in Figure 3), before feeding them into a branch of SINet, the MAPE, in this case, was 0.430% which is lower than while using single input. However, there was no benefit from simply

combining the two input representations into a single vector in the case of experimental target dataset. We surmise that this could be because the two line notations encode different representations with varying lengths for different compounds, and a concatenation of the representations was not sufficient

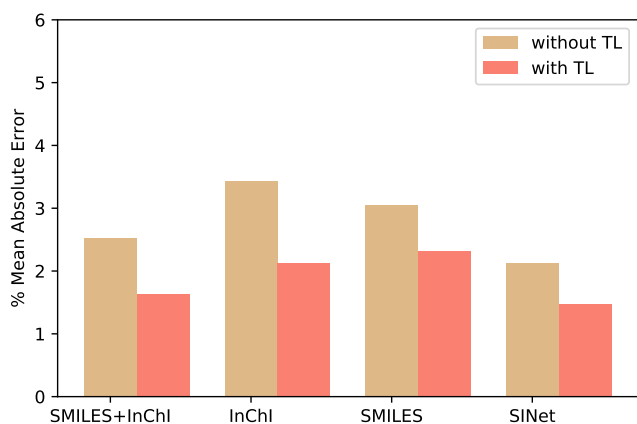


Fig. 5. Mean Absolute Error Percentage for the OPV DFT Dataset without and with Transfer Learning (TL)

for learning both notations. Our results recommend that a better way to incorporate multiple input representations such as SMILES and InChI, in this case, is to design the deep neural network to have different model components to handle each of them before the learned features can be combined to make the final output as in the case of SINet.

In addition, we can observe the impact of training data size on the prediction performance; the prediction error of SINet on Harvard CEP dataset is significantly lower than the prediction error of SINet on the two other relatively smaller datasets. This also justifies the use of large dataset as the source dataset while using transfer learning to a smaller target dataset.

C. Impact of Transfer Learning

We also investigated the impact of transfer learning from the DFT-computed dataset of Harvard CEP to the relatively smaller DFT-computed and experimental datasets from HOPV. For the experimental data having only 243 samples, the MAPE in case of SINet decreased significantly from 2.782% to 1.513% which is around half. We observed similar changes when using just one input or their simple combination as shown in Figure 4. For the target dataset of DFT-computed dataset with 344 samples, the error for SINet decreased from 2.118% to 1.478% (in Figure 5). Such a significant drop in the MAPE for both our target datasets illustrates the efficacy of using transfer learning from large datasets when doing predictive modeling on smaller datasets with a lesser number of samples. The experiments exhibit that when a model is trained on a large dataset (model parameters being initialized from a model trained on large source dataset), it already captures the required features from the dataset which makes it easy to learn the features present in target data from a similar domain on which it is fine-tuned.

V. CONCLUSION AND FUTURE WORKS

In this paper, we presented a novel approach of predictive modeling for HOMO values of donor molecules for the

generation of OPV candidates by leveraging both large DFT-computed dataset and relatively smaller DFT-computed and experimental datasets using both types of input representation—SMILES and InChI, using the concept of transfer learning with deep neural networks. For the source dataset, we leveraged the Harvard CEP dataset which contains millions of OPV candidates with the DFT-computed HOMO values. For the target dataset, we used the DFT-computed and experimental data from HOPV which contains relatively smaller data—344 and 243, respectively.

Our results demonstrate significant benefit from the use of both types of input representations as well as from transfer learning from a larger dataset. It showcases that leveraging machine learning with computational and experimental chemistry can play an essential role in the expedition of a systematic design of high-efficiency OPV materials, and holds significant promise as a potential solution to future energy needs. The search process for the donor cells with high HOMO values can be made faster by leveraging transfer learning from a larger calculated dataset to a small well-curated experiment-theory calibrated dataset, and this exposes an exciting area in materials discovery, and in particular for solar cell technology. Further, as our approach is based on simple text representations, it is easier for chemists to explore adding or removing subgroups to the chemical compounds to explore the impact on power efficiency instead of performing elaborate experiments.

As a future work, we believe Hierarchical Attention Networks (HANs) [48] that combine character level and word level sequences for text prediction can be harnessed to exploit the layer-based nomenclature of the InChI representation. Although we limited the scope of this current work for organic photovoltaic datasets, we believe that SINet would be helpful for the broad communities of cheminformatics and bioinformatics, as SMILES and InChI are popular representations for any type of organic molecule.

ACKNOWLEDGMENT

This work is supported in part by the following grants: NIST award 70NANB14H012, NSF award CCF-1409601; DOE awards DE-SC0014330, DE-SC0019358.

REFERENCES

- [1] T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, "Climate change 2013: The physical science basis," 2014.
- [2] A. Hoffman, "Computational chemistry in rational material design for organic photovoltaics," 2015.
- [3] G. Boyle *et al.*, *Renewable energy: power for a sustainable future*. Taylor & Francis, 1997.
- [4] J. A. Turner, "A realizable renewable energy future," *Science*, vol. 285, no. 5428, pp. 687–689, 1999.
- [5] L. Baxter, "Biomass-coal co-combustion: opportunity for affordable renewable energy," *Fuel*, vol. 84, no. 10, pp. 1295–1302, 2005.
- [6] NASA, "Nasa - clean energy," <https://www.nasa.gov/centers/ames/greenspace/clean-energy.html>, 2016.
- [7] J. Yu, Y. Zheng, and J. Huang, "Towards high performance organic photovoltaic cells: A review of recent development in organic photovoltaics," *Polymers*, vol. 6, no. 9, pp. 2473–2509, 2014.

- [8] C. Brabec, U. Scherf, and V. Dyakonov, *Organic photovoltaics: materials, device physics, and manufacturing technologies*. John Wiley & Sons, 2011.
- [9] O. A. Abdulrazzaq, V. Saini, S. Bourdo, E. Dervishi, and A. S. Biris, "Organic solar cells: a review of materials, limitations, and possibilities for improvement," *Particulate science and technology*, vol. 31, no. 5, pp. 427–442, 2013.
- [10] S. R. Forrest, "The limits to organic photovoltaic cell efficiency," *MRS bulletin*, vol. 30, no. 1, pp. 28–32, 2005.
- [11] M. C. Scharber, D. Mühlbacher, M. Koppe, P. Denk, C. Waldauf, A. J. Heeger, and C. J. Brabec, "Design rules for donors in bulk-heterojunction solar cells—towards 10% energy-conversion efficiency," *Advanced materials*, vol. 18, no. 6, pp. 789–794, 2006.
- [12] W. A. Warr, "Representation of chemical structures," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 1, no. 4, pp. 557–579, 2011.
- [13] E. O. Pyzer-Knapp, K. Li, and A. Aspuru-Guzik, "Learning from the harvard clean energy project: The use of neural networks to accelerate materials discovery," *Advanced Functional Materials*, vol. 25, no. 41, pp. 6495–6502, 2015.
- [14] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in neural information processing systems*, 2015, pp. 2224–2232.
- [15] G. B. Goh, N. O. Hodas, C. Siegel, and A. Vishnu, "Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties," *arXiv preprint arXiv:1712.02034*, 2017.
- [16] "Cepdata.csv.zip," <https://www.dropbox.com/s/3kqzt9u1ryflls0/CEPData.csv.zip?dl=0>, (Accessed on 10/15/2018).
- [17] S. J. Pan, Q. Yang *et al.*, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
- [18] S. Hoo-Chang, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Noguez, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, p. 1285, 2016.
- [19] "The harvard organic photovoltaics 2015 (hopv) dataset: An experiment-theory calibration resource." https://figshare.com/articles/HOPV15_Dataset/1610063, 2016, (Accessed on 09/22/2016).
- [20] I. Y. Kanal, S. G. Owens, J. S. Bechtel, and G. R. Hutchison, "Efficient computational screening of organic polymer photovoltaics," *The journal of physical chemistry letters*, vol. 4, no. 10, pp. 1613–1623, 2013.
- [21] C. N. Hoth, R. Steim, P. Schilinsky, S. A. Choulis, S. F. Tedde, O. Hayden, and C. J. Brabec, "Topographical and morphological aspects of spray coated organic photovoltaics," *Organic Electronics*, vol. 10, no. 4, pp. 587–593, 2009.
- [22] J. Xue, S. Uchida, B. P. Rand, and S. R. Forrest, "Asymmetric tandem organic photovoltaic cells with hybrid planar-mixed molecular heterojunctions," *Applied Physics Letters*, vol. 85, no. 23, pp. 5757–5759, 2004.
- [23] M. C. Scharber and N. S. Sariciftci, "Efficiency of bulk-heterojunction organic solar cells," *Progress in polymer science*, vol. 38, no. 12, pp. 1929–1940, 2013.
- [24] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of chemical information and computer sciences*, vol. 28, no. 1, pp. 31–36, 1988.
- [25] D. Toolkit, "Daylight chemical information systems," *Inc.: Aliso Viejo, CA*, 1997.
- [26] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi, "Inchi, the iupac international chemical identifier," *Journal of cheminformatics*, vol. 7, no. 1, p. 23, 2015.
- [27] D. Jha, S. Singh, R. Al-Bahrani, W.-k. Liao, A. Choudhary, M. De Graef, and A. Agrawal, "Extracting grain orientations from ebsd patterns of polycrystalline materials using convolutional neural networks," *Microscopy and Microanalysis*, vol. 24, no. 5, pp. 497–502, 2018.
- [28] A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, and R. Ramprasad, "Machine learning strategy for accelerated design of polymer dielectrics," *Scientific reports*, vol. 6, 2016.
- [29] D. Jha, L. Ward, A. Paul, W.-k. Liao, A. Choudhary, C. Wolverton, and A. Agrawal, "Elemnet: Deep learning the chemistry of materials from only elemental composition," *Scientific reports*, vol. 8, no. 1, p. 17593, 2018.
- [30] M. Mozaffar, A. Paul, R. Al-Bahrani, S. Wolff, A. Choudhary, A. Agrawal, K. Ehmann, and J. Cao, "Data-driven prediction of the high-dimensional thermal history in directed energy deposition processes via recurrent neural networks," *Manufacturing letters*, vol. 18, pp. 35–39, 2018.
- [31] A. Agrawal and A. Choudhary, "Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science," *APL Materials*, vol. 4, no. 5, p. 053208, 2016.
- [32] Z. Yang, Y. C. Yabansu, D. Jha, W.-k. Liao, A. N. Choudhary, S. R. Kalidindi, and A. Agrawal, "Establishing structure-property localization linkages for elastic deformation of three-dimensional high contrast composites using deep learning approaches," *Acta Materialia*, vol. 166, pp. 335–345, 2019.
- [33] A. Paul, P. Acar, W.-k. Liao, A. Choudhary, V. Sundararaghavan, and A. Agrawal, "Microstructure optimization with constrained design objectives using machine learning-based feedback-aware data-generation," *Computational Materials Science*, vol. 160, pp. 334–351, 2019.
- [34] G. B. Goh, N. Hodas, C. Siegel, and A. Vishnu, "Smiles2vec: Predicting chemical properties from text representations," 2018.
- [35] A. Paul, D. Jha, R. Al-Bahrani, W.-k. Liao, A. Choudhary, and A. Agrawal, "Chemixnet: Mixed dnn architectures for predicting chemical properties using multiple molecular representations," in *Proceedings of the Workshop on Molecules and Materials at the 32nd Conference on Neural Information Processing Systems*, 2018.
- [36] R. Sajeev, R. Athira, M. Nufail, K. J. Raj, M. Rakhila, S. M. Nair, U. A. Jaleel, and A. T. Manuel, "Computational predictive models for organic semiconductors," *Journal of Computational Electronics*, vol. 12, no. 4, pp. 790–795, 2013.
- [37] I. V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. A. Palyulin, E. V. Radchenko, N. S. Zefirov, A. S. Makarenko *et al.*, "Virtual computational chemistry laboratory—design and description," *Journal of computer-aided molecular design*, vol. 19, no. 6, pp. 453–463, 2005.
- [38] M. K. Riede, A. W. Liehr, M. Glatthaar, M. Niggemann, B. Zimmermann, T. Ziegler, A. Gombert, and G. Willeke, "Datamining and analysis of the key parameters in organic solar cells," in *Photonics Europe*. International Society for Optics and Photonics, 2006, pp. 61970H–61970H.
- [39] R. Olivares-Amaya, C. Amador-Bedolla, J. Hachmann, S. Atahan-Evrenk, R. S. Sánchez-Carrera, L. Vogt, and A. Aspuru-Guzik, "Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics," *Energy & Environmental Science*, vol. 4, no. 12, pp. 4849–4861, 2011.
- [40] "Chemaxon – software for chemistry and biology," <https://www.chemaxon.com/>, 2016, (Accessed on 10/15/2016).
- [41] O. A. von Lilienfeld, R. Ramakrishnan, M. Rupp, and A. Knoll, "Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties," *International Journal of Quantum Chemistry*, vol. 115, no. 16, pp. 1084–1093, 2015.
- [42] F. Weigend and R. Ahlrichs, "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: Design and assessment of accuracy," *Physical Chemistry Chemical Physics*, vol. 7, no. 18, pp. 3297–3305, 2005.
- [43] G. Barone, D. Duca, A. Silvestri, L. Gomez-Paloma, R. Riccio, and G. Bifulco, "Determination of the relative stereochemistry of flexible organic compounds by ab initio methods: conformational analysis and boltzmann-averaged gao 13c nmr chemical shifts," *Chemistry—A European Journal*, vol. 8, no. 14, pp. 3240–3245, 2002.
- [44] F. Chollet *et al.*, "Keras," 2015.
- [45] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.
- [46] G. Landrum, "Rdkit: Open-source cheminformatics," vol. 3, no. 04, p. 2012, 2006.
- [47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.
- [48] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480–1489.