# Prediction of novel charge transport layers for efficient perovskite solar cells

**Anthonipillai Jenestin Anton, Armadorou Konstantina Kalliopi, Kessler Raluca-Ana**

## Abstract

This project aims to identify novel small organic molecules as efficient hole (HTL) and electron transport layers (ETL) for their application in perovskite solar cells (PSCs) with high efficiency. A dataset including both HTLs and ETLs was created and divided into training and testing sets. Machine learning algorithms like Random Forest, XGBoost, and Chemprop were trained on the training set to predict power conversion efficiencies (PCEs) for the two different charge transport layers. Testing on the separate set yielded 63% efficiency for electron transport layers and 21% for hole transport layers. Overall, this study shows the effectiveness of machine learning for designing and evaluating the optoelectronic performance of novel organic materials, facilitating the development of more efficient, affordable and scallable PSCs along with other renewable energy technologies.

## 1 Introduction

Solar cells based on hybrid organic-inorganic lead halide perovskites (PSCs) have been attracting increasing attention since their conception back in 2009 [1], achieving a rapid increase in power conversion efficiency from 3.8% up to 25.8% [2]over the span of 15 years. The key factor for the rapid progress of the perovskite solar cell technologies stems from the unique properties of the perovskite material, such as its tunable bandgap, large absorption coefficient, long charge carrier diffusion lengths and high charge carrier mobilities [3–5]. The perovskite absorber is situated between two charge transport layers for each type of charge carrier, namely electrons and holes, which are known as electron transport layer (ETL) and the hole transport layer (HTL) respectively. These layers are carefully selected to to match the energy difference at the interface between the perovskite and the ETL or HTL, to ensure that the two types of carrier can be efficiently transported to the respective electrode contacts [6].

Consequently, the photophysical, chemical and electronic properties of ETL and HTL can influence considerably the performance of the final device. This issue has led to the development of a wide range of functional materials from various classes that can be employed as ETL or HTL [7–11] , including transition metal oxides, conductive polymers and small organic molecules, as it is shown in Fig. 1. [8]

However, the experimental methods for the development and fabrication of such materials, especially small organic molecules, often include multiple-step reactions, which negatively impact the time and cost of the synthesis. At the same time, their development is based on previous experimental results where the device exhibited high efficiency, without a predefined systematicity.

This trial-and-error process can be efficiently circumvented by the employment of Machine Learning (ML) models, in order to evaluate existing small molecules from publicly available databases or even probe the efficiency of new novel molecules, without having to synthesize them beforehands. Indeed, this approach has been implemented before in the field of perovskite solar cells, concenring the perovskite absorber itself or the parameters relevant to scale-up, but extensive studies related to the charge transport layers are scarcely reported [12–17].

AI4Chemistry course project at EPFL 2022. https://github.com/schwallergroup/ai4chem_course

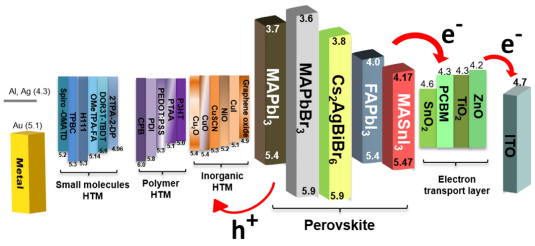


Figure 1: Most efficient materials employed as perovskite absorber, ETL and HTL.

## 2 Task

In this project, our objective is to find a connection between the molecular structures of small organic molecules acting as ETLs and HTLs and their properties with their performance in solar cells. By doing so, our aim is to develop a model capable of predicting the suitability of various small organic molecules from a given database as either HTLs or ETL, increasing in this way the number of available candidates which can be used.

An additional objective is the creation of an application that can facilitate the search process of ETL and HTL. By giving the name of the desired molecule, the application will give its corresponding SMILE representation, its 3D structure and its PCE. This particular application has the potential to help researchers and users to quickly access the information regarding specific molecules of interest. In the future, with further improvements, this application can become capable to predict efficient and novel HTL and ETL, thereby minimizing the time and resources required for synthesizing materials. This, in turn, would lead to a reduction in material waste and its impact on the environment.

## 3 Data

The dataset employed for the training and testing of our model was manually extracted from the current literature reviews concerning charge transport materials development [7–11,18,19] for the HTLs, as well as from the Perovskite Database [20] for the ETLs. The hyperparameters contained in the dataset include the name and the SMILES of each compound, the energy difference between the HOMO or LUMO of the compound and the valence band or conduction band of the perovskite respectively, as well as their electronic characteristics, such as $J_{sc}$, $V_{oc}$, FF, and PCE. The carrier mobility of the compounds was not included, because of incompletely reported data in the literature. Similarly, for the HTL, the effect of dopants was not explicitly calculated in our model, since the majority of the compounds include the same dopants to enhance their performance, namely Li-TFSI, TBP, FK209. For the ETL, where there was the possibility, most of the data was selected without dopants or hole blocking layers. In addition, for the ETL in the the Perovskite Database, it was observed that some blocking layers were erroneously referred to as ETLs.

## 4 Methods

Once the dataset was generated, it was divided into two different sets: a training set and a testing set. The propose of the training set is to train a machine learning algorithm, letting it to learn patterns and make predictions. On the other hand, the testing set is employed to measure the algorithm's performance, validating its efficiency and accuracy.

In this project, three different machine learning supervised algorithms were employed. Among these algorithms, two are based on a decision tree structure, namely Random Forest and XGBoost, while the third algorithm used is an artificial neural network called ChemProp. These algorithms were chosen thanks to their characteristics and advantages which are suitable for the propose of this project.

### 4.1 Decision tree based algorithms

The reason for using a decision tree based algorithm is because it considers various characteristics of the small organic molecules, such as the difference between the HOMO or the LUMO level of the small organic molecule, for HTLs and ETLs respectively, and the valence band or conduction band of the perovskite, the chemical structure etc. and makes the correlation between these features and the performance obtained with these molecules.

**Random forest:** This methods works by combining multiple decision trees and forming an ensemble model that can provide more accurate predictions. In this project, random forest was used to analyze the different features of the small molecules. Each decision tree in the random forest is trained on a random subset of the data and considers a random subset of features at each split. During prediction, the random forest accumulates the predictions from all the trees to produce the final regression result, being able to capture complex relationships between the different features of the small molecules and their performance[21].

**XGBoost:** This method works by sequentially training an group of decision trees and optimizing their collective performance. During the training process, XGBoost focuses on minimizing the loss function by constantly adding new decision trees to the group, where each tree corrects the mistakes made by the previous ones. In our case, by using the XGBoost algorithm on our data set, it is possible to train an ensemble of decision trees to predict the performance of our small organic molecules.[21]

### 4.2 Artificial neural network based algorithm

Artificial Neural Networks (ANNs) are a class of machine learning models which consist of interconnected nodes, called neurons, organized into layers. Each neuron takes inputs, applies a transformation, and then produces an output that is passed to the next layer. Graph Convolutional Neural Network (GCN) is a specific type of ANN architecture which is designed to work on graph-structured data, in our case molecular graphs. GCNs utilize the connectivity information of the graph to learn representations of nodes.

## 5 Results

A two-layer Graph Convolutional Network (GCN) was employed for PCE coefficient prediction, trained on a modest dataset (approximately 200 points). High Root Mean Squared Error (RMSE) of 31.375 indicates subpar performance, likely influenced by the constrained size of the training data and the simple model architecture. The absence of thorough hyperparameter tuning may have also limited the model's ability to accurately represent data complexity. These results emphasize the requirement for ample data, more complex architectures, and optimized hyperparameters in GCN implementations for supervised learning.

Chemprop model showed that given the small size of the dataset, the model achieved an R-squared value of 0.0398, explaining approximately 3.98% of the variance in PCE values within the test data. These results highlight the intricate nature of predicting PCE and the potential limitations of both the Chemprop model and the available dataset for this task.

In this project, we followed a structured approach, as it is shown in Fig. 2 in order to predict Power Conversion Efficiency (PCE) values from Simplified Molecular Input Line Entry System (SMILES) strings. This involved data preprocessing where SMILES strings were extracted and processed, followed by feature extraction where characteristics of the SMILES strings were manually obtained first and afterwards the SMILES with the descriptors were transformed using the RDKit toolkit. The data was then partitioned into training (60%), validation (20%), and test sets (20%). The feature were not normalized, as tree-based methods are not sensitive to the scale of the input features. Indeed, they are based on hierarchical decisions and not on distances or gradients. The Random Forest Regressor and XGBoost Regressor were trained and fine-tuned using the training and validation sets, respectively. The models performance was finally evaluated on the test set, providing a comprehensive understanding of their predictive capabilities.

The test scores (see table 1) show that Random Forest Regression performs better than XGBoost in this case. Random Forests tend to overfit less than XGBoost through the random subsetting of data and features, which promotes diverse learning across the ensemble of trees. Analyzing the test

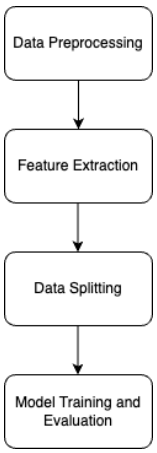


Figure 2: Machine Learning Pipeline for Power Conversion Efficiency Prediction

scores reveals that our best model, Random Forest Regression, tends to overfit, particularly on the HTL data. The HTL dataset shows less variation in PCE values compared to the ETL data. This lack of variation can make it hard for the model to capture the true patterns in the data and may cause overfitting, where the model performs well on the training data but not on new, unseen data. Looking at the $R^2$-Score for the XGBoost, a high $R^2$ on the training set and a negative $R^2$ on the test set is a strong indicator of overfitting. Overfitting is a modeling error that occurs when a function is too closely fit to a limited set of data points, so the model learns the noise in the training data, which negatively impacts the model's ability to generalize from the training set to unseen data in the test set.

| **Method** | **Data Set** | **$R^2$-Score** |
|---|---|---|
| XGBoost | HTL | -0.232 |
| Random Forest Regression | HTL | 0.205 |
| XGBoost | ETL | 0.498 |
| Random Forest Regression | ETL | 0.621 |

Table 1: R-Scores of different methods on test sets

## 6 Discussions

To be able to understand we looked at the descriptors which have the highest impact on our model and which are presented in the Table 2 below:

Table 2: The first three descriptors which influences the performance of the model. A higher score indicates a greater impact from the respective descriptor.

| Descriptor | Physical meaning | Score for HTLs | Score for ETLs |
|---|---|---|---|
| LabuteASA | Accessible surface area | **0.059377** | 0.003893 |
| EState_VSA3 | Intrinsic valence state | 0.047809 | **0.042932** |
| PEOE_VSA7 | Orbital electronegativity of atoms | 0.038060 | 0.004212 |

The meaning of the following descriptors is:

**LabuteASA** calculates the accessible surface area (ASA) of a molecule. The term refers to the surface surrounding a molecule, which is described as the hypothetical center of a solvent sphere in contact with the molecule's van der Waals surface. ASAs can vary in types, ranging from relative solvent accessibility to absolute surface areas[22,23].

**The PEOE (partial equalization of orbital electronegativity)** aims to address two issues: (i) accounting for the electrostatic interactions of charged atoms within a molecule, and (ii) estimating the

orbital electronegativities of atoms in a molecular environment based on the orbital electronegativities of free atoms. By evaluating the charge shifts in each bond individually, only neighboring atoms are considered in each cycle, leading to the prediction of connectivity-dependent atomic charges and residual electronegativities for each atom in a molecule. Remarkable correlations have been observed between the atomic charges calculated using PEOE and experimental measurements of charge-sensitive atomic properties, such as core electron binding energies (ESCA shifts) and resonance shifts in NMR [24,25].

**The electrotopological state (E-state) index** is a chemometric topological index that incorporates both electronic characteristics and the topological environment of each non-hydrogen atom in a molecule. By combining information about the atom's valence electronic structure and its presence in the molecular topology, the E-state index can identify atoms or molecular fragments that have a significant influence on molecular properties. In the molecular graph, each atom is assigned an E-state variable, which represents the atom's electronic state modified by the electronic influences from all other atoms in the molecule, taking into account the molecule's topological characteristics. As a result, the E-state assigned to a specific atom (or atom type) varies across different molecules and is contingent upon the detailed structure of the molecule [26,27].

Upon examining the presented data in the table, it can be seen that the primary descriptor for the HTL is the ASA (surface area), while for the ETL the main descriptor is EState. Considering the descriptors explanation, it can be concluded that even though the surface area (ASA) could be useful for our model, it is not the most crucial descriptor, and therefore should not be assigned the highest importance. In our case, the Estate and EPOP are more valuable as they make the correlation between the molecular structure and the electronic properties. For our materials, this correlation is critical because they have to assure the charge transport from the semiconductor to the electrodes. Thus, the difference between the descriptors' contribution on the computational model may be one of the reasons why the result for ETL are better than the ones for HTL.

## 7 Limitations and conclusions

A major limitation of our model stems from the initial dataset that was collected and on which the model was tested and trained on. As it is seen from the bar chart in Fig. 3 , while there are more datapoints for HTLs and both HTL and ETLs have similar PCE values, there is a difference in their distribution. More specifically, the ETLs exhibit an even distribution of their PCE values in the large 0-20% , while the performance of most HTLs ranges from 10 to 20%, thus creating discrepancies in the original data. Some other limitations of our model concern the fact that some highly efficient inorganic materials, such as the ETLs $TiO_2$ and $SnO_2$ were not included in the dataset. Moreover, the HTLs shared many similarities in their structure, such as the type of their core (carbazole, thiophene and furan cyclic structures), the substituents (triphenylamine and methoxy groups) or the overall shape of the molecule (dendrimer-shaped or star-shaped) [10]. It is also worth mentioning that most HTLs require the use of dopants, such as the ones mentioned in the Data section, which however were not explicitly included in our model. Finally, there are numerous cases in the literature where important data are not reported, especially as far as the energy levels and the hole mobilities of newly synthesized charge transport layers are concerned.

In conclusion, as illustrated in the "Results", our model was able to achieve quite a good result for the ETLs (63%), while for the HTLs it was less efficient (21%). It is likely that this variation in results can be attributed to the different in the descriptions' contribution, along with the limitations described above.

The implementation of our model for the prediction of the efficiency of small organic molecules as electron transport layers can help advance the field of solar technology. Indeed, by combining our model with large publicly available databases, such as PubChem or Zinc20, it is possible to perform fast PCE screening of potential novel ETLs. As a result, we aim to minimize the amount of time required to probe new materials in PSCs by the usual trial-and-error method, as well as decrease the amount of reagents and solvents required for extensive and arduous organic synthesis of multiple small organic molecules with various functionalizations. In this way, a new multidisciplinary field combining photonics and structure-property prediction can emerge. [28]

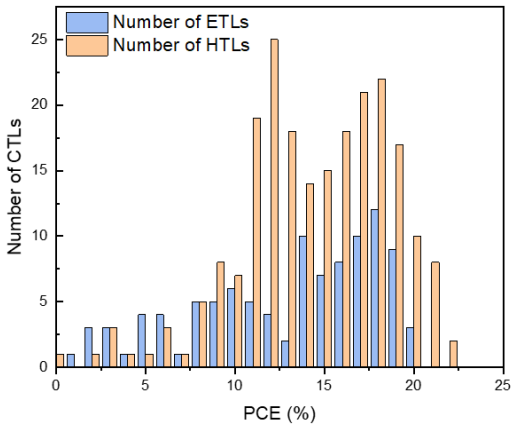


Figure 3: Distribution of HTL and ETL dataset according to PCE.

## References


[1] Akihiro Kojima, Kenjiro Teshima, Yasuo Shirai, and Tsutomu Miyasaka. Organometal halide perovskites as visible-light sensitizers for photovoltaic cells. *Journal of the American Chemical Society*, 131(17):6050–6051, 2009.

[2] NREL Best Research-Cell Efficiency Chart. https://www.nrel.gov/pv/cell-efficiency.html. Accessed: 01.06.2023.

[3] T. Jesper Jacobsson, Juan Pablo Correa-Baena, Meysam Pazoki, Michael Saliba, Kurt Schenk, Michael Grätzel, and Anders Hagfeldt. Exploration of the compositional space for mixed lead halogen perovskites for high efficiency solar cells. *Energy and Environmental Science*, 9(5):1706–1724, 2016.

[4] Christian Wehrenfennig, Giles E. Eperon, Michael B. Johnston, Henry J. Snaith, and Laura M. Herz. High charge carrier mobilities and lifetimes in organolead trihalide perovskites. *Advanced Materials*, 26(10):1584–1589, 2014.

[5] Guichuan Xing, Nripan Mathews, Shuangyong Sun, Swee Sien Lim, Yeng Ming Lam, Michael Grätzel, Subodh Mhaisalkar, and Tze Chien Sum. Long-Range Balanced Electron- and Hole-Transport Lengths in Organic-Inorganic $CH_3NH_3PbI_3$. *Science*, 342(6156):344–347, 2013.

[6] Istiak Hussain, Hoang Phong Tran, Jared Jaksik, Justin Moore, Nazmul Islam, and M. Jasim Uddin. Functional materials, device architecture, and flexibility of perovskite solar cell. *Emergent Materials*, 1(3-4):133–154, 2018.

[7] Anastasia Soultati, Apostolis Verykios, Konstantina-Kalliopi Armadorou, Marinos Tountas, Veroniki P Vidali, Kalliopi Ladomenou, Leonidas Palilis, Dimitris Davazoglou, Athanassios G Coutsolelos, Panagiotis Argitis, and Maria Vasilopoulou. Interfacial engineering for organic and perovskite solar cells using molecular materials. *Journal of Physics D: Applied Physics*, 53(26):263001, 2020.

[8] Zinab H. Bakr, Qamar Wali, Azhar Fakharuddin, Lukas Schmidt-Mende, Thomas M. Brown, and Rajan Jose. Advances in hole transport materials engineering for stable and efficient perovskite solar cells. *Nano Energy*, 34:271–305, 2017.

[9] Laura Calió, Samrana Kazim, Michael Grätzel, and Shahzada Ahmad. Hole-Transport Materials for Perovskite Solar Cells. *Angewandte Chemie International Edition*, 55(47):14522–14545, 2016.

[10] Anurag Krishna and Andrew C. Grimsdale. Hole transporting materials for mesoscopic perovskite solar cells-towards a rational design? *Journal of Materials Chemistry A*, 5(32):16446–16466, 2017.

[11] Wenxiao Zhang, Ying Chiao Wang, Xiaodong Li, Changjian Song, Li Wan, Khurram Usman, and Junfeng Fang. Recent Advance in Solution-Processed Organic Interlayers for High-Performance Planar Perovskite Solar Cells. *Advanced Science*, 5(7), 2018.

[12] Rishi E. Kumar, Armi Tiihonen, Shijing Sun, David P. Fenning, Zhe Liu, and Tonio Buonassisi. Opportunities for machine learning to accelerate halide-perovskite commercialization and scale-up. *Matter*, 5(5):1353–1366, 2022.

[13] Mahdi Hasanzadeh Azar, Samaneh Aynehband, Habib Abdollahi, Homayoon Alimohammadi, Nooshin Rajabi, Shayan Angizi, Vahid Kamraninejad, Razieh Teimouri, Raheleh Mohammadpour, and Abdolreza Simchi. SCAPS Empowered Machine Learning Modelling of Perovskite Solar Cells: Predictive Design of Active Layer and Hole Transport Materials. *Photonics*, 10(3):271, 2023.

[14] Shijing Sun, Noor T.P. Hartono, Zekun D. Ren, Felipe Oviedo, Antonio M. Buscemi, Mariya Layurova, De Xin Chen, Tofunmi Ogunfunmi, Janak Thapa, Savitha Ramasamy, Charles Settens, Brian L. DeCost, Aaron G. Kusne, Zhe Liu, Siyu I.P. Tian, Ian Marius Peters, Juan Pablo Correa-Baena, and Tonio Buonassisi. Accelerated Development of Perovskite-Inspired Materials via High-Throughput Synthesis and Machine-Learning Diagnosis. *Joule*, 3(6):1437–1451, 2019.

[15] Kate Higgins, Sai Mani Valleti, Maxim Ziatdinov, Sergei V. Kalinin, and Mahshid Ahmadi. Chemical Robotics Enabled Exploration of Stability in Multicomponent Lead Halide Perovskites via Machine Learning. *ACS Energy Letters*, 5(11):3426–3436, 2020.

[16] Qiuling Tao, Pengcheng Xu, Minjie Li, and Wencong Lu. Machine learning for perovskite materials design and discovery. *npj Computational Materials*, 7(1), 2021.

[17] Filippo De Angelis. The Impact of Machine Learning in Energy Materials Research: The Case of Halide Perovskites. *ACS Energy Letters*, 8(2):1270–1272, 2023.

[18] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019.

[19] Song Li, Yong Li Cao, Wen Hua Li, and Zhi Shan Bo. A brief review of hole transporting materials commonly used in perovskite solar cells. *Rare Metals*, 40(10):2712–2729, 2021.

[20] The Perovksite Database. https://www.perovskitedatabase.com. Accessed: 01.06.2023.

[21] XGBoost versus Random Forest. https://www.qwak.com/post/xgboost-versus-random-forest. Accessed: 01.06.2023.

[22] Tiejun Dong, Tong Gong, and Wenfei Li. Accurate Estimation of Solvent Accessible Surface Area for Coarse-Grained Biomolecular Structures with Deep Learning. *Journal of Physical Chemistry B*, 125(33):9490–9498, 2021.

[23] Syed Ali, Md. Hassan, Asimul Islam, and Faizan Ahmad. A Review of Methods Available to Estimate Solvent-Accessible Surface Areas of Soluble Proteins in the Folded and Unfolded States. *Current Protein & Peptide Science*, 15(5):456–476, 2014.

[24] Wilfried J. Mortier, Karin Van Genechten, and Johann Gasteiger. Electronegativity equalization: application and parametrization. *Journal of the American Chemical Society*, 107(4):829–835, 1985.

[25] M Marsili and J Gasteiger. Pi Charge Distribution from Molecular Topology and re Orbital Electronegativity. *Croatica Chemica Acta*, 53(4):601–614, 1980.

[26] Lowell H Hall and Lemont B Kier. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *Journal of chemical information and computer sciences*, 35:1039–1045, 1995.

[27] Electronic Supporting Information. https://www.rsc.org/suppdata/c7/me/c7me00094d/c7me00094d1.pdf. Accessed: 01.06.2023.

[28] Marcos del Cueto, Charles Rawski-Furman, Juan Aragó, Enrique Ortí, and Alessandro Troisi. Data-driven analysis of hole-transporting materials for perovskite solar cells performance. *The Journal of Physical Chemistry C*, 126(31):13053–13061, 2022.