Add-on for High Throughput Screening in Material Discovery for Organic Electronics: “Tagging” Molecules to Address the Device Considerations

A. Ashtiani Abdi, F. Nourmohammadian*, Y. Mohammadi, M. R. Saeb

1. Department of Organic Colorants, Institute for Color Science and Technology, P.O. Box: 16765-654, Tehran, Iran
2. Center of Excellence for Color Science and Technology, Institute for Color Science and Technology, P.O. Box: 654-16765, Tehran, Iran
3. Petrochemical Research and Technology Company (NPC-rt), National Petrochemical Company (NPC), P. O. Box: 14358-84711, Tehran, Iran
4. Department of Resin and Additives, Institute for Color Science and Technology, P.O. Box: 16765-654, Tehran, Iran

ARTICLE INFO

Article history:
Received: 9 Dec 2018
Final Revised: 14 Feb 2019
Accepted: 16 Feb 2019
Available online: 12 Mar 2019

Keywords:
Organic electronics
Bulk heterojunction solar cell
Machine learning
Artificial intelligence
Molecule tagging.

ABSTRACT

This work reflects the worth of intelligent modeling in controlling the nanostructure morphology in manufacturing organic bulk heterojunction (BHJ) solar cells. It suggests the idea of screening the pool of material design possibilities inspired by machine learning. To fulfill this goal, a set of experimental data on a BHJ solar cell with a donor structure of diketopyrrolopyrrole (DPP) and backbone of benzothiadiazole (BT) are fed into a home-written artificial neural network (ANN)/genetic algorithm (GA) hybrid code to optimize film-casting parameters. The annealing temperature, spin coating spin rate, and donor/acceptor ratio taken from available literature are applied to give the machine chance of learning trends in the power conversion efficiency (PCE). DPP-BT structures virtually born in the mind of machine are then screened for resemblance survey to receive a tag of desired characteristic. The results enable device manufacturers to identify the sensitivity of designed molecules to specific film casting conditions, while homologous structures may result in similar responses against design variables. Prog. Color Colorants Coat. 12 (2019), 107-120 © Institute for Color Science and Technology.

1. Introduction

While utilization of solar energy has opened a new era of technology [1], a particular attention was directed towards organic photovoltaic (OPV) devices [2], more particularly bulk heterojunction (BHJ) solar cells as promising photovoltaic (PV) with low cost and high applicability [3-6]. BHJ devices are dynamically under development thanks to their beneficial features such as flexibility, transparency, printability, and roll-to-roll processing [7]. Nevertheless, serious challenges still represent to design and manufacture of devices having adequate efficiency and durability; what was the reason behind recent advances in the field for developing new generations of solar cells [8-11].

BHJ device maturation experienced tremendous progress in recent years, classically through two experimental routes [12]; (i) Material: design and synthesis of high-performance light-harvesting materials [13, 14]. An active layer of an ordinary single-junction BHJ cell consisting of an inhomogeneous mixture of at least two polymeric or small molecule component with well-aligned energy levels. Such a photoelectronic character contributed from these materials enabled absorbing photons and producing excitons; (ii) Device: manipulation of device architecture and fabrication techniques [15-18], where...
Nevertheless, there have been bewilderments in the process, known as prospective study. In this sense, produced in retrospective ways in an intelligent phase separation of the materials within the active layer time could facilitate crystallization of P3HT to reach a manipulation of such parameters, e.g. for P3HT/PCBM wide variety of possibilities would be checked out by deterioration of PCE by annealing [31, 36]. Thus, a desired nanostructure, but some reports indicated sources of bias requires prolonged exposure to the [37].

imitation of PCE by blending theories and experiments manipulating device manufacture parameters by AA [20-23]. While the $V_{oc}$ is known to be the reflection of energy levels of donor and acceptor and $J_{sc}$ is determined by extend of light absorbed by the active layer, the FF can be affected by many factors, which complexity interact with each other [24]. One of those most influencing factors is morphology of the nanoscale structure that depends on the structural characteristics of the materials in the mixture. Therefore, for a given molecular component optimizing cell efficiency depends on the degree of control over nanostructure formed during film formation [25-28].

The effects of fabrication parameters on PCE of BHJ were the subject of intensive recent studies, where the viscosity of the solution, the type of solvent co-solvents as additives [29-31], donor to acceptor ratio [32], film casting technique [33] and treatments of the layers varying temperature and solvent annealing were recognized as the most influential factors [3, 34]. Nevertheless, there have been bewilderments in the manipulation of such parameters, e.g. for P3HT/PCBM component annealing at a specific temperature [35] or time could facilitate crystallization of P3HT to reach a desired nanostructure, but some reports indicated deterioration of PCE by annealing [31, 36]. Thus, a wide variety of possibilities would be checked out by manipulating device manufacture parameters by imitation of PCE by blending theories and experiments [37].

Precise estimation of an outcome to avoid/lessen sources of bias requires prolonged exposure to the process, known as prospective study. In this sense, there is a continued need for looking back to manipulate fabrication parameters for a desired outcome imitated earlier, known as a retrospective study [38]. To deal with such a huge amount of data produced in retrospective ways in an intelligent manner, experimental data should be integrated into unique patterns by the aid of artificial intelligence (AI), data mining and machine learning [39].

Artificial neural network (ANN) as prime techniques in AI has recently implemented in the field of organic electronics. The exclusive feature of this technique in the field of organic electronics is shown to model the properties that are not possible to explicitly model [40, 41], or helped to lower the computational cost of quantum mechanics and molecular dynamics [42-44].

Inspired from Quantitative Structure–Activity Relationships (QSAR) with their almost old history in designing drugs, a fresh line of thought has emerged recently for computer-aided researches in the field of organic electronics [45-48]. A growing body of literature has presented improvements in aspects of machine learning in photovoltaics mainly by high-throughput screening (HTS) [49, 50], from generating and handling numerous material structures in machine, digging technics and models [51-54], to delivering results [55-57]. However, all these studies has tended to focus on generate high-performance ‘material’ structures and do not address ‘device’ related reflections.

By pointing out the high variation in PCE of the benchmark device of P3HT/PCBM [58], Alan Aspuru-Guzik underlines that the current screening studies are directed toward filtering those ‘designed molecules’ with undesired electronic levels and do not deal with the big impact of device optimizations on PCE [55]. Therefore, complexities in sensitivity of PCE to fabrication parameters highlights a tremendous surge in the use of computer-aided approaches to address device considerations in the search for higher PCEs. Consequently, few AI studies have addressed fabrication PV devices. The combined use of artificial neural network (ANN) and genetic algorithm (GA) methods as an advanced prospective approach based on an in-house computer code appeared successful in anticipation of levels of fabrication parameters needed for maximization of both durability and PCE of co-sensitized dye solar cells with a very low error in the optimization of outcomes [59].

Looking at the HTS computational approach, we propose here a computer-aided protocol as add-on for screening designed materials for organic electronics to find target devise on the bedrock of machine learning concept. The term ‘tagging’ here has come to be used to refer the process of giving the identification cards to the molecules born in the mind of intelligent code were used to mimic the signature of an available device with optimized PCE. The donor:acceptor ratio, annealing temperature, and spin coating spin rate are selected as
typical influential parameters governing the PCE of fullerene BHJ solar cells [36]. A robust computer code was developed to imitate and mimic precisely and very quickly fluctuations in the performance parameters until reaching the best solution (highest PCE) with minimum error. The main outcome of the optimization would be the ability to predict the sensitivity of PCEs of similar designed DPP-BT structures to fabrication parameters, thereby to add new information/patterns about film casting signature of the resulting structures. A paradigm shift from prospective to retrospective analyses or vice versa in the field of PV device makes possible identifying molecular features on the performance of solar cells and allows for designing active layers very efficiently.

2. Model Development

Artificial intelligence approaches were implemented to unveil the complex nature of the relationship between different kind of problems of solar cells [39, 60]. Blending ANN with GAs methods, an in-house computer code developed and implemented in this work. The use of ANN and GAs approaches manifests construction of a mathematical model and optimization of the problem, respectively, towards a set of solutions satisfying the predetermined targets.

We had recently new class of donor structures based on DPP-BT backbone under scrutiny with potential efficiencies of 5.5 to 6.8 % in fullerene BHJs and 8.5 to 10.5 % for non-fullerene (NF) BHJs [61]. To study the behavior of these molecules by means of the proposed AI technique of this research, we need an experimental dataset of another DPP-BT based molecule to feed into the model. The most important factors for choosing the dataset in such this study, regardless of the extent of its efficiencies and performance, is the quantity of data points and the similarity of the two designed and reference molecular backbones. The experimental data point has been used from BT(TDPPPertTTC6)s:PC71BM devices from reference [36]. This dataset consists of 27 different film casting conditions on a molecule with a backbone of DPP-BT by full factorial experimental design with three levels of three intended casting variables; annealing temperature, donor: acceptor ratios and spin coating spin rate.

The data were first classified and nominated as input variables (x₁ to x₉) and targets (y₁ to y₄). For convenience, the donor to acceptor ratio was used in the form of donor content, i.e. 0.3 donor content is equivalent of donor to acceptor ratio 30:70. Since the multiplication of y₁, y₂, and y₃ provides a practical sense, it is defined as y₄ (PCE). The developed code identifies each original data used for training the ANN and tests the rest of the data as a “scenario”. Following each step, the program considers them as "experiment"; however, in the end, it reports the best solution in accordance with the initial scenario corresponding to the experimental runs.

For deep learning the data, we implemented the multi-layer perceptron (MLP) ANN technique. Since the ANN model cannot principally optimize the problem, it was combined with GAs enabling it for optimization and determination of the unknown parameters of the network (weights and biases). In doing so, a powerful hybrid tool was achieved enjoying a very high level of accuracy and the complexity of the problem will be controlled by the excellence of developed code with high potential for learning.

To simultaneously adjust J_sc, V_oc, and FF criteria, Non-dominated Sorting Genetic Algorithm-II (NSGA-II) was implemented [62, 63]. This algorithm is the same as the single objective, though with a different mechanism for sorting and choosing the best chromosomes.

Figure 1 demonstrates sequential steps considered in developing the computer code, as well as the structure of chromosome defined for prediction of J_sc, V_oc, FF, and PCE criteria. The detailed information about single-objective and multi-objective optimization parts of our code, the ANN architecture and explanation of how the number of layers and nodes are defined, the type of activating function and method to preventing the network to over fitting can be found in our previous publications [64, 65].

The program was written in PASCAL programming environment (Lazarus IDE) and FPC 2.6.2. was used as compiler. Modeling was then performed on a desktop computer with Intel Core i7-3770K (3.50 GHz), 32 GB of memory (2133 MHz), and under Windows 7 Ultimate 64-bit operating system.

According to the report on error analyses and network statistics of the model presented in Table 1, the maximum 1 % training errors were obtained for responding variables with regard to the selected inputs by the model. This confirms that, in spite of the low number of data inputs, the model is capable to correctly converge the error during gradient descent process.
Table 1: Error analyses the performance of the model in training and testing the data points.

<table>
<thead>
<tr>
<th>Performance</th>
<th>$y_1 (J_{sc})$</th>
<th>$y_2 (V_{oc})$</th>
<th>$y_3 (FF)$</th>
<th>$y_4 (PCE_{SO})^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Test</td>
<td>Training</td>
<td>Test</td>
</tr>
<tr>
<td>Training MSE</td>
<td>0.0004</td>
<td>0.0398</td>
<td>0.0004</td>
<td>0.0200</td>
</tr>
<tr>
<td>Training Error (%)</td>
<td>0.9999</td>
<td>9.9776</td>
<td>0.9931</td>
<td>7.0660</td>
</tr>
<tr>
<td>Maximum Error (Scenario)</td>
<td>11</td>
<td>18</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>Correlation Coefficient</td>
<td>0.9784</td>
<td>0.9887</td>
<td>0.9841</td>
<td>0.9925</td>
</tr>
<tr>
<td>R-Squared</td>
<td>0.9573</td>
<td>0.9775</td>
<td>0.9684</td>
<td>0.9851</td>
</tr>
<tr>
<td>Coefficient of Efficiency</td>
<td>0.9568</td>
<td>0.9771</td>
<td>0.9679</td>
<td>0.9847</td>
</tr>
<tr>
<td>Goodness of Fit (%)</td>
<td>79.22</td>
<td>84.88</td>
<td>82.07</td>
<td>87.61</td>
</tr>
</tbody>
</table>

* PCE single objective

Figure 1: Flowchart of the training, fitting and reporting process of the model along with the second phase including multi-objective optimization.
3. Results and Discussion

3.1. PCE: Single and Multi-objective studies

According to the corresponding model, there are three methods to calculate the power conversion efficiency. The first one is the single objective (PCE$_{S.O.}$) modeling of the 27 PCE data points, which are produced by the multiplication of the experimental $J_{sc}$, $V_{oc}$ and FF values in datasets. Statistical coefficients in Table 1 show that the model is more capable to fit the trend of this parameter ($y_4$) compared with the other three factors ($y_1$ to $y_3$). It means uncertainties about fitting the model to components of PCE is reduced by multiplication.

Another method to evaluate the theoretical final efficiency of a cell is to multiply the predicted values of modeled $J_{sc}$, $V_{oc}$, and FF (PCE$_p$). Figure 2 A and B demonstrates the preferred ranges of inputs to reach higher values of PCE based on single objective modeling (PCE$_{S.O.}$) and product of modeled efficiency components (PCE$_p$). Both plots show identical region to reach high PCE values (more than 0.69). Moreover, the correlation ($r=0.890$) between PCE$_{S.O.}$ and PCE$_p$ is noteworthy as depicts that the model is reliable to predict the PCE either by firstly modeling the $J_{sc}$, $V_{oc}$, and FF or single objective modeling the PCE directly (Figure 2 C).

Figure 2: Illustrations of two methods of evaluating the PCE, single objective prediction from the ANN model (A), product of modeled components of the PCE (B), correlation of the results of these two methods (C).
The third method, the multi-objective optimization to find the PCE is constructed according to the highest possible values for each of the three outputs ($y_1$ to $y_3$) in order to find the Pareto points. Figure 3 A shows that many compositions of the obtained outputs can be considered as Pareto points. However, it is evident that the intended points are those that maximize the product of the three outputs (i.e. $PCE_{M.O.}$). Figure 3 B shows that the obtained PCE is higher for the maximum values of both $J_{sc}$ and FF and the optimized medium range of $V_{oc}$. The capability of obtaining a single PCE by multiple recipes can be used as a controlling tool aimed at lowering the fabrication costs. By defining a manufacturing figure of merit (FOM) like equation 4, it is possible to rank the suggested recipes.

$$FOM = \frac{PCE}{Annealing\, temperature \times spin\, rate}$$ (4)

To address the donor:acceptor ratio, in case of HTS where the synthetic complexity (SC) is a parameter to filter the designed structures [66, 67], it is possible to practice SC in FOM as an effective parameter related to the final price of the device.

Table 2 shows the recipes suggested by the three methods of modeling the PCE, highest achieved PCE from the 27 experimental data points ($PCE_{exp}$) and the recipe with highest FOM within results of multi-objective optimization and experiments. It is important to make sure that the machine truly learns the experimental data from the database when it comes to comparing the designed structures in HTS with the previously examined molecules. Comparing the individual parameters in modeled $PCE_{M.O.}$ with the $PCE_{exp}$ shows that the model can readily learn from the limited number of experimental data and slightly improve it. It should be notices that sometimes round-off toleration may cause an optimum point lie somewhere outside the studied range of variables, but here optimum data are as closely as possible in the neighborhood of maximum possible levels, successfully done by the smart code.

<table>
<thead>
<tr>
<th>Table 2: Recipes suggested by the methods of modeling the PCE aimed at maximizing cell efficiency or lowering the manufacturing cost.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
</tr>
<tr>
<td>-------------------------------</td>
</tr>
<tr>
<td>$PCE_{S.O.}$ (maximizing PCE)</td>
</tr>
<tr>
<td>$PCE_{P}$</td>
</tr>
<tr>
<td>$PCE_{M.O.}$ (maximizing $J_{sc}$, $V_{oc}$ and FF)</td>
</tr>
<tr>
<td>$PCE_{exp}$ 27 points of experimental dataset</td>
</tr>
</tbody>
</table>

Figure 3: Pareto front for multi-objective optimization aim at highest possible values for $y_1$ to $y_3$. |
A remarkable result to emerge from Table 2 and in agreement with correlation in Figure 2 C, is come from comparing the PCE_{S.O.} and PCE_{p}. Training ANN with \( J_{sc}, V_{oc}, \) and FF as outputs is important for both: 1) delivering the sensitivities as far as machine learning concerns. More details on this will be given in the next section, and 2) learning the trend of these parameters to variations of inputs. More details on this topic will disclose in our upcoming article [68]. Consequently, while the individual optimization of the \( J_{sc}, V_{oc}, \) and FF is important, the close result from trained ANN with PCE is beneficial for supervising the reliability of trained neural networks to make sure that they are not over-fitted and not learned the noises. On the other hand, in case of training larger databases, training the model by PCE is significant when some data points (\( J_{sc}, V_{oc}, \) or FF) are missed or for crosschecking the validity of imported data into the database by comparing the PCE_{S.O.} and PCE_{p}.

### 3.2. Sensitivity analysis

The effect of each three input variables on the outputs can be described using sensitivity analysis [69]. Here we implemented the one-factor-at-a-time (OTA) method that means the extent of possible change of an output per one input, while keeping the other two inputs constant. A simple form of this test is computing the percentage of improvement of a specific response compared to the lowest result (base level) when sweeping one of the inputs. Keep in mind that while the continuous space of ANN model of responses let us sweeping the variables in fine mesh grids, in the full factorial experimental design where the levels of each variable are limited (there are only three levels in this case), such this analysis will result in rough outcomes. In a preliminary study of results, the most obvious point is the \( V_{oc} \), which takes a much lower effect by the three inputs (Figure 4). It is in agreement with the theory stating that the \( V_{oc} \) is mainly a function of electronic states of active layer materials (donor and acceptor).

While this figure demonstrates an estimate of the ultimate potential of each input to improve the value of outputs, but it cannot address the direction of these variations. A sensitivity analysis for the purpose of BHJ should deliver meaningful insight into the capability of changing film casting parameters to make either positive (+) or negative (-) influences on efficiency. Moreover, it is sensible to choose the mildest film casting conditions as base levels or reference. Moving base level from minimum output value to the mildest condition will make a distinction between positive and negative impacts. In view of machine learning and using this parameter in screening practices, the sensitivity of outputs e.g. PCE to inputs of different natures (temperature, time, rate, etc.) needs to be dimensionless. Equation 5 employed as a sensitivity analysis in this study.

![Figure 4: Sensitivity of responses to the input variables.](www.SID.ir)
The base levels for inputs ($x_0$) are chosen to be the lowest spin rate of spin coater, lowest temperature for post-annealing and lowest donor content. The reference values for outputs ($y_0$) are corresponding response to the $x_0$. Whilst the sign of the denominator in equation (5) is always positive, as all the $x_i$ values are greater than the base level and the percent of change in output. Therefore, the sign of sensitivity can be positive or negative. In this way, two surfaces for positive and negative sensitivity can be obtained in a 3-dimensional space of sensitivity of an output ($J_{sc}$, $V_{oc}$, FF, PCE_{s.o.} and PCE_{p}) to the variation of one of the inputs (z-axis), and the other two inputs as x and y-axis.

Figure 5 shows sensitivities for PCE_{p} to the inputs. By the first glances at Figure 5, it is obvious that span of improvement of PCE by optimizing the spin rate and annealing temperature is not as wide as the positive impact of changing donor content.

The individual outcomes of sensitivity analysis can be interpreted visually from box-and-whisker plots [70] in Figure 6. This kind of plot together with data point’s presentation is useful to distinguish distributions of data points and visually judge about outliers data in datasets. As expected, broad span of sensitivities in Figure 6 A demonstrate that the donor:acceptor ratio is the most influential parameter in optimization of PCE. It is apparent from Figure 6 B that for this specific BHJ cell, increasing the spin rate during casting film of active layer on the substrate will not put any negative impact on the final cell functions. Figure 6 C shows a clear controversial trend in sensitivities of $J_{sc}$/FF and $V_{oc}$ by increasing annealing temperature. Correlations between positive and negative sensitivity of PCE_{p}/PCE_{s.o.} and $J_{sc}$ and FF is in accordance with Figure 4 and confirms the synergic governing effect of $J_{sc}$ and FF on PCE.

$\text{Sensitivity} = \frac{\text{Percent of change in output}}{\text{percent of change in input}} = \frac{y_i - y_0}{y_0} \frac{x_0}{x_i}$ (5)

Figure 5: 3-dimensional surfaces of positive and negative sensitivities of PCE to donor:acceptor ratio (A), spin rate (B) and annealing temperature (C). The color bar is denoting sensitivity level.
Figure 6: Box plots together with the data point's presentation of sensitivities. Upper and lower bounds of boxes and the vertical line inside the boxes are locating 75%, 25% and median of data respectively. The small square is the mean value. The whiskers are the lines extended from the quartiles to 0.5% and 95% of data. The star points are the lowest and highest values in the data points.
Figure 7 shows the three parameters that can be employed to compare the parameters regarding their impacts on outputs; average, extremum (maximum for positive sensitivity, minimum for negative sensitivity) and positive:negative ratio. The latter parameter is calculated in a way to show the direction of the overall impact. That is absolute value of \((\text{positive sensitivity})/\text{negative sensitivity}\)^n which n is 1 when the positive effect is more than negative effect and n is -1 and the sign of ratio is set to be negative.

Returning to the question posed at the beginning of this study, it is now possible for machine to address questions like ‘which factor is the most influencing one to improve the PCE?’ by sorting the sensitivities to the input variables. Despite the simplicity of delivering sensitivity analysis by OAT, but in case of J-V results of photovoltaics, there are interactions between $J_{sc}$, $V_{oc}$, and FF. In other words, positive sensitivity in one of them does not mean positive sensitivity of efficiency. Therefore, it is needed to take the PCE as the sole parameter for screening the input variables. Accordingly, for the cells with donor molecules with structures similar to BT(TDPP$_{EH}$TTT)$_2$ and PC$_{71}$BM as acceptor, it is possible to attach ‘tag’s like these to the designed DPP-BT structure:

- Unlike normal expectations, post thermal annealing the cells will deteriorate the efficiency.
- Focus on changing the donor to acceptor ratio. However, be cautious, it is a double-edged sword. Changing the donor:acceptor ratio at some points will show great positive impacts (greatest between all three input variables) on PCE, but it also can show equally same great negative impacts.

- Feel free to increase the spin rate. Higher spin rates during casting active layer by spin coater will have a safe increasing effect on the cell efficiency.

Although the simple ranking of the sensitivity parameters could easily deliver conclusions on the importance of the input variables, for the bigger datasets it is needed to employ more efficient decision-making techniques [71].

4. Concluding remarks

The artificial intelligence study in this work showed the hybrid ANN-GA as a reliable model to evaluate the trend of BHJ working factors ($J_{sc}$, FF, $V_{oc}$, and PCE) by changing the cell fabrication parameters. We showed how to deliver an exact sensitivity analysis by interpolating in a three-dimensional space of discrete points of experimental data limited to three levels of three variables. The model architecture has shown a decent performance to predict the complex behavior of a BHJ efficiency.

However, the very important outcome of this study is showing the capabilities of AI techniques in examining device aspects of organic electronics. In this regard, we suggest the next directions towards completing the puzzle of addressing the device considerations in organic photovoltaics:
- The first of first step is to construct a thorough database of experimental database including all the aspects of device recipe manufacturing variables and resulting device related characteristics. This database will be bigger and more comprehensive compared with the current databases [72, 73].

- By the concept of molecular fingerprint or other machine-readable parameters, it is possible to split the data based on their similarities. There are valuable manuscripts to read about it [53, 74].

- Then it is time to use the ANN/GA model of this study to train the model by the data from the database.

- Finally, it is time to import the proposed data from HTS studies [72, 75]. The model is ready to deliver the comprehensive information around the device manufacturing operations, sorting by the final price, ease of synthesis, sensitivity of PCE.

5. References


44. F. Pereira, K. Xiao, D. A. Latino, C. Wu, Q. Zhang, J. Aires-de-Sousa, Machine learning methods to predict density functional theory B3LYP energies of HOMO
47. A. A. White, Big data are shaping the future of materials science, *MRS Bull.*, 38(2013), 594-595.
72. S. A. Lopez, E. O. Pyzer-Knapp, G. N. Simm, T. Lutzow, K. Li, L. R. Seress, J. Hachmann, A. Aspuru-


How to cite this article: