molecules

MDPI

Article

Energy Level Prediction of Organic Semiconductors for Photodetectors and Mining of a Photovoltaic Database to Search for New Building Units

Jehad Saleh [1], Sajjad Haider [1], Muhammad Saeed Akhtar [2], Muhammad Saqib [3,*], Muqadas Javed [3], Sayed Elshahat [4] and Ghulam Mustafa Kamal [3]

[1] Chemical Engineering Department, College of Engineering, King Saud University, P.O. Box 800, Riyadh 11421, Saudi Arabia
[2] School of Chemical Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea
[3] Institute of Chemistry, Khwaja Fareed University of Engineering & Information Technology, Rahim Yar Khan 64200, Pakistan
[4] Physics Department, Faculty of Science, Assiut University, Assiut 71516, Egypt
* Correspondence: muhammad.saqib@kfueit.edu.pk

Abstract: Due to the large versatility in organic semiconductors, selecting a suitable (organic semiconductor) material for photodetectors is a challenging task. Integrating computer science and artificial intelligence with conventional methods in optimization and material synthesis can guide experimental researchers to develop, design, predict and discover high-performance materials for photodetectors. To find high-performance organic semiconductor materials for photodetectors, it is crucial to establish a relationship between photovoltaic properties and chemical structures before performing synthetic procedures in laboratories. Moreover, the fast prediction of energy levels is desirable for designing better organic semiconductor photodetectors. Herein, we first collected large sets of data containing photovoltaic properties of organic semiconductor photodetectors reported in the literature. In addition, molecular descriptors that make it easy and fast to predict the required properties were used to train machine learning models. Power conversion efficiency and energy levels were also predicted. Multiple models were trained using experimental data. The light gradient boosting machine (LGBM) regression model and Hist gradient booting regression model are the best models. The best models were further tuned to achieve better prediction ability. The reliability of our designed approach was further verified by mining the photovoltaic database to search for new building units. The results revealed that good consistency is obtained between experimental outcomes and model predictions, indicating that machine learning is a powerful approach to predict the properties of photodetectors, which can facilitate their rapid development in various fields.

Keywords: machine learning; energy levels prediction; semiconductor photodetectors; regression models; pearson correlations

Citation: Saleh, J.; Haider, S.; Akhtar, M.S.; Saqib, M.; Javed, M.; Elshahat, S.; Kamal, G.M. Energy Level Prediction of Organic Semiconductors for Photodetectors and Mining of a Photovoltaic Database to Search for New Building Units. *Molecules* **2023**, *28*, 1240. https://doi.org/10.3390/molecules28031240

Academic Editors: Muhammad Khalid and Muhammad Nadeem Arshad

Received: 11 January 2023
Revised: 24 January 2023
Accepted: 24 January 2023
Published: 27 January 2023

1. Introduction

The world is a place of discovery and billions of devices containing multiple sensors have been commercialized. Photodetectors or photosensors primarily work as optical receivers for the conversion of light into electrical signals. The photodetector has become a vital part of modern devices with a broad range of applications, including environmental monitoring, optical communication, health monitoring, image sensing, defense system and for safety purposes in industries [1]. In the modern age, silicon (Si), germanium (Ge), and indium gallium arsenide (InGaAs)-based inorganic photodetectors (PDs) have been popular in the market due to their stable performance, high quantum efficiency, high sensitivity/detectivity, response speed or responsivity. Despite having several advantages,

inorganic photodetectors have limitations that cannot be ignored, such as complex fabrication procedures, manufacturing cost (e.g., requirement of high vacuum environment, high processing temperature and complex growth) and brittleness [2,3]. Moreover, inorganic photodetectors are rigid in nature, which limits their utilization in flexible environments or applications [4,5].

In recent years, organic photodetectors have emerged as a promising alternative for inorganic photodetectors [6]. These show various superiorities, including operation with a cooling-free effect, compatibility with flexible devices, detectable spectral response, and easy processing, which makes them potential candidates in wearable electronics. The spectral response can be achieved by altering the material used for organic semiconductors in the case of organic photodetectors (OPDs) [7,8]. In comparison, to inorganic semiconductors, organic semiconductors are composed of carbon-based molecules, which makes the organic photodetectors more environmentally friendly and biocompatible. In addition, various polymers and small molecules have been used for organic semiconductors [9,10]. These compounds generate a positive impact on electronic and optoelectronic devices [11]. Much progress has been made in recent years in the development of organic semiconductor devices [12].

A narrow bandgap is required for high-performance organic semiconductor photodetectors [13–15]. For decades, semiconductors with a large bandgap have been utilized. When the light of high energy falls on the target material, the excitons generated in donor front layers are unable to separate into free charges properly. Only low-energy photons with the power of long penetration depth can reach the donor-acceptor interface and then successfully generate the free charge. Therefore, it is an area of ongoing interest to predict energy levels of organic semiconductors for photodetectors and mining of the photovoltaic database to search for new building units. Traditionally, new designs can be achieved by utilizing the knowledge gained from laborious and multistep synthesis procedures, expensive device optimization and characterization. However, these trial-and-error methods do not guarantee success in the end. Moreover, it is hard to predict the performance of the materials before performing expensive experimental work. To this end, computer-aided material designing, discovery, and screening are of utmost importance.

Computational science is a popular field of science, which can be effectively applied to solve the complex problems of various systems and to finally find the solutions for such scientific problems [16–18]. Computational methods can analyze, screen, and predict data through mathematical algorithms. These methods can be applied to various fields of science [19–22]. During the past decade, researchers have been focusing on developing predictive computer models. These models can help them to analyze challenging problems [23]. Machine learning is a modern research tool [24]. Machine learning analysis (MLA) is based on pattern recognition by reducing the size of data and those parameters which can be learned by the computer. In machine learning, the results can be obtained by analyzing previously reported studies [25]. Moreover, several properties can be studied without understanding the chemistry or physics behind these properties [26]. The recent advancement in MLA includes the successful prediction of properties of the materials, materials discovery, drug development and material designing. Molecular fingerprinting and similarity analysis is now a common feature. Various algorithms can be used to train the models to obtain good accuracy. Machine learning can be used to predict the values of the highest-occupied molecular orbital (HOMO) and lowest-unoccupied molecular orbital (LUMO and power conversion efficiency (PCE) with accuracy. In recent studies, MLA has been used to design efficient molecules for organic photovoltaic (OPV) applications, and the structures of these molecules pass through subsequent successful experimental testing. Efficient organic semiconductor materials have been designed by using this approach, which accelerates the developing process in a time-saving manner [27].

Herein, a machine learning-based approach was applied to predict the energy level of organic semiconductors for photodetectors. Multiple models were trained, and their respective parameters were adjusted. As a result, the models with the highest accuracy

were chosen for conducting further analyses. Moreover, detailed data about energy levels (HOMO and LUMO) were visualized to show the trend (hidden pattern) in the data. The parameter's feature importance was also evaluated for training machine learning models. In addition, Pearson correlation and Shapiro ranking was applied to demonstrate the correlation between different parameters. A similarity analysis was performed to find the similarities between reference structure and structure in the database. Furthermore, mining of the photovoltaic database was used to search for new building units.

2. Results and Discussions

The performance of varied materials depends on their chemistry [28–30]. Chemical data can help to understand their behavior [31,32]. The hidden pattern of data can provide much useful information [33].

2.1. Molecular Descriptors

Molecular descriptors are the mathematical representation of the molecules used to train the models based on machine learning (Table 1) [34–36]. Molecular descriptors can be generated with the help of different algorithms. These can be derived from the chemical structure of the molecules. Physical and chemical properties or information can be described quantitatively with the help of the numerical value of molecular descriptors [37,38]. Almost thousands of molecular descriptors were calculated. Then, these were shortlisted in several unique ways. Molecular descriptors are based on independent properties. They can help researchers to perform similarity tests by using different models such as RDkit in molecular libraries. Based on the similarities present in the descriptor values, the molecules with the same physical and chemical properties can be evaluated.

Table 1. Molecular descriptors with respective categories and descriptions.

No	Molecular Descriptor	Category	Description
1	SM5_X	2D matrix-based descriptors	Spectral moment of order 5 from chi matrix
2	RCI	Ring descriptors	Ring complexity index
3	nR05	Ring descriptors	Number of 5-membered rings
4	RFD	Ring descriptors	Ring fusion density
5	NNRS	Ring descriptors	Normalized number of ring systems
6	DECC	Topological indice	Eccentric
7	ETA-D-AlphaB		Eta delta alpha b index
8	SdssC	Atom-type E-state indices	Sum of dssC E-states
9	SpAD_AEA(dm)	Edge adjacency indices	Spectral absolute deviation from augmented edge adjacency mat. weighted by the dipole moment
10	TI2-LN	2D matrix-based descriptors	Second Mohar index from Laplace matrix

2.2. Regression Analysis

The performance of machine learning is strongly dependent on algorithms [39]. To identify which kind of variable has a strong effect on the topic of interest, these methods are effectively reliable. The regression analysis provides information on the way these factors influence each other, how to determine the factors that are of most importance, and which factors can be ignored. Regression analysis uses various algorithms of machine learning. The data can be integrated into two parts: the testing set and the training set. These data are of different ratios, 70% 30%, 60% 40%. The best one is training: test ratio. Afterwards, by analyzing the values of predicted PCE and experimental PCE, the correlation between these two was calculated. The obtained results are plotted in the form of a graph.

2.3. Pearson Ranking Correlation

In machine learning algorithms, Pearson correlation is the most widely used correlation for numerical variables. To effectively measure the degree of relationship (linear association or correlation) between two different variables, this correlation can be used. It shows how far the data points are from the line of best fit. For this method to work effectively, the variables should be normally distributed. The direction and strength of two variables can be measured by the number between 1 and −1, where −1 indicates negative correlation, 1 represents positive correlation, and 0 indicates no correlation.

2.4. HOMO Prediction

The value from 1 to 0 shows a positive correlation between the HOMO and other molecular descriptors. The value from 0 to −1 shows the negative correlation between HOMO and other molecular descriptors. The 0 value shows no correlation between the variables. The red color indicates a positive correlation, while blue color indicates a negative correlation. As shown in Figure 1, DEEC, SdssC, NNRS, Nr05, ti2-L, and SM5-X all are red in color, these molecular descriptors show a positive correlation with the HOMO lies on the x-axis. In contrast, SPMAD-AEAdm, RFD, RCI and Eta-D-AlphaB show a negative correlation with the HOMO of the x-axis. In addition, TI2-L shows a strong positive correlation with the HOMO lying on the y-axis.

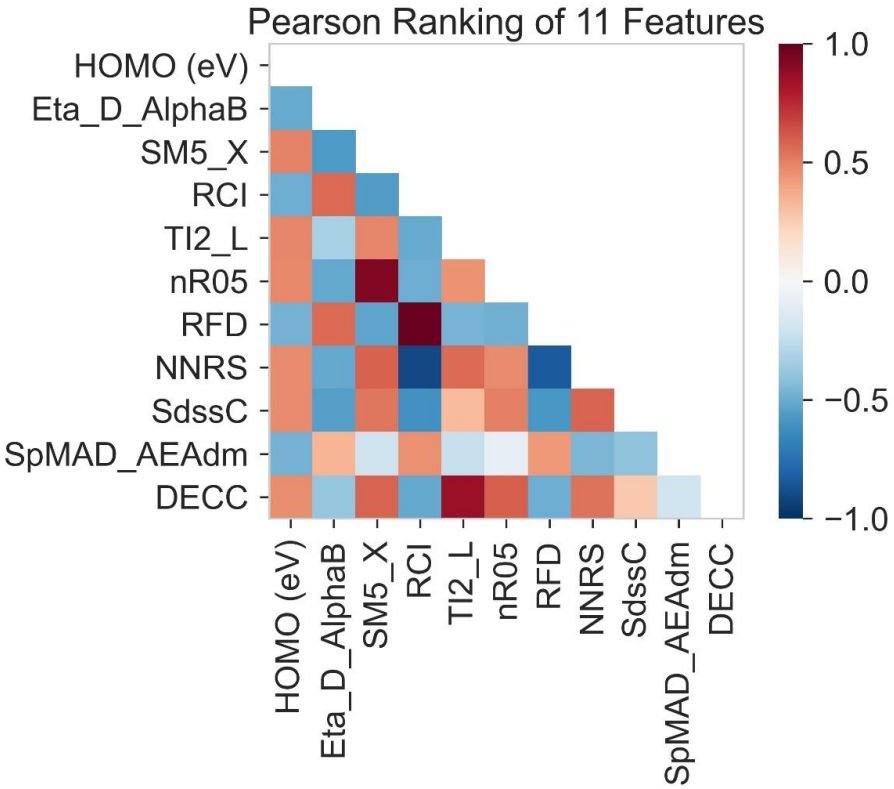

Figure 1. Correlation between HOMO and molecular descriptors.

For model training, calculated molecular descriptors are used as a source of input. The chemistry of donor molecules is represented by the help of molecular descriptors. The numerical form is used for presenting the chemical structure of materials in the numerical form presented by the molecular descriptors. Figure 2 shows that the Eta-D-AlphaB, RCI, RFD, and SPMAD-AEAdm show the negative dependent Pearson correlation. On the other hand, other molecular descriptors such as DECC, SdssC, NNRS, Nr05, TI2-L, and SM5-X show that they are positively dependent features.

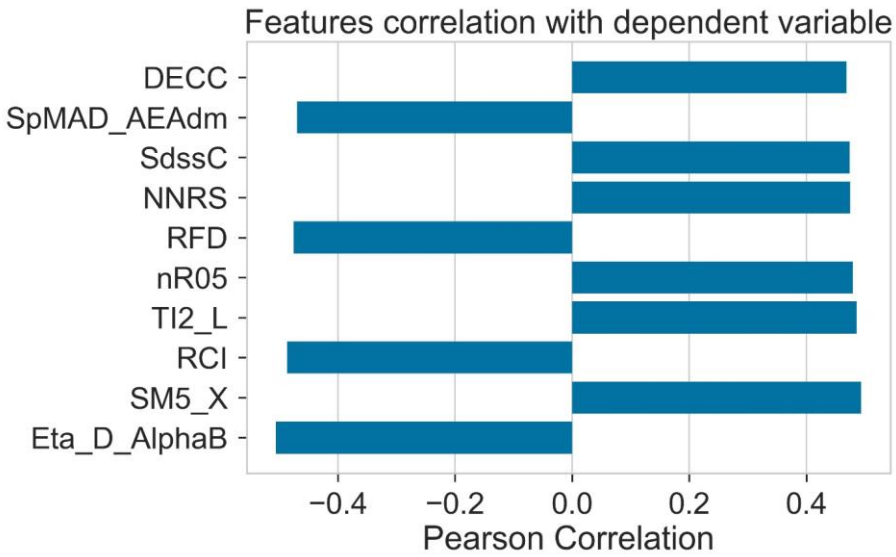

Figure 2. Correlation of features with HOMO.

2.5. Shapiro Ranking

The Shapiro ranking test is also known as the Shapiro–Wilk test. This test is usually performed to test the normality in statistics. Martin Wilk and Samuel Sanford Shapiro published this in 1965. This ranking uses the Shapiro–Wilk algorithm, which is generated by the Yellowbrick Python package. This is a one-dimensional feature ranking. To assess the normality of distribution, it considers a single feature at a time. The results are shown in the form of a bar plot, which shows the features with the maximum score on the one side and the features using the average score on the other side. Figure 3 shows that Eta-D-AlphaB shows the least amount of distribution according to this ranking.

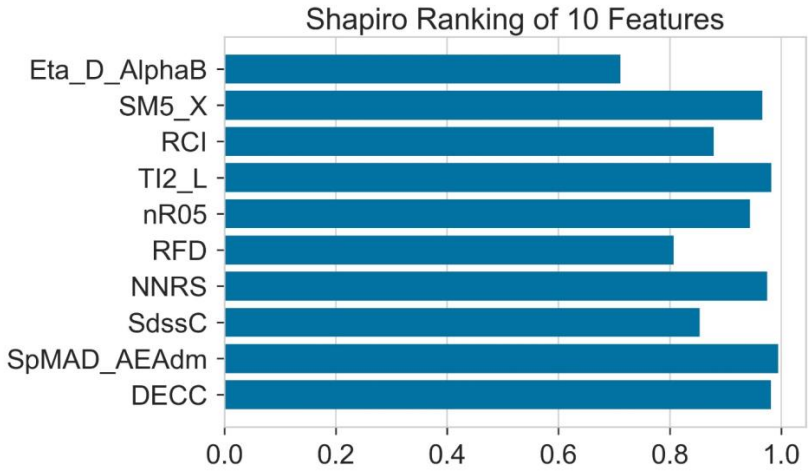

Figure 3. The normality of all the features analyzed by Shapiro test.

All the molecular descriptors are not able to give the performance value equally during the time of model training [40,41]. Consequently, to evaluate the performance ability of each feature or molecular descriptor, it is important to calculate the relative importance of all the molecular descriptors used. The feature present with a high value of relative importance shows that it can contribute most to the algorithm used in machine learning. Moreover, the feature of high relative importance shows that these are considered helpful for predicting machine learning models. Figure 4 indicates that the molecular descriptor RCI shows the least value of relative importance and its contribution to the algorithms is extremely low. In contrast, ETA-D-AlphaB shows a high value of relative importance,

and its contribution is greater than all the other features for the prediction of the machine learning models. The variety of the features shows different relative importance.

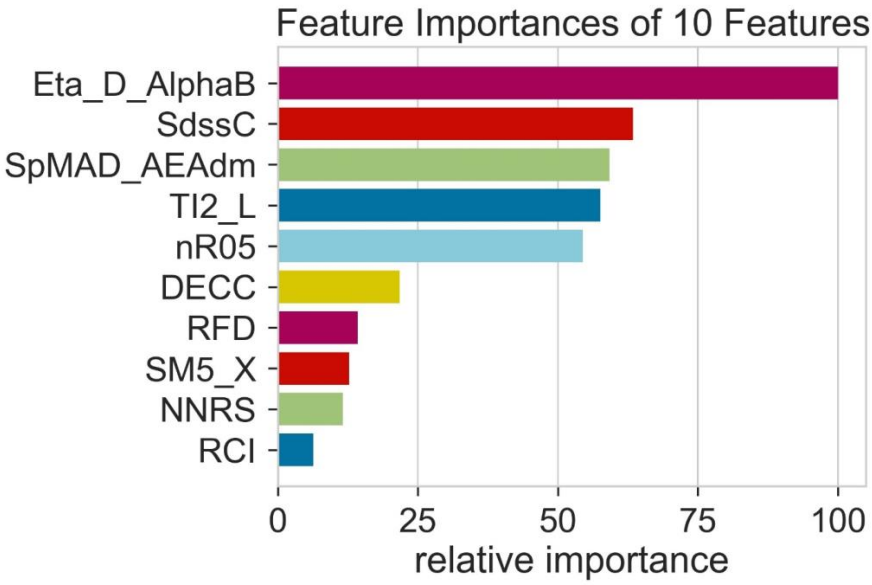

Figure 4. The relative importance of features.

Different models are tested for their predictive capability (Table 2). The light gradient boosting model (LGBM) and Hist gradient boosting are used for further analysis. A residual plot helps to identify problems associated with regression analysis. In the residual plot, the target variable is present on the x-axis and the residuals are on the y-axis. The deviation of the predicted value from the actual value is indicated by the residual values. If the data point is away from the zero line, the prediction value will differ from the actual values. The residual plot for the LGBM regression model is shown in Figure 5. The residual plot for the Hist gradient boosting model is shown in Figure 6. The behavior of LGBM regression models is like that of the Hist gradient boosting regression model. R^2 is the coefficient of determination for the test and trained value. R^2 for the test and train residues is not remarkably high: near zero, which is considered accurate. So, the results of both models are good. Dependence on expensive experimental techniques can be decreased by finding accurate results by machine learning models. The more similarities in the predicted and experimental values show that the model or method used was precise and accurate. The easy and fast prediction of results can speed up the design process of new structures of donor materials.

Table 2. R^2, mean absolute error (MAE) and root mean square error (RMSE) values of various models for HOMO prediction.

Model	Train R^2	Test R^2	Train MAE (eV)	Test MAE (eV)	Train RMSE (eV)	Test RMSE (eV)
Hist Gradient Boosting Regressor	0.912	0.820	0.136	0.146	0.163	0.176
LGBM Regressor	0.906	0.863	0.137	0.142	0.165	0.172
Random Forest Regressor	0.853	0.801	0.144	0.148	0.174	0.180
Decision Tree Regressor	0.752	0.683	0.150	0.155	0.183	0.193
Extra Trees Regressor	0.723	0.652	0.152	0.159	0.189	0.194
AdaBoost Regressor	0.623	0.560	0.161	0.172	0.1950	0.239
K-Neighbors Regressor	0.620	0.564	0.161	0.171	0.1950	0.237
Linear Regression	0.610	0.550	0.162	0.173	0.1960	0.243

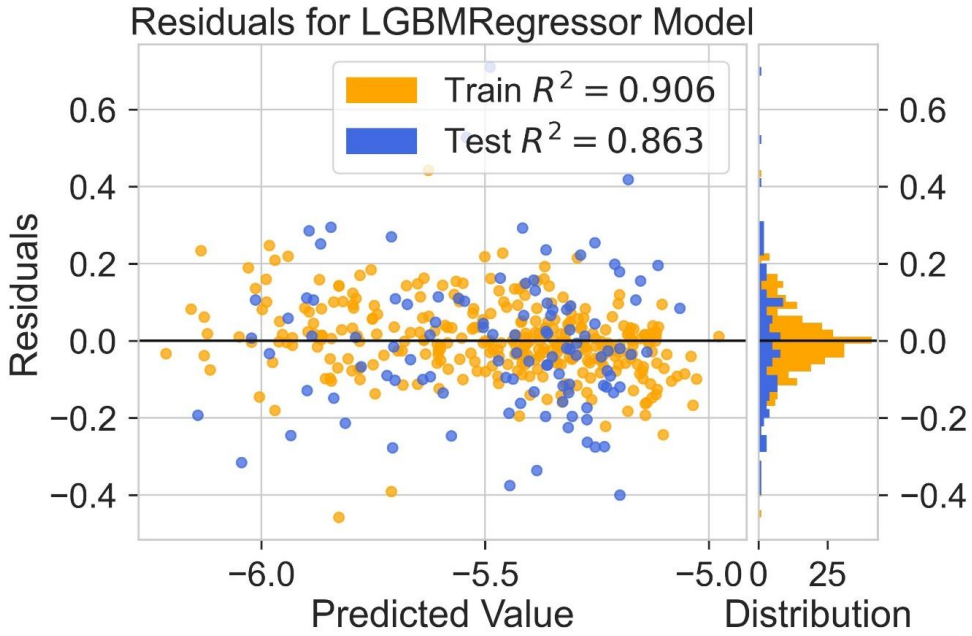

Figure 5. Residual for LGBM regression model.

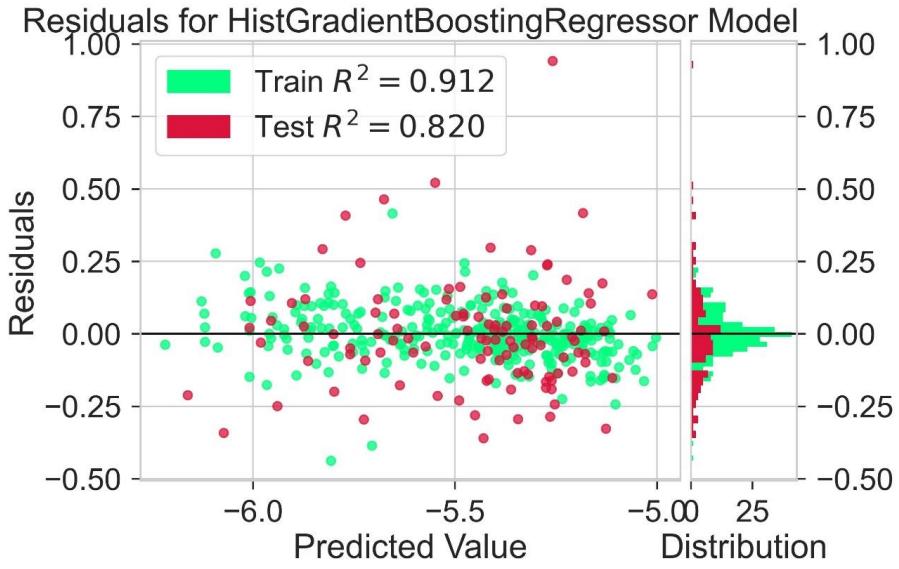

Figure 6. Residuals for Hist gradient booting regression model.

A scattered plot between the experimental value and expected value of HOMO using the LGBM Regression model and Hist gradient regression model is shown in Figure 7 and Figure 8, respectively. The scatter plot is drawn between the residuals for models and the experimental or predicted value. The majority of values are in the low range, close to zero, which is a clear indication of accurate results. The values for train residues and the values for the test residues are also close to zero. The results show that both LGBM Regression model and the Hist Gradient Boosting Regression model are the best models for regression analysis.

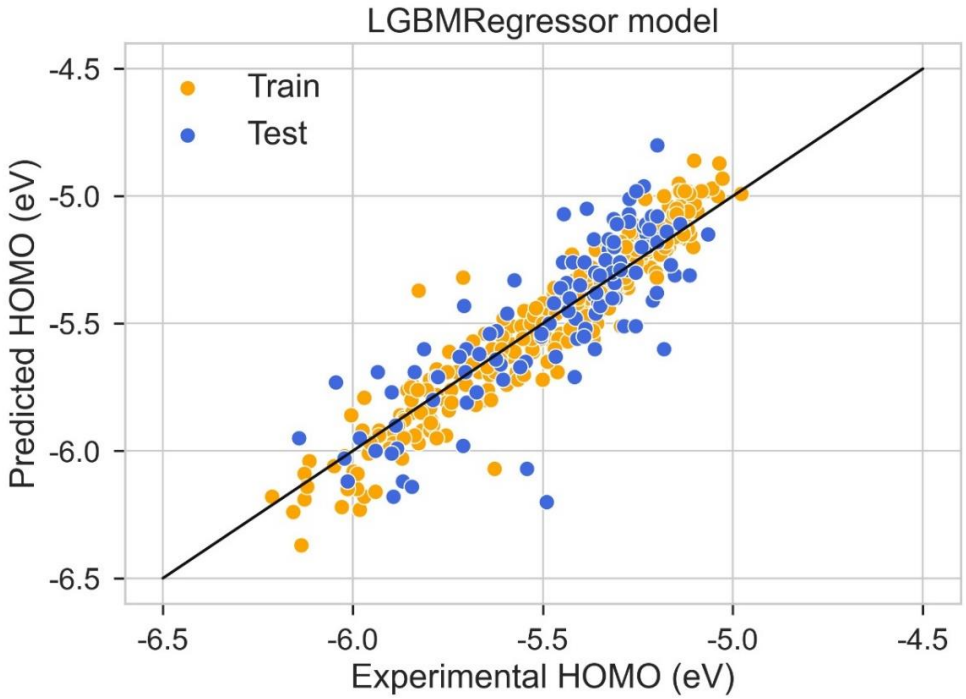

Figure 7. Scatter plot between experimental and predicted HOMO using LGBM regression model.

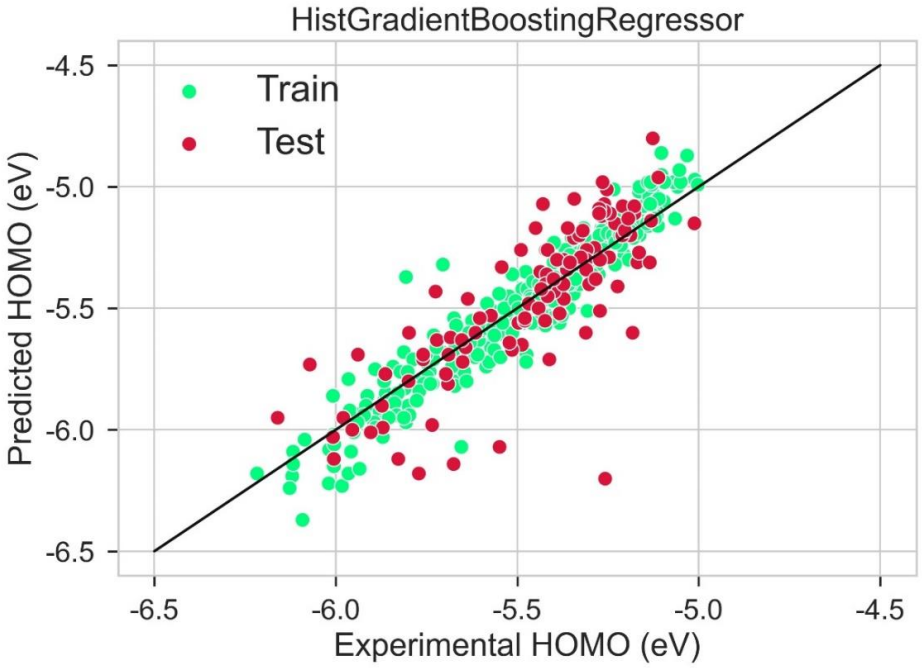

Figure 8. Scatter plot between experimental and predicted HOMO using Hist gradient booting regression.

2.6. LUMO Prediction

In Pearson ranking, a correlation between the LUMO and molecular descriptors is determined (Figure 9). The value of the Pearson ranking shows that the value from 0 to +1 shows a positive correlation. The molecular descriptors that fall in the value of 0 to +1 are indicated by red color. There is no correlation at zero point. The molecular descriptors having a blue appearance indicate a negative correlation. The negative correlation ranges from 0 to −1.

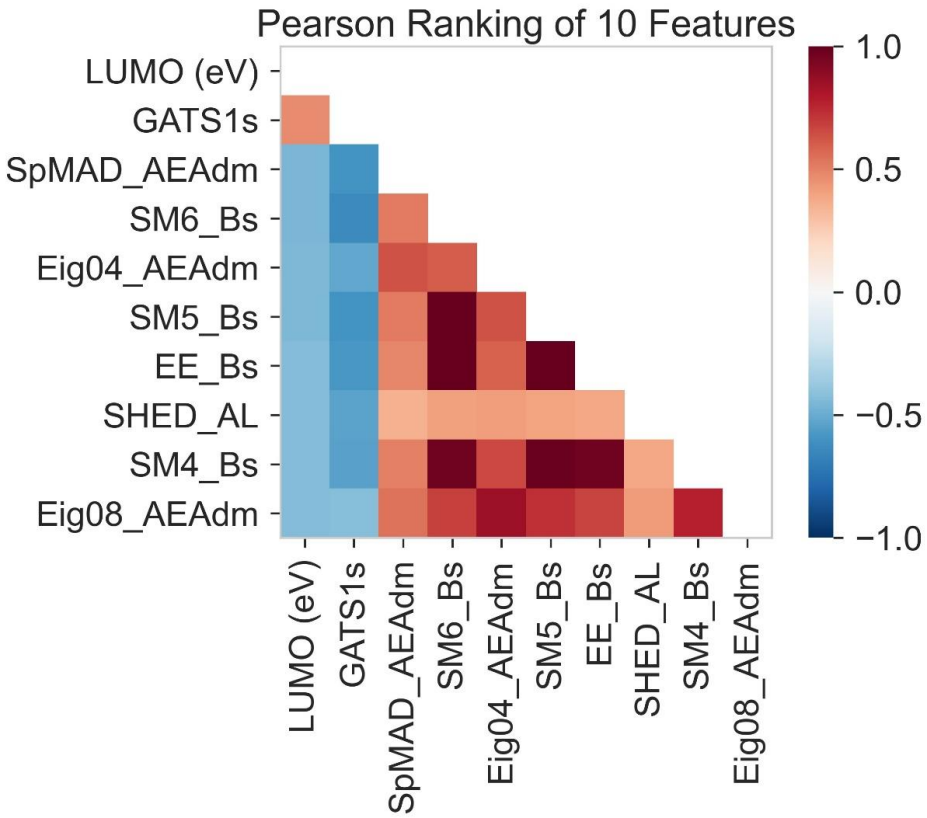

Figure 9. Correlation between LUMO and molecular descriptors.

GATS1s is the molecular descriptor (Table 3) presented on the y-axis. It indicates a positive correlation with LUMO lying on the x-axis, while the other molecular descriptors present on the y-axis show a negative correlation because their blue color indicates that the values of these molecular descriptors must lie between 0 and −1. In contrast, SPMAD-AEAdm, Eig04_EA(dm), EE_B(s), SM4_B(s), SM5_B(s), SM6_B(s), SHED-AL and Eig08_EA(dm) are the molecular descriptors present on the x-axis. These indicate a positive correlation with the LUMO present on the y-axis (red in color).

Table 3. Molecular descriptors with respective categories and descriptions.

No	Molecular Descriptor	Category	Description
1	SpAD_AEA(dm)	Edge adjacency indices	Spectral absolute deviation from augmented edge mat. weighted by dipole moment
2	GATS1s	2D autocorrelations	Geary autocorrelation of lag 1 weighted by I-state
3	Eig04_EA(dm)	Edge adjacency indices	Eigenvalue n. 4 from edge adjacency mat. weighted by dipole Moment
4	EE_B(s)	2D matrix-based descriptor	Estrada-like index (log function) from Burden matrix weighted by I-State
5	SM4_B(s)	2D matrix-based descriptors	Spectral moment of order 4 from Burden matrix by I-State
6	SM5_B(s)	2D matrix-based descriptors	Spectral moment of order 5 from Burden matrix by I-State
7	SM6_B(s)	2D matrix-based descriptors	Spectral moment of order 6 from Burden matrix by I-State
8	Eig08_EA(dm)	Edge adjacency indices	Eigenvalue n. 8 from edge adjacency mat. weighted by dipole Moment
9	SHED-AL		SHED Acceptor Lipophilic

A source of input is considered a calculated molecular descriptor for the model training. The chemistry of donor molecules is represented by the help of molecular descriptors. Figure 10 shows that molecular descriptors such as SPMAD-AEAdm, Eig04_EA(dm), EE_B(s), SM4_B(s), SM5_B(s), Eig08_EA(dm), SM6_B(s), and Eig08_EA(dm) indicate the negative dependent Pearson correlation. This negative correlation is determined by noting that these molecular descriptors lie between 0 and _1. In LUMO's case, only one molecular descriptor, GATS1s, shows a positive dependent correlation, with a value from 0 to +1.

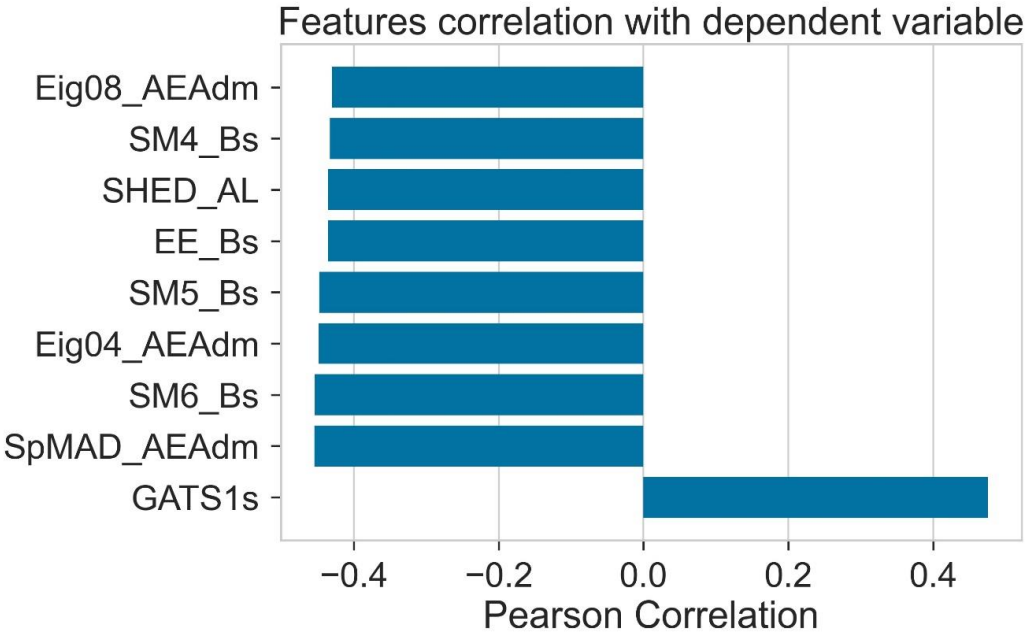

Figure 10. Correlation of features with LUMO.

To find the normality of the distribution of molecular descriptors, Shapiro ranking deals with a single molecular descriptor one at a time. In Figure 11, GATS1s shows the least normality according to Shapiro ranking. On the other hand, SPMAD-AEAdm and SM4-BS show the greatest normality according to this ranking.

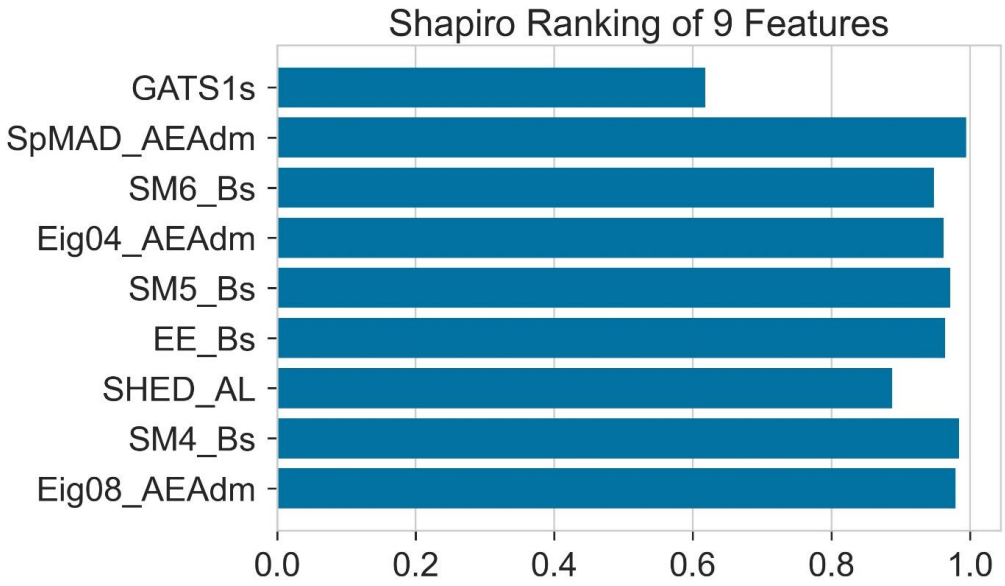

Figure 11. The normality of all the features analyzed by Shapiro test.

The relative importance of features tells us about the performing ability of different molecular descriptors. During the training of the model, all the molecular descriptors present are not able to perform at an equal level. It is important to calculate the relative importance of all the molecular descriptors to check the performance ability of each. So, the relative importance of features helps to evaluate the performing ability of different molecular descriptors. The molecular descriptor whose relative importance is high shows that it can be used mostly for the prediction of results. Additionally, the feature with the highest value of relative performance among all the features is considered helpful in training algorithms used in machine learning. Figure 12 shows that the molecular descriptor SM5-Bs shows the least value of relative importance and its contribution to the algorithms is extremely low. On the other hand, the molecular descriptors GATS1s, SPMAD-AEAdm, Eigo4-AEAdm and SHED-AL show high values of relative importance. The variety of features shows different relative importance.

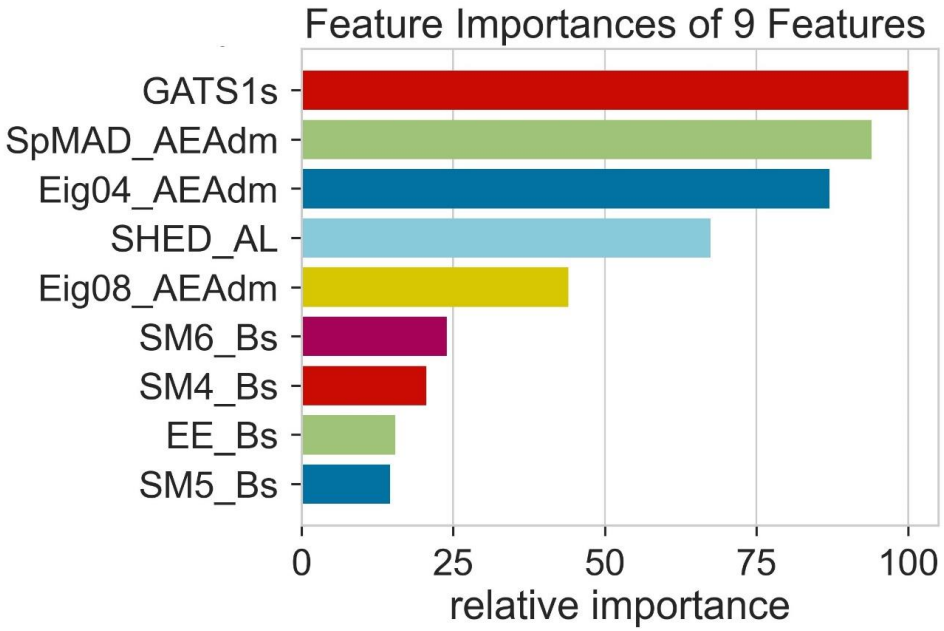

Figure 12. The relative importance of features.

A variety of regression models are used for the prediction of results [42]. R^2 values are given in Table 4. LGBM and Hist gradient boosting models consider the best working models for the prediction of LUMO. An analysis of different molecular descriptors is carried out by using these regression models. A residual plot is used to identify problems with regression analysis. In the residual plot, the relationship between the test value and the train value is predicted. The target variable is present on the x-axis, and the residuals are on the y-axis. If the value of train data is near the value of test data, then the chances of accurate results increase. If the values of the test and train data are not near each other or near the zero line, it indicates that the prediction value will differ further from the actual values. The residual plot for the LGBM regression model is shown in Figure 13. The residual plot for the Hist gradient boosting model is shown in Figure 14. The obtained results show that the behavior of LGBM regression models is like that of the Hist Gradient Boosting regression model. The coefficient of determination for the test and trained value is indicated by the symbol R^2. These regression models show that the value of R^2 is near zero. So, the results of both models are considered good enough. By using machine learning models, accurate results can be achieved. This is helpful to avoid expensive experimental techniques.

Table 4. R^2, mean absolute error (MAE) and root mean square error (RMSE) values for LUMO prediction.

Model	Train R^2	Test R^2	Train MAE (eV)	Test MAE (eV)	Train RMSE (eV)	Test RMSE (eV)
Hist Gradient Boosting Regressor	0.843	0.667	0.070	0.074	0.084	0.089
LGBM Regressor	0.831	0.602	0.071	0.075	0.085	0.090
Random Forest Regressor	0.820	0.601	0.072	0.076	0.087	0.092
Decision Tree Regressor	0.732	0.583	0.075	0.078	0.093	0.097
Extra Trees Regressor	0.723	0.570	0.076	0.080	0.095	0.097
AdaBoost Regressor	0.652	0.540	0.081	0.086	0.098	0.120
Linear Regression	0.612	0.504	0.082	0.087	0.099	0.121
K-Neighbors Regressor	0.610	0.520	0.081	0.087	0.098	0.122

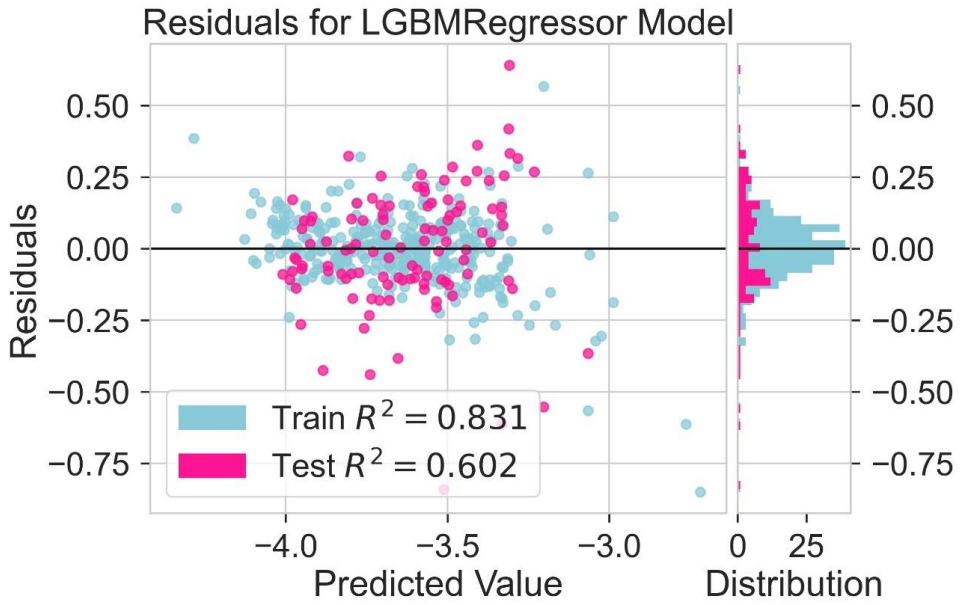

Figure 13. Residual for LGBM regression model for LUMO prediction.

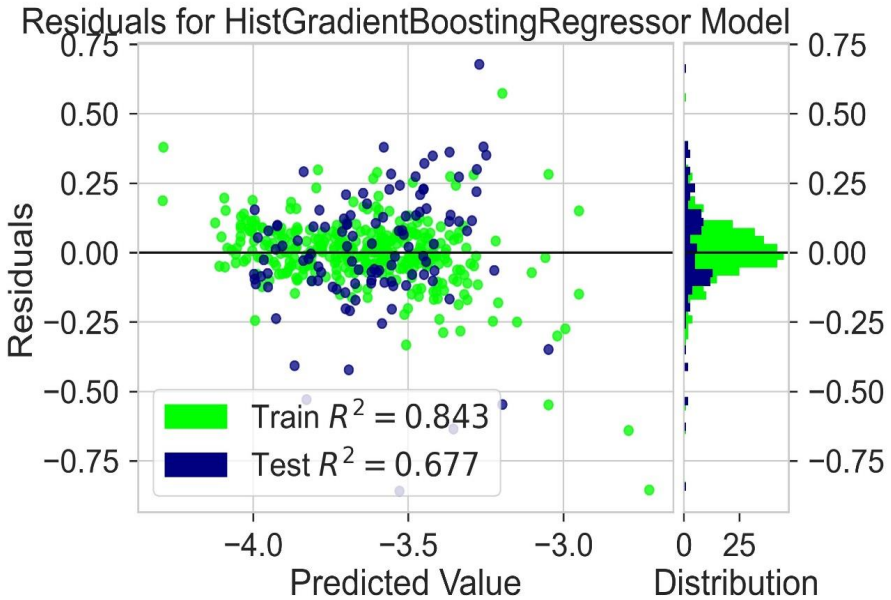

Figure 14. Residuals for Hist gradient booting regression model for LUMO prediction.

A scattered plot between the experimental value and expected value of LUMO using the LGBM regression model and Hist gradient regression model is shown in Figure 15 and Figure 16, respectively. The scatter plot is drawn between the residuals for the models and the experimentally predicted value. It is mostly used to find problems with regression models. For data points above the line, residuals are positive. For the data points below the line, the residuals are negative. The closer the value of the data points to 0, the more accurate it is for results. The scatter plot of LGBM and Hist gradient regression models shows that most of the values lie in the low range, near the value of zero, indicating accurate results.

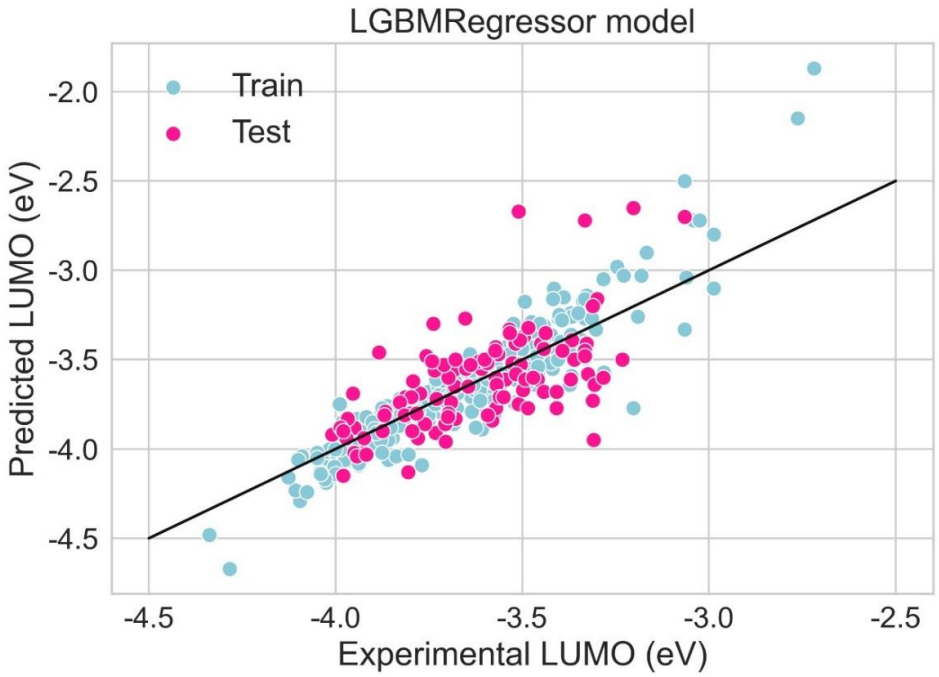

Figure 15. Scatter plot between experimental and predicted HOMO using LGBM regression model.

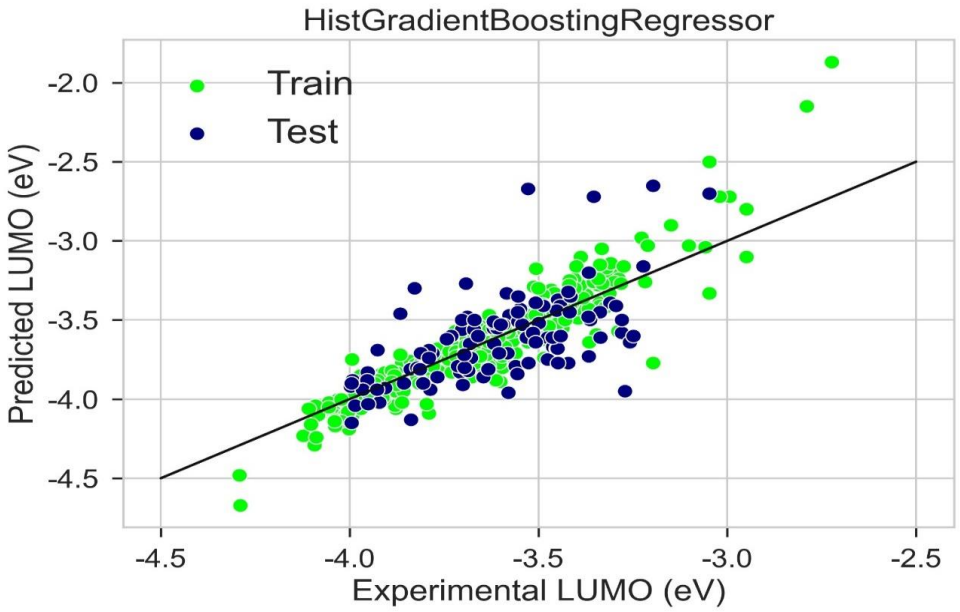

Figure 16. Scatter plot between experimental and predicted LUMO using Hist gradient booting.

2.7. Database Mining

The Clean Energy Project (CEP) is a database that contains thousands of organic molecules. These molecules can be used for various photovoltaics applications [43,44]. A similarity analysis is performed to find suitable building units. O4TIC is a low-band gap molecule. O4TIC contains a carbon–oxygen bridged-type ladder with strong electron-donating capability with the oxygen atoms conjugation effect. The further band gap of the molecule is decreased with the attachment of a more electron-rich group instead of the central phenyl group, which increases the donating capability of the molecule [43,45]. The linear side increases the crystallinity, which in turn increases the mobility of the electrons. The top search hits for O4TIC references are given in Figure 17. The building blocks found are not overly similar to O4TIC; however, most are suitable for the design of polymers for organic solar cells. The top search hits for middle O4TIC are given in Figure 18. All the structures are unique and possible to synthesize.

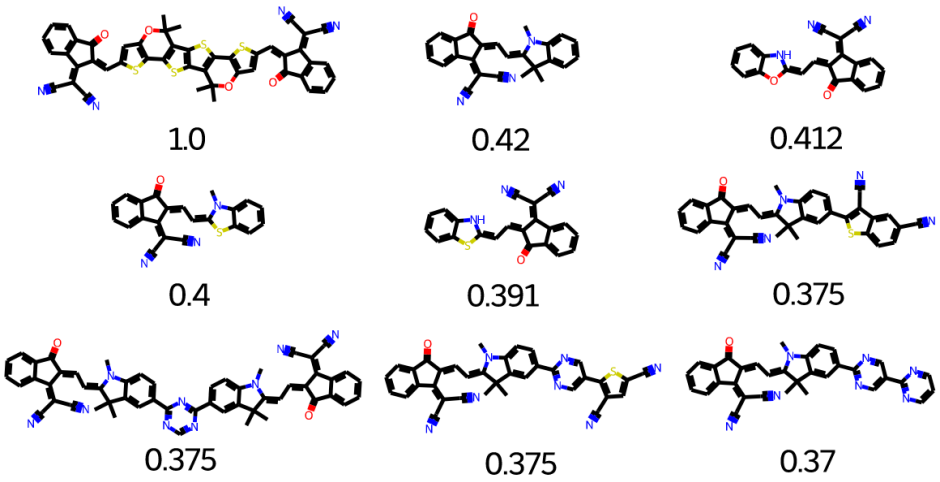

Figure 17. Top search hit for O4TIC.

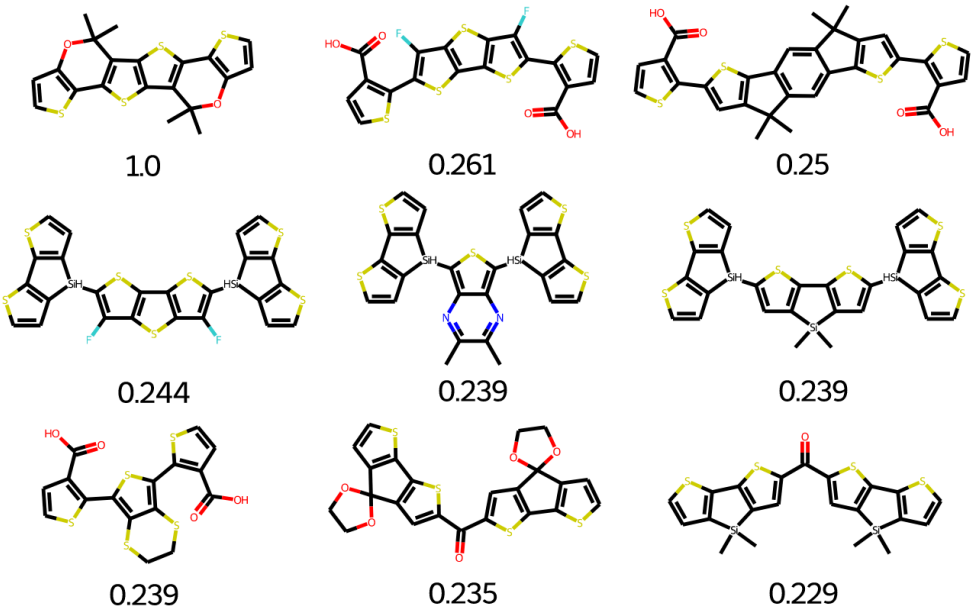

Figure 18. Top search hit for middle O4TIC.

The top search hits for Y5 are given in Figure 19. Many groups can be used for polymer designing. After a minor structural modification, other groups can also be useful candidates. The top search hits for Y5 middle are given in Figure 20. Molecular core of Y5 is used as

an electron deficient group. Y5 core structure is considered as high performance (NFA) non-fullerene acceptor. Y5 can be applied to both inverted and conventional OPV devices because of versatility of Y5. OSCs based on NFA can achieve longer device life-time with greater photochemical and thermal stability [46]. The combination of Y5 electron deficient with five different donor polymers could lead to enhanced efficiency [47].

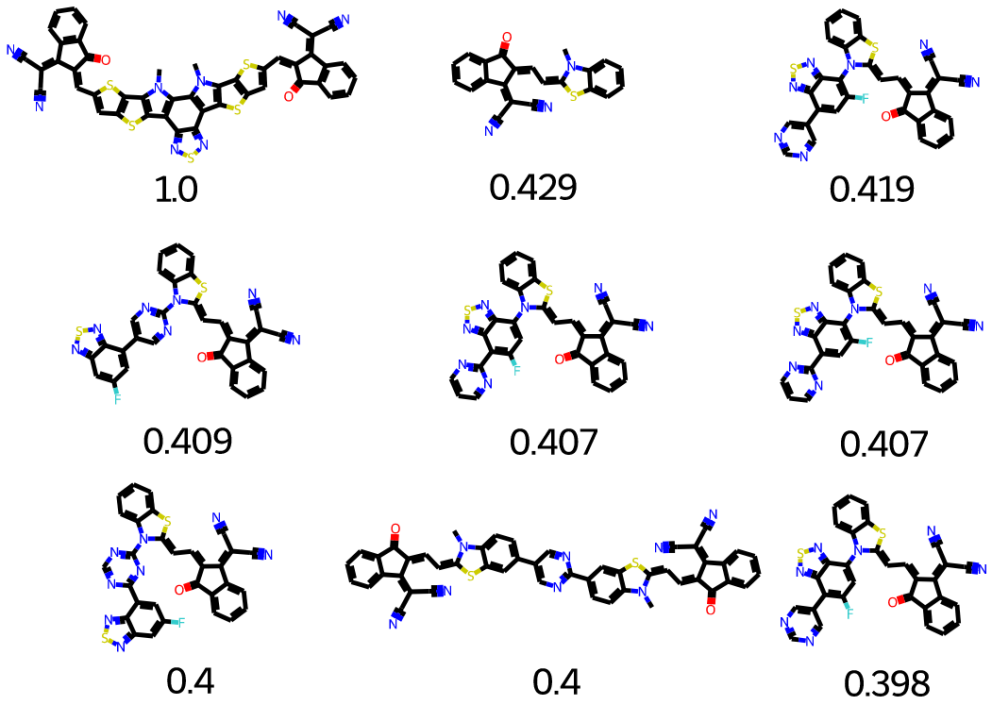

Figure 19. Top search hit for Y5.

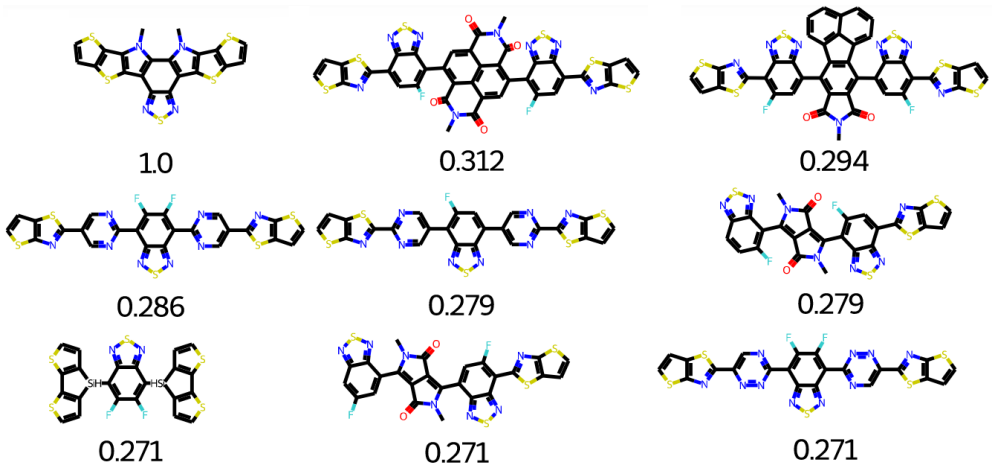

Figure 20. Top search hit for Y5 middle.

The organic building units are varied because various positions are available to add or connect plenty of heteroatoms [48–51]. This is carried out to produce or synthesize countless organic molecules that are better in their characteristics than the previous ones. To design new polymers or organic semiconductors for organic photodetectors, electron-deficient and electron-rich groups can be used. Hundreds of building blocks can be selected based on the addition of terminal groups and the availability of the position for alkyl chains. Many organic semiconductor materials can be designed by connecting new building units. A suitable combination of electron-rich and electron-deficient results in the formation of an electron hole, which leads to an increase in conjugation [52–55].

3. Methodology

3.1. Dataset

The data for machine learning were collected from research papers. The volume of data was enough for good machine learning models. The data are based on energy levels and photovoltaic parameters. The performance of the machine learning model strongly depends on the quality and quantity of the data [56].

3.2. Molecular Descriptor Calculation

Several types of molecular descriptors of molecules were calculated using Dragon software [57]. About 4000 descriptors were generated. The best descriptors were short-listed using univariate regression. These descriptors were used for training machine learning models.

3.3. Training the Model

We have imported the necessary packages of Python such as Scikit-learn, Pandas, Scipy, Numpy, Seaborn, and Matplotlib. These packages are necessary for data visualization and analysis. The calculated descriptors and target properties in comma-separated value (CSV) files were imported with the help of the Pandas module.

3.4. Similarity Analysis

A similarity analysis was performed using RDKit [58]. The similarity analysis is a straightforward method to find the similarities between reference structure and structure in the database. For this purpose, pharmacophores, distances, fingerprints, etc., can be used. In our work, Tanimoto similarity was used. For this purpose, ECFP4 fingerprints were selected.

4. Conclusions

In summary, data on large photovoltaic properties were collected from already reported experimental studies and subsequently utilized to train machine learning models. Among the multiple trained models, the LGBM regression model and Hist gradient booting regression model demonstrated the best predictive capability. Moreover, HOMO and LUMO energy levels were successfully predicted. The results revealed that good consistency was obtained between experimental outcomes and model predictions. In addition, Pearson correlation and Shapiro ranking was applied to demonstrate the correlation between different parameters. Furthermore, a similarity analysis was performed to find the similarities between reference structure and structure in the database. The reliability of our designed approach was also verified by mining the photovoltaic database to search for new building units. This indicates that machine learning is a powerful approach to predict the properties of photodetectors, which can facilitate their rapid development in various fields. Fast screening or searching of new building units with minimal computational costs could significantly reduce experimentation (trial and error methods) costs by narrowing down the search for potential candidates.

Author Contributions: Conceptualization, M.S. and J.S.; Methodology, M.J., S.H.; Software, M.S.A.; Validation, M.S. and J.S.; Formal Analysis, S.E.; Investigation, M.J.; Resources, J.S.; Data Curation, S.H.; Writing—Original Draft Preparation, M.S., M.J. and J.S.; Writing—Review & Editing, S.H., M.S.A.; Visualization, S.E.; Supervision, M.S.; Project Administration, M.S.; Funding Acquisition, J.S.; Writing—Review & Editing, G.M.K. All authors have read and agreed to the published version of the manuscript.

Funding: The authors sincerely appreciate funding from Researchers Supporting Project number (RSP2023R399), King Saud University, Riyadh, Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Associated data used in this paper can be obtained from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sulaman, M.; Yang, S.; Song, T.; Wang, H.; Wang, Y.; He, B.; Dong, M.; Tang, Y.; Song, Y.; Zou, B. High performance solution-processed infrared photodiode based on ternary PbSxSe1−x colloidal quantum dots. *RSC Adv.* **2016**, *6*, 87730–87737. [CrossRef]
2. Sulaman, M.; Song, Y.; Yang, S.; Li, M.; Saleem, M.I.; Chandraseakar, P.V.; Jiang, Y.; Tang, Y.; Zou, B. Ultra-sensitive solution-processed broadband photodetectors based on vertical field-effect transistor. *Nanotechnology* **2019**, *31*, 105203. [CrossRef] [PubMed]
3. Sulaman, M.; Song, Y.; Yang, S.; Saleem, M.I.; Li, M.; Perumal Veeramalai, C.; Zhi, R.; Jiang, Y.; Cui, Y.; Hao, Q.; et al. Interlayer of PMMA Doped with Au Nanoparticles for High-Performance Tandem Photodetectors: A Solution to Suppress Dark Current and Maintain High Photocurrent. *ACS Appl. Mater. Interfaces* **2020**, *12*, 26153–26160. [CrossRef] [PubMed]
4. Sulaman, M.; Yang, S.; Bukhtiar, A.; Fu, C.; Song, T.; Wang, H.; Wang, Y.; Bo, H.; Tang, Y.; Zou, B. High performance solution-processed infrared photodetector based on PbSe quantum dots doped with low carrier mobility polymer poly(N-vinylcarbazole). *RSC Adv.* **2016**, *6*, 44514–44521. [CrossRef]
5. Sulaman, M.; Yang, S.; Bukhtiar, A.; Tang, P.; Zhang, Z.; Song, Y.; Imran, A.; Jiang, Y.; Cui, Y.; Tang, L.; et al. Hybrid Bulk-Heterojunction of Colloidal Quantum Dots and Mixed-Halide Perovskite Nanocrystals for High-Performance Self-Powered Broadband Photodetectors. *Adv. Funct. Mater.* **2022**, *32*, 2201527. [CrossRef]
6. Hussain, R.; Hassan, F.; Khan, M.U.; Mehboob, M.Y.; Fatima, R.; Khalid, M.; Mahmood, K.; Tariq, C.J.; Akhtar, M.N. Molecular engineering of A–D–C–D–A configured small molecular acceptors (SMAs) with promising photovoltaic properties for high-efficiency fullerene-free organic solar cells. *Opt. Quantum Electron.* **2020**, *52*, 364. [CrossRef]
7. Saleem, M.I.; Yang, S.; Zhi, R.; Sulaman, M.; Chandrasekar, P.V.; Jiang, Y.; Tang, Y.; Batool, A.; Zou, B. Surface Engineering of All-Inorganic Perovskite Quantum Dots with Quasi Core−Shell Technique for High-Performance Photodetectors. *Adv. Mater. Interfaces* **2020**, *7*, 2000360. [CrossRef]
8. Sulaman, M.; Song, Y.; Yang, S.; Hao, Q.; Zhao, Y.; Li, M.; Saleem, M.I.; Chandraseakar, P.V.; Jiang, Y.; Tang, Y.; et al. High-performance solution-processed colloidal quantum dots-based tandem broadband photodetectors with dielectric interlayer. *Nanotechnology* **2019**, *30*, 465203. [CrossRef]
9. Hussain, R.; Mehboob, M.Y.; Khan, M.U.; Khalid, M.; Irshad, Z.; Fatima, R.; Anwar, A.; Nawab, S.; Adnan, M. Efficient designing of triphenylamine-based hole transport materials with outstanding photovoltaic characteristics for organic solar cells. *J. Mater. Sci.* **2021**, *56*, 5113–5131. [CrossRef]
10. Khalid, M.; Khan, M.U.; Ahmed, S.; Shafiq, Z.; Alam, M.M.; Imran, M.; Braga, A.A.C.; Akram, M.S. Exploration of promising optical and electronic properties of (non-polymer) small donor molecules for organic solar cells. *Sci. Rep.* **2021**, *11*, 21540. [CrossRef]
11. Babics, M.; Bristow, H.; Zhang, W.; Wadsworth, A.; Neophytou, M.; Gasparini, N.; McCulloch, I. Non-fullerene-based organic photodetectors for infrared communication. *J. Mater. Chem. C* **2021**, *9*, 2375–2380. [CrossRef]
12. Liao, X.; Xie, W.; Han, Z.; Cui, Y.; Xia, X.; Shi, X.; Yao, Z.; Xu, X.; Lu, X.; Chen, Y. NIR Photodetectors with Highly Efficient Detectivity Enabled by 2D Fluorinated Dithienopicenocarbazole-Based Ultra-Narrow Bandgap Acceptors. *Adv. Funct. Mater.* **2022**, *32*, 2204255. [CrossRef]
13. Mahmood, A. Photovoltaic and Charge Transport Behavior of Diketopyrrolopyrrole Based Compounds with A–D–A–D–A Skeleton. *J. Cluster Sci.* **2019**, *30*, 1123–1130. [CrossRef]
14. Janjua, M.R.S.A. How Does Bridging Core Modification Alter the Photovoltaic Characteristics of Triphenylamine-Based Hole Transport Materials? Theoretical Understanding and Prediction. *Chem. Eur. J.* **2021**, *27*, 4197–4210. [CrossRef] [PubMed]
15. Janjua, M.R.S.A. Photovoltaic properties and enhancement in near-infrared light absorption capabilities of acceptor materials for organic solar cell applications: A quantum chemical perspective via DFT. *J. Phys. Chem. Solids* **2022**, *171*, 110996. [CrossRef]
16. Mahmood, A.; Khan, S.U.-D.; Rehman, F.U. Assessing the quantum mechanical level of theory for prediction of UV/Visible absorption spectra of some aminoazobenzene dyes. *J. Saudi Chem. Soc.* **2015**, *19*, 436–441. [CrossRef]
17. Mahmood, A.; Khan, S.U.-D.; Rana, U.A.; Tahir, M.H. Red shifting of absorption maxima of phenothiazine based dyes by incorporating electron-deficient thiadiazole derivatives as π-spacer. *Arab. J. Chem.* **2019**, *12*, 1447–1453. [CrossRef]
18. Khalid, M.; Ali, A.; Khan, M.U.; Tahir, M.N.; Ahmad, A.; Ashfaq, M.; Hussain, R.; Morais, S.F.d.A.; Braga, A.A.C. Non-covalent interactions abetted supramolecular arrangements of N-Substituted benzylidene acetohydrazide to direct its solid-state network. *J. Mol. Struct.* **2021**, *1230*, 129827. [CrossRef]
19. Mahmood, A.; Saqib, M.; Ali, M.; Abdullah, M.I.; Khalid, B. Theoretical investigation for the designing of novel antioxidants. *Can. J. Chem.* **2013**, *91*, 126–130. [CrossRef]
20. Mahmood, A.; Abdullah Muhammad, I.; Nazar Muhammad, F. Quantum Chemical Designing of Novel Organic Non-Linear Optical Compounds. *Bull. Korean Chem. Soc.* **2014**, *35*, 1391–1396. [CrossRef]

21. Khalid, M.; Ali, A.; Abid, S.; Tahir, M.N.; Khan, M.U.; Ashfaq, M.; Imran, M.; Ahmad, A. Facile Ultrasound-Based Synthesis, SC-XRD, DFT Exploration of the Substituted Acyl-Hydrazones: An Experimental and Theoretical Slant towards Supramolecular Chemistry. *ChemistrySelect* **2020**, *5*, 14844–14856. [CrossRef]
22. Siddiqui, W.A.; Khalid, M.; Ashraf, A.; Shafiq, I.; Parvez, M.; Imran, M.; Irfan, A.; Hanif, M.; Khan, M.U.; Sher, F.; et al. Antibacterial metal complexes of o-sulfamoylbenzoic acid: Synthesis, characterization, and DFT study. *Appl. Organomet. Chem.* **2022**, *36*, e6464. [CrossRef]
23. Khalid, M.; Ali, A.; Asim, S.; Tahir, M.N.; Khan, M.U.; Curcino Vieira, L.C.; de la Torre, A.F.; Usman, M. Persistent prevalence of supramolecular architectures of novel ultrasonically synthesized hydrazones due to hydrogen bonding [X–H···O.; X=N]: Experimental and density functional theory analyses. *J. Phys. Chem. Solids* **2021**, *148*, 109679. [CrossRef]
24. Mebed, A.M.; Jafri, H.M.; Hakamy, A.; Abd-Elnaiem, A.M.; Sulaman, M.; Elshahat, S. Multidimensional modeling assisted mining of GDB17 chemical database: A search for polymer donors for organic solar cells and machine learning assisted performance prediction. *Int. J. Quantum Chem* **2022**, *122*, e26991. [CrossRef]
25. Janjua, M.R.S.A.; Irfan, A.; Hussien, M.; Ali, M.; Saqib, M.; Sulaman, M. Machine-Learning Analysis of Small-Molecule Donors for Fullerene Based Organic Solar Cells. *Energy Technol.* **2022**, *10*, 2200019. [CrossRef]
26. Mahmood, A.; Wang, J.-L. A time and resource efficient machine learning assisted design of non-fullerene small molecule acceptors for P3HT-based organic solar cells and green solvent selection. *J. Mater. Chem. A* **2021**, *9*, 15684–15695. [CrossRef]
27. Mahmood, A.; Irfan, A.; Wang, J.-L. Machine learning and molecular dynamics simulation-assisted evolutionary design and discovery pipeline to screen efficient small molecule acceptors for PTB7-Th-based organic solar cells with over 15% efficiency. *J. Mater. Chem. A* **2022**, *10*, 4170–4180. [CrossRef]
28. Khan, S.U.-D.; Mahmood, A.; Rana, U.A.; Haider, S. Utilization of electron-deficient thiadiazole derivatives as π-spacer for the red shifting of absorption maxima of diarylamine-fluorene based dyes. *Theor. Chem. Acc.* **2014**, *134*, 1596. [CrossRef]
29. Sharif, H.M.A.; Cheng, H.-Y.; Haider, M.R.; Khan, K.; Yang, L.; Wang, A.-J. NO Removal with Efficient Recovery of N2O by Using Recyclable Fe3O4@EDTA@Fe(II) Complex: A Novel Approach toward Resource Recovery from Flue Gas. *Environ. Sci. Technol.* **2019**, *53*, 1004–1013. [CrossRef]
30. Sharif, H.M.A.; Farooq, M.; Hussain, I.; Ali, M.; Mujtaba, M.A.; Sultan, M.; Yang, B. Recent innovations for scaling up microbial fuel cell systems: Significance of physicochemical factors for electrodes and membranes materials. *J. Taiwan Inst. Chem. Eng.* **2021**, *129*, 207–226. [CrossRef]
31. Tahir, M.H.; Mubashir, T.; Shah, T.-U.-H.; Mahmood, A. Impact of electron-withdrawing and electron-donating substituents on the electrochemical and charge transport properties of indacenodithiophene-based small molecule acceptors for organic solar cells. *J. Phys. Org. Chem.* **2019**, *32*, e3909. [CrossRef]
32. Sharif, H.M.A.; Mahmood, N.; Wang, S.; Hussain, I.; Hou, Y.-N.; Yang, L.-H.; Zhao, X.; Yang, B. Recent advances in hybrid wet scrubbing techniques for NOx and SO2 removal: State of the art and future research. *Chemosphere* **2021**, *273*, 129695. [CrossRef] [PubMed]
33. Mahmood, A.; Irfan, A.; Wang, J.-L. Developing Efficient Small Molecule Acceptors with sp2-Hybridized Nitrogen at Different Positions by Density Functional Theory Calculations, Molecular Dynamics Simulations and Machine Learning. *Chem. Eur. J.* **2022**, *28*, e202103712. [CrossRef] [PubMed]
34. Mahmood, A.; Irfan, A. Computational analysis to understand the performance difference between two small-molecule acceptors differing in their terminal electron-deficient group. *J. Comput. Electron.* **2020**, *19*, 931–939. [CrossRef]
35. Mahmood, A.; Irfan, A.; Ahmad, F.; Ramzan Saeed Ashraf Janjua, M. Quantum chemical analysis and molecular dynamics simulations to study the impact of electron-deficient substituents on electronic behavior of small molecule acceptors. *Comput. Theor. Chem.* **2021**, *1204*, 113387. [CrossRef]
36. Khalid, M.; Khan, M.U.; Razia, E.-t.; Shafiq, Z.; Alam, M.M.; Imran, M.; Akram, M.S. Exploration of efficient electron acceptors for organic solar cells: Rational design of indacenodithiophene based non-fullerene compounds. *Sci. Rep.* **2021**, *11*, 19931. [CrossRef]
37. Khalid, M.; Momina; Imran, M.; Rehman, M.F.U.; Braga, A.A.C.; Akram, M.S. Molecular engineering of indenoindene-3-ethylrodanine acceptors with A2-A1-D-A1-A2 architecture for promising fullerene-free organic solar cells. *Sci. Rep.* **2021**, *11*, 20320. [CrossRef]
38. Khan, M.U.; Hussain, R.; Mehboob, M.Y.; Khalid, M.; Ehsan, M.A.; Rehman, A.; Janjua, M.R.S.A. First theoretical framework of Z-shaped acceptor materials with fused-chrysene core for high performance organic solar cells. *Spectrochim. Act A Mol. Biomol. Spectrosc.* **2021**, *245*, 118938. [CrossRef]
39. Mahmood, A.; Irfan, A.; Wang, J.-L. Machine Learning for Organic Photovoltaic Polymers: A Minireview. *Chin. J. Polym. Sci.* **2022**, *40*, 870–876. [CrossRef]
40. Mahmood, A.; Abdullah, M.I.; Khan, S.U.-D. Enhancement of nonlinear optical (NLO) properties of indigo through modification of auxiliary donor, donor and acceptor. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2015**, *139*, 425–430. [CrossRef]
41. Mahmood, A.; HussainTahir, M.; Irfan, A.; Khalid, B.; Al-Sehemi, A.G. Computational Designing of Triphenylamine Dyes with Broad and Red-shifted Absorption Spectra for Dye-sensitized Solar Cells using Multi-Thiophene Rings in π-Spacer. *Bull. Korean Chem. Soc.* **2015**, *36*, 2615–2620. [CrossRef]
42. Mahmood, A.; Wang, J.-L. Machine learning for high performance organic solar cells: Current scenario and future prospects. *Energy Environ. Sci.* **2021**, *14*, 90–105. [CrossRef]

43. Zhang, Y.; Ji, Y.; Zhang, Y.; Zhang, W.; Bai, H.; Du, M.; Wu, H.; Guo, Q.; Zhou, E. Recent Progress of Y6-Derived Asymmetric Fused Ring Electron Acceptors. *Adv. Funct. Mater.* **2022**, *32*, 2205115. [CrossRef]
44. Yang, J.; Xiao, B.; Tang, A.; Li, J.; Wang, X.; Zhou, E. Aromatic-Diimide-Based n-Type Conjugated Polymers for All-Polymer Solar Cell Applications. *Adv. Mater.* **2019**, *31*, 1804699. [CrossRef] [PubMed]
45. Mahmood, A.; Irfan, A.; Wang, J.-L. Molecular level understanding of the chalcogen atom effect on chalcogen-based polymers through electrostatic potential, non-covalent interactions, excited state behaviour, and radial distribution function. *Polym. Chem.* **2022**, *13*, 5993–6001. [CrossRef]
46. Guo, Q.; Guo, Q.; Geng, Y.; Tang, A.; Zhang, M.; Du, M.; Sun, X.; Zhou, E. Recent advances in PM6:Y6-based organic solar cells. *Mater. Chem. Front.* **2021**, *5*, 3257–3280. [CrossRef]
47. Nie, Q.; Tang, A.; Guo, Q.; Zhou, E. Benzothiadiazole-based non-fullerene acceptors. *Nano Energy* **2021**, *87*, 106174. [CrossRef]
48. Mahmood, A.; Irfan, A. Effect of fluorination on exciton binding energy and electronic coupling in small molecule acceptors for organic solar cells. *Comput. Theor. Chem.* **2020**, *1179*, 112797. [CrossRef]
49. Khan, M.U.; Hussain, R.; Yasir Mehboob, M.; Khalid, M.; Shafiq, Z.; Aslam, M.; Al-Saadi, A.A.; Jamil, S.; Janjua, M.R.S.A. In Silico Modeling of New "Y-Series"-Based Near-Infrared Sensitive Non-Fullerene Acceptors for Efficient Organic Solar Cells. *ACS Omega* **2020**, *5*, 24125–24137. [CrossRef]
50. Khan, M.U.; Khalid, M.; Arshad, M.N.; Khan, M.N.; Usman, M.; Ali, A.; Saifullah, B. Designing Star-Shaped Subphthalocyanine-Based Acceptor Materials with Promising Photovoltaic Parameters for Non-fullerene Solar Cells. *ACS Omega* **2020**, *5*, 23039–23052. [CrossRef]
51. Mahmood, A.; Hu, J.-Y.; Xiao, B.; Tang, A.; Wang, X.; Zhou, E. Recent progress in porphyrin-based materials for organic solar cells. *J. Mater. Chem. A* **2018**, *6*, 16769–16797. [CrossRef]
52. Khan, M.U.; Khalid, M.; Hussain, R.; Umar, A.; Mehboob, M.Y.; Shafiq, Z.; Imran, M.; Irfan, A. Novel W-Shaped Oxygen Heterocycle-Fused Fluorene-Based Non-Fullerene Acceptors: First Theoretical Framework for Designing Environment-Friendly Organic Solar Cells. *Energy Fuels* **2021**, *35*, 12436–12450. [CrossRef]
53. Mehboob, M.Y.; Hussain, R.; Khan, M.U.; Adnan, M.; Umar, A.; Alvi, M.U.; Ahmed, M.; Khalid, M.; Iqbal, J.; Akhtar, M.N.; et al. Designing N-phenylaniline-triazol configured donor materials with promising optoelectronic properties for high-efficiency solar cells. *Comput. Theor. Chem.* **2020**, *1186*, 112908. [CrossRef]
54. Khan, M.U.; Mehboob, M.Y.; Hussain, R.; Fatima, R.; Tahir, M.S.; Khalid, M.; Braga, A.A.C. Molecular designing of high-performance 3D star-shaped electron acceptors containing a truxene core for nonfullerene organic solar cells. *J. Phys. Org. Chem.* **2021**, *34*, e4119. [CrossRef]
55. Mahmood, A.; Hu, J.; Tang, A.; Chen, F.; Wang, X.; Zhou, E. A novel thiazole based acceptor for fullerene-free organic solar cells. *Dyes Pigm.* **2018**, *149*, 470–474. [CrossRef]
56. Irfan, A.; Hussien, M.; Mehboob, M.Y.; Ahmad, A.; Janjua, M.R.S.A. Learning from Fullerenes and Predicting for Y6: Machine Learning and High-Throughput Screening of Small Molecule Donors for Organic Solar Cells. *Energy Technol.* **2022**, *10*, 2101096. [CrossRef]
57. Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. DRAGON software: An easy approach to molecular descriptor calculations. *MATCH Commun. Math. Comput. Chem.* **2006**, *56*, 237–248.
58. Landrum, G. RDKit: Open-Source Cheminformatics. Available online: http://www.rdkit.org.