STM JOURNALS

International Journal of
Cheminformatics

Volume 4, Issue 1, 2026
January—June
DOI (Journal): 10.37591/IJCI

https://journals.stmjournals.com/ijci/

Research IJCI

DFT/Data Guided Predictive Modelling of Absorption Maxima in the OLED Rubrene Derivatives

Anjana K.S.[1,2], Rencemon T. Thoppil[1,2], Sneha Anna Sunny[1,2], Renjith Thomas[1,2,*], Anila Skariah[3]

Abstract

This study investigates the optical properties of rubrene derivatives to develop an accurate predictive model for absorption maxima using computational chemistry and chemoinformatic techniques. We benchmarked various quantum chemical methods, identifying that the M06-2X/aug-cc-pVDZ method in dichloromethane (DCM) provided the strongest correlation with experimental data. Key molecular descriptors such as band gap, ionization potential, and electrophilicity index were calculated and analyzed using principal component analysis (PCA) to identify significant factors influencing absorption maxima. A multiple linear regression model was then developed and validated using test molecules, achieving an R^2 value of 0.7512. The predictive model demonstrated average accuracy in forecasting absorption maxima, aligning well with experimentally observed values. This study offers some insights into the structure–property relationships in rubrene derivatives and provides a reliable computational approach for guiding the design of new OLED materials.

Furthermore, the influence of substituent effects on the electronic distribution within the rubrene framework was systematically examined, revealing that electron-donating groups tend to induce bathochromic shifts, while electron-withdrawing groups lead to hypsochromic behavior. Solvent effects were also evaluated using implicit solvation models, confirming the critical role of polarity in modulating excited-state properties. The robustness of the developed model was assessed through cross-validation and external validation datasets, ensuring its generalizability across structurally diverse derivatives. In addition, correlation analysis highlighted the combined contribution of frontier molecular orbital energies and global reactivity descriptors in determining optical responses. These findings not only enhance the understanding of photophysical behavior in rubrene-based systems but also establish a cost-effective screening strategy for the rational design and optimization of high-performance organic optoelectronic materials.

Keywords: Rubrene, DFT, machine learning, absorption maxima, predictive modelling

***Author for Correspondence**
Renjith Thomas
E-mail: renjith@sbcollege.ac.in

[1]Professor, Department of Chemistry, St Berchmans College (Autonomous), Changanassery, Kerala, India
[2]Student, Centre for Computational and Theoretical Chemistry, St Berchmans College (Autonomous), Changanassery, Kerala, India
[3]Faculty, Department of Economics, St Berchmans College (Autonomous), Changanassery, Kerala, India

Received Date: March 31, 2026
Accepted Date: April 22, 2026
Published Date: April 30, 2026

Citation: Anjana K.S., Rencemon T. Thoppil, Sneha Anna Sunny, Renjith Thomas, Anila Skariah. DFT/Data Guided Predictive Modelling of Absorption Maxima in the OLED Rubrene Derivatives. International Journal of Cheminformatics. 2026; 4(1): 41–56p.

INTRODUCTION

In the field of optoelectronics, organic light-emitting diodes have gained significant attention due to their potential for use in various applications such as display technologies and eco-friendly lighting sources since the pioneering work of Tang and Van Slyke in 1987 [1, 2]. There are potential applications of OLEDs beyond traditional display technologies, such as in OLED-based wearable healthcare [8], several other biomedical applications [9] and automotive market [10]. Unlike traditional inorganic LEDs, which typically use semiconductor materials like gallium nitride, OLEDs utilize organic molecules or polymers as the emissive layer [11]. Among the organic materials, rubrene (5, 6, 11, 12-

tetraphenyltetracene) and its derivatives stand out due to their unique molecular structure and exceptional charge carrier mobility for holes, making them promising candidates for OLED development [12,13].

Extensive research has demonstrated the potential of rubrene-based OLEDs, showcasing high power efficiency and favorable charge transport properties [14, 15]. The rubrene layer improves charge carrier injection, reduces metal penetration, and enhances device performance in OLED applications with an optimal thickness of around 8nm [16]. However, predicting the optoelectronic properties of rubrene derivatives poses challenges due to their complex molecular structures and vast variations. To address this, we combine computational chemistry and machine learning to predict the behavior of rubrene derivatives efficiently.

Computational chemistry and machine learning can be combined to make predictions in molecular modeling, catalysis, and drug design. Here, we used this combination to predict the behavior of rubrene derivatives [17–19].The integration of machine learning enhances our predictive capabilities by analyzing and extracting patterns from large datasets, enabling us to make accurate predictions about the behavior of rubrene derivatives based on their molecular structures and properties [20, 21]. Through the synergy of machine learning and computational chemistry, we can predict the behavior of rubrene derivatives with much greater efficiency than experimental studies, thereby aiding in the development of rubrene derivatives as OLED materials.

Theoretical calculations, including Density Functional Theory (DFT) and Time-Dependent DFT (TD-DFT), are vital for discovering the intricate properties of OLED materials [22–25]. In 2020, research conducted by Zhang et al. [26] provided insight into how substituents affect the molecular structure and electronic properties of rubrene using theoretical methods such as DFT and TD-DFT. Liang Zhao et al. [27] also utilized theoretical studies to analyze the impact of conjugation and perpendicular phenyl groups on the two-photon absorption cross section (δmax) of rubrene derivatives. These studies highlight the importance of understanding the structure-property relationships of rubrene derivatives through theoretical studies.

Machine learning techniques provide a powerful solution for data analysis and prediction in various fields, including chemistry [28–36]. In the case of investigating the OLED properties of rubrene, machine learning can be highly beneficial [37]. By training machine learning models on a dataset of rubrene derivatives with known OLED properties, we can extract patterns and relationships between molecular structures and properties [29, 38].

The experimental absorbance values of 8 rubrene derivatives (Figure 1) from the study conducted by Paraskar et al. [39] and Gaozhan Xie [40] became the foundation of our work. To facilitate this study, we can construct the necessary dataset through the utilization of Density Functional Theory (DFT) and Time-Dependent Density Functional Theory (TD-DFT) methods. Linear regression [41, 42] a fundamental statistical technique used to model the relationship between a dependent variable and one or more independent variables, can be applied to the dataset to identify the factors influencing the absorption maxima of rubrene derivatives. This research intersects quantum mechanics, machine learning, and OLED technology [43, 44] offering insights into accelerated material discovery.

METHODOLOGY

The study aimed at understanding the optical properties of rubrene and its derivatives, which are crucial for enhancing OLED technology. A thorough literature review was conducted to gather experimental absorbance data for various rubrene derivatives. This data served as the benchmark for our work. Our methodology integrated computational chemistry, particularly Density Functional Theory (DFT) [45–47] and Time-Dependent Density Functional Theory (TD-DFT), along with machine learning, specifically linear regression analysis.

Initially, we selected eight rubrene derivatives for our analysis, six of which were sourced from a study by Paraskar et al., which extensively explored the structural and electronic properties of rubrene and its derivatives. The remaining two derivatives were obtained from a study conducted by Xie on the synthesis and characterization of azaacenes. The absorbance data from these studies formed the foundation of our research. The most closely correlating theoretically generated absorption maxima of eight rubrene derivatives to the experimentally calculated absorption maxima were determined. Geometric optimization of rubrene derivatives was carried out using DFT with the B3LYP functional and the 6-31G(d) basis set in the Gaussian 09 software [48, 49]. Frequency analysis was conducted to confirm that there were no negative frequencies present, ensuring the stability of the optimized structures. Subsequently, TD-DFT calculations were conducted to simulate the absorption spectra of the rubrene derivatives under various conditions, including vacuum and solvent atmospheres (specifically, dichloromethane). Different combinations of basis sets and functionals were utilized to obtain the absorption spectra. Specifically, absorption spectra were obtained in vacuum using CAM-B3LYP/6-31G(d) and in a solvent atmosphere of dichloromethane using CAM-B3LYP/6-31G(d), CAM-B3LYP/cc-pVDZ, M06-2X/6-31G(d), M06-2X/cc-pVDZ, and M06-2X/aug-cc-pVDZ.

To assess the accuracy of our theoretical predictions, we compared the computed absorption maxima with experimental data. It was found that the theoretical data obtained using M06-2X/aug-cc-pVDZ showed the best correlation with the experimental data, and this method was subsequently used to determine the absorption spectra for the newly designed rubrene derivatives. Various variables suspected to affect the absorption maxima were then generated from FMO analysis, Multiwfn (3.7) software, [50] optimized files, and other mathematical calculations. Factors such as the Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbital (LUMO) energies were extracted from the checkpoint files generated during calculations. These parameters were used to calculate key descriptors like band gap, ionization potential, electron affinity, electronegativity, chemical potential, hardness, softness, etc. as follows [51–56].

- Band gap = $E_{LUMO} - E_{HOMO}$ (1)
- Ionization potential (I) $= -E_{HOMO}$ (2)
- Electron Affinity (A) $= -E_{LUMO}$ (3)
- Electronegativity (χ) $= \frac{I+A}{2}$ (4)
- Chemical potential (μ) $= -(\chi)$ (5)
- Chemical hardness (η) $= \frac{I-A}{2}$ (6)
- Chemical softness (s) $= \frac{1}{2\eta}$ (7)
- Electrophilicity index ($\boldsymbol{\omega}$) $= \frac{\mu^2}{2\eta}$ (8)
- Electron accepting capability ($\boldsymbol{\omega}$ +) $= \frac{(I+3A)^2}{16(I-A)}$ (9)
- Electron donating capability ($\boldsymbol{\omega}$ -) $= \frac{(3I+A)^2}{16(I-A)}$ (10)
- Net electrophilicity index ($\Delta\boldsymbol{\omega}$ +) $= (\omega^+ - \omega^-)$ (11)
- $^{\Delta E}_{\text{Back donation}} = -\frac{\eta}{4}$ (12)
- Optical Softness (σ_0) $= \frac{1}{\Delta E}$ (13)
- Nucleophilicity Index (N) $= \frac{1}{\omega}$ (14)
- Maximum charge transfer capability (ΔN_{max}) $= \frac{I+A}{2(I-A)}$ (15)

Furthermore, we investigated molecular moments, including dipole moment, quadrupole moment, and octopole moment, obtained from wave function files generated during calculations. These

moments contribute to the overall polarity and electronic distribution within the molecules, which may influence their optical properties.

Once we collected and organized all the factors related to rubrene derivatives' properties, a dataset was compiled in Excel. Using this dataset, we initially conducted principal component analysis [57, 58], a statistical technique that simplifies complex data while retaining trends and patterns. Subsequently, linear regression analysis [59] was performed using Jamovi software [60] to determine which factors significantly influenced the absorbance values of rubrene derivatives. After identifying the influential factors, multiple regression analysis was conducted to pinpoint the major contributors to absorption max and derived a linear regression equation illustrating their contributions.

The general form of a multiple linear regression model [61-65] is given by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon \quad (16)$$

Where;

- Y is the dependent variable
- $X_1, X_2, \ldots, X_n$ are independent variables,
- β_0 is the intercept,
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients,
- ϵ is the error term

In our study the multivariable regression model utilized the Absorption maxima as the dependent variable and eight independent variables: Band Gap, Optical Softness, ΔE Back Donation, Nucleophilicity Index, Softness, Hardness, Chemical Potential, and Electrophilicity Index.

The small dataset of eight derivatives is a limitation; therefore, the validation of our model with 17 additional test molecules aimed to mitigate this issue. However, further validation with more extensive and diverse datasets is needed for a comprehensive assessment [26,65,]. Following molecular optimization and TD-DFT analysis similar to the experimental data, the major contributing factors affecting absorption max were identified for each of these test molecules. These factors were then incorporated into the previously obtained linear equation, and the modelled and theoretically calculated absorption maxima were compared to assess the accuracy of the model.

RESULTS & DISSCUSSIONS

In our current study, we have conducted calculations on eight derivatives of rubrene to predict their suitability for OLED applications (See Figure 2 for the optimized structures). Initially, we optimized

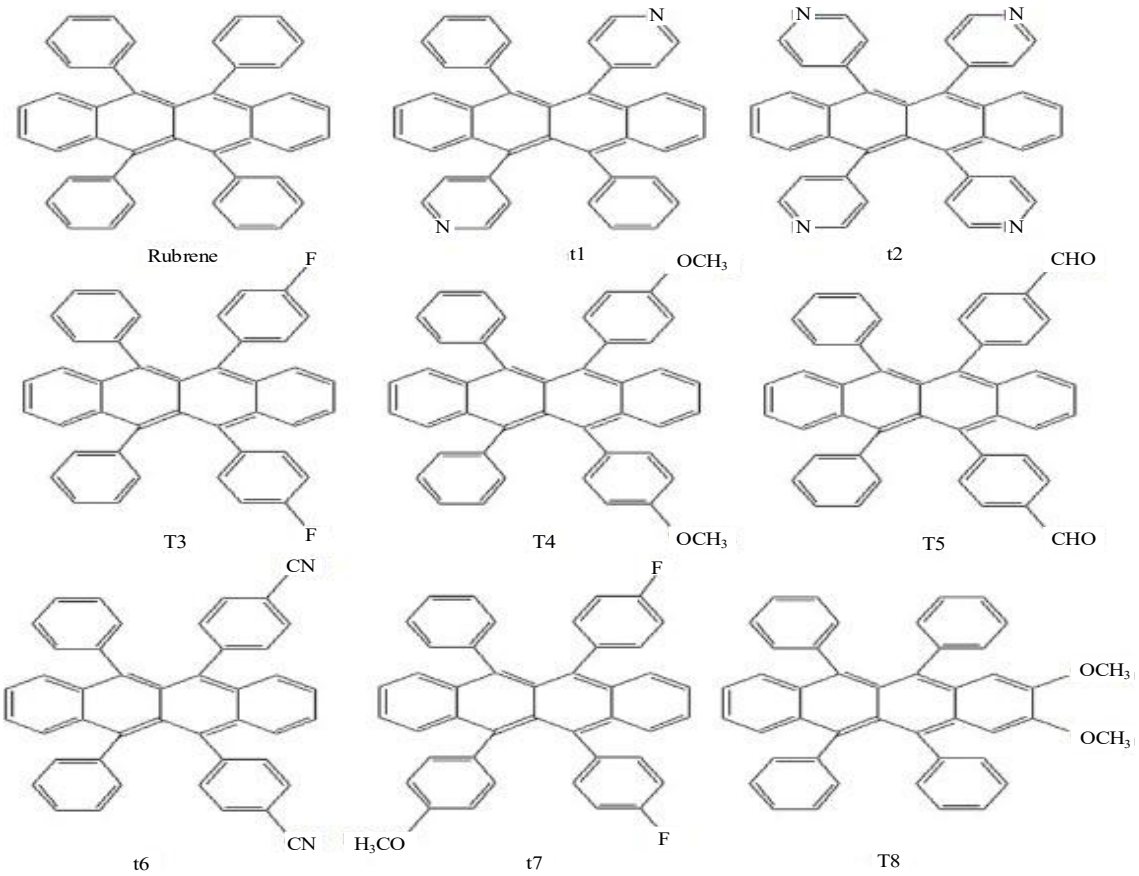
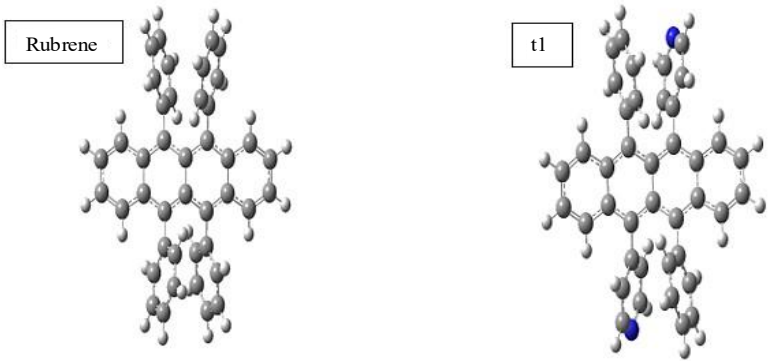


Figure 1. Molecular structure of rubrene and its derivatives.



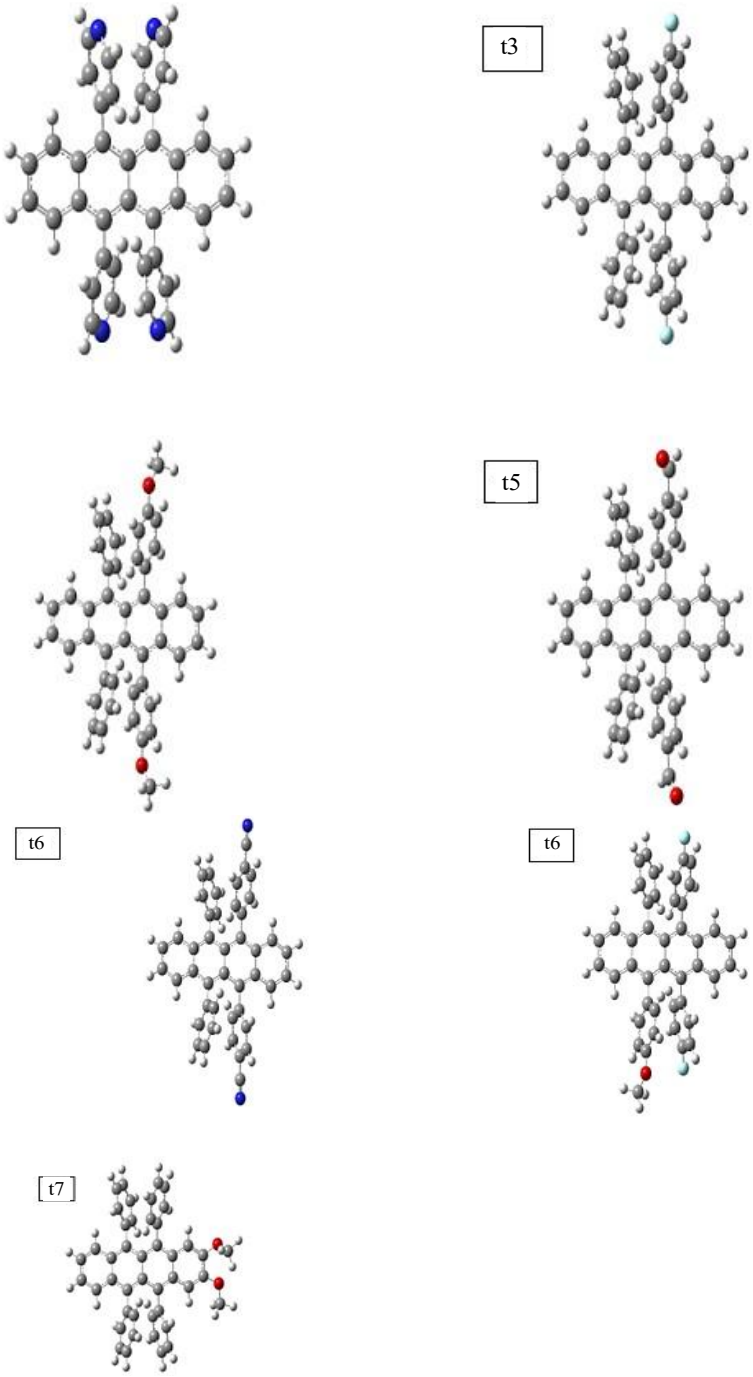


Figure 2. Optimized structures of the molecules.

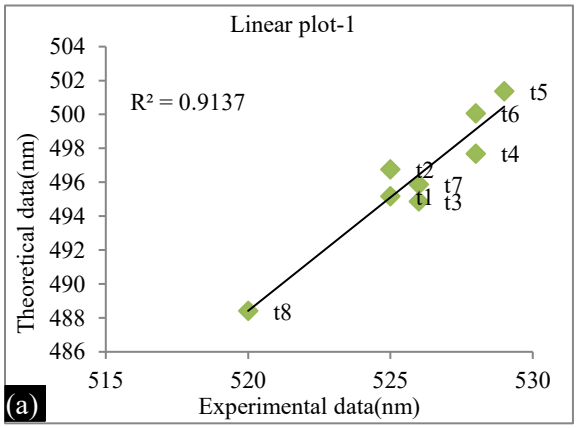
the rubrene derivatives and discovered that the optimized structures (Figure 3) exhibited zero imaginary frequencies, confirming their stability. To determine the absorption spectra of the rubrene derivatives in various conditions, including vacuum and solvent atmospheres, we utilized TD-DFT calculations. The values of the absorption maxima in each method are provided in Table 1 for reference. Upon analysis, it appears that the M06-2X/aug-cc-pVDZ method (solvent: DCM) shows the most agreement with the experimental data across the majority of the rubrene derivatives. This is indicated by the values being closest to the experimental λmax values for the majority of derivatives.

Based on this comparison, it can be suggested that the M062X/aug-cc-pVDZ method (solvent: DCM) demonstrates a strong correlation with the experimental absorbance data for rubrene

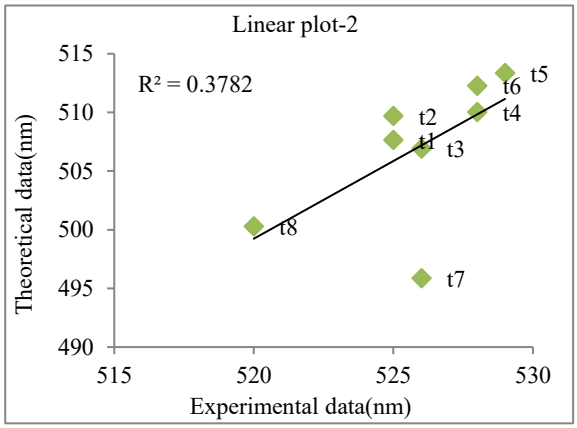
derivatives, making it a reliable option for predicting the absorbance maxima of rubrene derivatives in dichloromethane (DCM) solvent. The absorbance spectra obtained by this method is also shown in the Figure S1.

The (R^2) coefficient of determination values obtained from the linear regression plots (Figure 5) can provide insights into the strength of the correlation between the theoretical absorbance values and the experimental data. The R^2 value ranges from 0 to 1, with a value closer to 1 indicating a stronger correlation between the predicted and observed values.

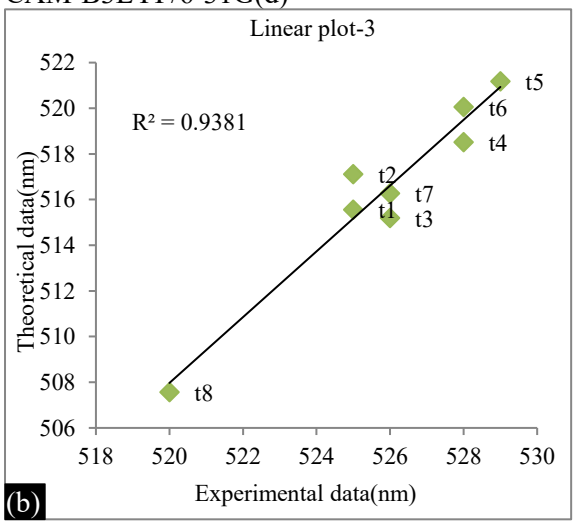
Upon examining the R^2 values for each method we can observe that CAM-B3LYP/cc-pVDZ (solvent: DCM), M062X/cc-pVDZ (solvent: DCM), and M062X/aug-cc-pVDZ (solvent: DCM) methods have relatively higher R^2 values of 0.9376 compared to the other methods.



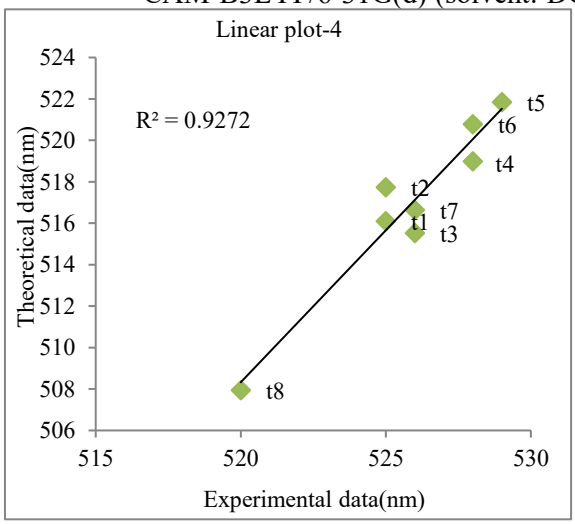
CAM-B3LYP/6-31G(d)



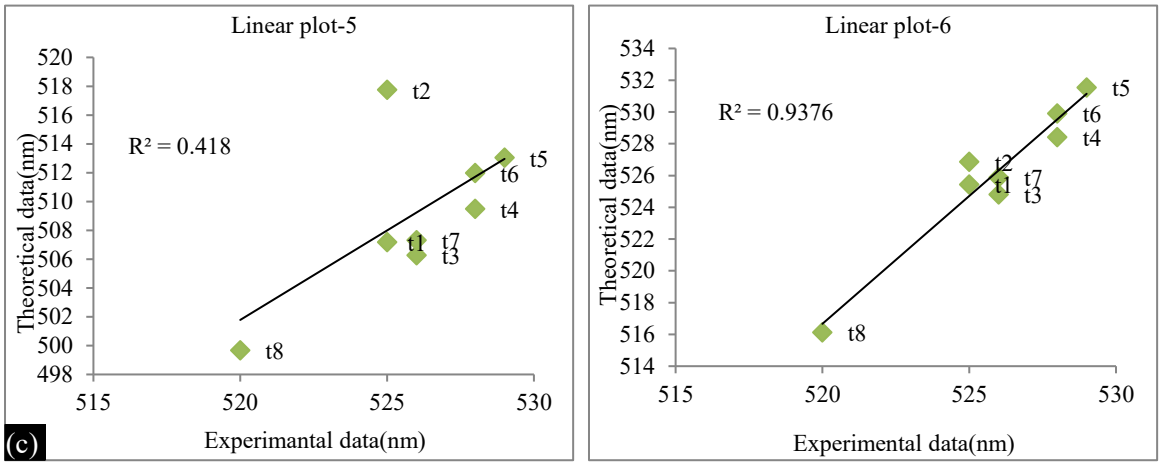
CAM-B3LYP/6-31G(d) (solvent: DCM)



CAM-B3LYP/cc-pVDZ (solvent: DCM)



M062X/cc-pVDZ (solvent: DCM



M062X/6-31G(d) (solvent: DCM) M062X/aug-cc-pVDZ (solvent: DCM):

Figure 3. (a–c) Linear regression plots by different methods.

Based on both the tabulated absorbance values and the R-squared (R^2) analysis, it is evident that the method M062X/aug-cc-pVDZ (solvent: DCM) demonstrates a stronger correlation with the experimental data. Hence, it can be considered as the benchmark method for further analysis. In conclusion, both the table and the R^2 value analysis highlight the robust correlation of M062X/aug-cc-pVDZ (solvent: DCM) with the experimental values, solidifying its reliability for predicting the absorbance maxima of rubrene derivatives.

Furthermore, the decision to exclude the absorbance values of parent rubrene from the analysis is justified. Parent rubrene was omitted because it does not exhibit significant push-pull effects as observed in its substituted derivatives. Therefore, its absorbance values do not align well with the rest of the derivatives, making it unsuitable for comparison and analysis in this context.

After identifying the method that correlates with the experimental data, several procedures were employed to determine the variables affecting the absorption maxima. From the checkpoint files of the eight derivatives of rubrene, the values of Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbital (LUMO) energies were obtained, from which several parameters were calculated using the equations outlined in the methodology. Additionally, values of various moments were obtained from the Multiwfn software. All these variables were tabulated for analysis (Table 2).

Principal Component analysis (Table S1) of this dataset provides valuable insights into the major components affecting the absorption maxima by reducing the dimensionality of data while preserving as much as variable as possible. The PCA model provided includes several components of interest:

Component loadings indicate how much each variable contributes to the principal components. In this PCA, two main components are considered: Component 1 has high loadings for HOMO (0.942), LUMO (0.880), Molecular Mass (0.909), and other variables, suggesting that it largely represents electron-related properties and molecular size. Component 2 is significantly influenced by properties like Band Gap (0.976), Chemical Potential (0.976), and Hardness (0.976), reflecting the stability and reactive capabilities of the molecules. The uniqueness of a variable indicates the proportion of the variance in that variable that isn't shared with other variables, and high values (e.g., Dipole Moment: 0.7188) suggest that not all variability is captured by the two main components.

The Bartlett's Test of Sphericity checks (Table S2) whether the correlation matrix is an identity matrix, which would indicate that variables are unrelated. The extremely low p-value (< 0.001) in this analysis suggests that there is a significant correlation among variables, justifying the use of PCA. The Kaiser-Meyer-Olkin (KMO) measure (Table S3) tests whether the partial correlations among

Table 1. Absorption maxima obtained by different methods.

Rubrene Derivatives	λ_{max} (expt)	CAM-B3LYP/ 6-31G(d)	CAM-B3LYP/ 6-31G(d) (solvent: DCM)	CAM-B3LYP/ cc-pVDZ (solvent: DCM)	M062X/ cc-pVDZ (solvent: DCM)	M062X/ 6-31G(d) (solvent: DCM)	M062X/aug-cc-pVDZ (solvent: DCM)
t1	525	495.17	507.65	515.56	516.10	507.17	525.43
t2	525	496.76	509.71	517.11	517.74	517.14	526.86
t3	526	494.88	506.92	515.20	515.52	506.26	524.8
t4	528	497.69	510.03	518.52	518.98	509.48	528.42
t5	529	501.37	513.37	521.18	521.84	513.03	531.53
t6	528	500.07	512.28	520.06	520.78	511.97	529.91
t7	526	495.88	495.87	516.27	516.64	507.03	525.88
t8	520	488.42	500.29	507.57	507.94	499.67	516.13

Table 2. Dataset obtained from different procedures.

Rubrene derivatives	**λmax**	**HOMO**	**LUMO**	**Band Gap**	**I**	**A**	**χ**	**μ**
t1	525.43	-6.12	-1.94	4.18	6.12	1.94	4.030	1.044
t2	526.86	-6.25	-2.08	4.17	6.25	2.08	4.166	1.042
t3	524.8	-6.02	-1.85	4.18	6.02	1.85	3.934	1.044
t4	528.42	-5.93	-1.77	4.15	5.93	1.77	3.850	1.038
t5	531.53	-6.08	-1.95	4.13	6.08	1.95	4.015	1.034
t6	529.91	-6.13	-1.99	4.15	6.13	1.99	4.058	1.036
t7	525.88	-5.99	-1.83	4.17	5.99	1.83	3.911	1.042
t8	516.13	-5.92	-1.70	4.21	5.92	1.70	3.809	1.053

Rubrene derivatives	**η**	**S**	**ω**	**ΔN_{max}**	**Dipole Moment**	**Energy**	**Molecular Mass**	**Quadrupole Moment**
t1	2.088	1.044	1.137	16.825	26.505	-44852.01	4.97758E+11	618.94
t2	2.085	1.042	1.133	17.370	0.009	-45724.56	4.99598E+11	1561.64
t3	2.088	1.044	1.137	16.427	2.822	-49377.08	5.29417E+11	599.13
t4	2.077	1.038	1.120	15.991	17.296	-50208.09	5.51827E+11	1381.82
t5	2.067	1.034	1.104	16.598	41.766	-50143.39	5.48073E+11	1395.66
t6	2.073	1.036	1.113	16.823	10.302	-48996.99	5.42486E+11	2605.86
t7	2.084	1.042	1.131	16.299	21.013	-52491.59	5.57371E+11	557.62
t8	2.107	1.053	1.169	16.050	29.835	-50207.91	5.51827E+11	756.77

Rubrene derivative	**Octopole Moment**	**Polariza-bility**	**Oscillator strength**	**ω+**	**ω-**	**Δ ω**	**$\Delta E_{Back\ Donation}$**	**N**	**σ_0**
t1	16549.9	18490.7	0.2588	37.224	107.47	-70.24	-0.522	0.879	-1.916
t2	11602.4	17957.3	0.2506	40.674	113.10	-72.42	-0.521	0.883	-1.919
t3	8138.30	18910.3	0.2639	34.888	103.48	-68.58	-0.522	0.879	-1.916
t4	11654.4	20327.9	0.2917	32.838	99.26	-66.42	-0.519	0.893	-1.926
t5	26484.5	20315.0	0.3065	36.743	105.37	-68.62	-0.517	0.906	-1.935
t6	22845.7	20340.9	0.3019	37.855	107.59	-69.73	-0.518	0.899	-1.930
t7	15228.5	19547.5	0.2761	34.314	102.23	-67.92	-0.521	0.884	-1.920
t8	7401.85	21174.7	0.2482	32.003	99.63	-67.62	-0.527	0.856	-1.899

variables are small. Here, a KMO value of 0.500 suggests that the sampling adequacy is borderline

acceptable. The eigenvalues measure (Table S4) the amount of variation retained by each principal component. The first two components have eigenvalues of 9.43763 and 6.09743, respectively, indicating they capture most of the variance in the data. Component 1 explains about 52.4313% of the variance, Component 2 covers an additional 33.8746%, cumulatively, about 86.3069% of the variance is explained by the first two components. This high percentage of explained variance suggests that these two components successfully capture most of the information in the dataset.

A scree plot (Figure 4) is used to determine the number of components to keep in a PCA model. It plots the eigenvalues in a descending order against the number of components. The 'elbow' in the scree plot typically indicates where the remaining components stop adding significant value. Here, the scree plot shows a clear elbow after the second component, reinforcing the decision to focus on two components.

Thus the high loadings of molecular properties related to electron behavior and size on Component 1 suggest that this component could be interpreted as representing the "electronic and size characteristics" of molecules. In contrast, Component 2 might be considered as encapsulating "stability and reactivity characteristics" due to its high loadings for properties like Band Gap and Hardness. The PCA performed on this dataset effectively reduces its dimensionality by condensing a large set of variables into two principal components that explain over 86% of the variance. These components provide significant insights into the electronic, size, stability, and reactivity characteristics of the molecules studied. In practical terms, this PCA can help chemists and researchers reduce the complexity of their data, allowing them to focus on two broad underlying factors that encapsulate most of the information in the original variables. This can be particularly useful in initial screens of molecular datasets where researchers are trying to identify underlying patterns or groupings in molecular properties.

After identifying the principal components that affects the absorption maxima, the dataset was analyzed using Jamovi software, a statistical tool designed to identify significant variables. Model fit measures (Table S5) were then assessed to gain insights into the relationship between variables and the absorption maxima of rubrene derivatives. The coefficient of determination (R-squared) indicated that over 96% of the variability in absorption maxima can be attributed to changes in certain variables, including band gap, optical softness, chemical potential, chemical hardness and softness, electrophilicity index, $\Delta E_{\text{Back donation}}$, and nucleophilicity index. This suggests a strong linear association between absorption maxima and these variables. Adjusted R-squared values remained high for models incorporating the aforementioned independent variables, indicating that the model's explanatory power was not inflated by unnecessary variables. The root mean square error (RMSE) values, such as 0.764 for band gap, indicated relatively small average deviations between observed and predicted values, suggesting a good model fit. This pattern was consistent across variables like Optical softness, Chemical pot, chemical hardness, chemical softness electrophilicity index, $\Delta E_{\text{Back donation}}$, and nucleophilicity index.

The overall model F-test assessed whether the regression model as a whole provided a better fit to the data than a model with no predictors. Variables with p-values less than the significance level (typically 0.05) were deemed statistically significant. In this analysis, Band Gap, Optical softness, chemical pot, chemical hardness, chemical softness electrophilicity index , $\Delta E_{\text{Back donation}}$, and nucleophilicity index, octopole moment, and oscillator strength were identified as significant variables, with band gap, optical softness, chemical pot, chemical hardness, chemical softness electrophilicity index , $\Delta E_{\text{Back donation}}$, and nucleophilicity index being particularly significant with p-values less than .001.

The linear regression analysis found a clear link between absorption maxima and certain key traits like band gap, optical softness, chemical potential, chemical hardness and softness, electrophilicity index, ΔE Back donation, and nucleophilicity index. These findings offer valuable insights into how

the absorption properties of rubrene derivatives can be influenced by specific molecular features. Moreover, this knowledge can be utilized to develop a multiple linear regression model for predicting absorption maxima, aiding further research and development in this area.

After conducting multiple regression analysis using a dataset containing absorption maxima as the dependent variable and Band Gap, Optical Softness, ΔE Back Donation, Nucleophilicity Index, Softness, Hardness, Chemical Potential, and Electrophilicity Index as independent variables, a linear regression model was obtained as follows. (Please see Table S6 for data)

Absorption Maxima = -11491.5-278.5 (Band Gap) + 30.0 (Optical Softness) + 24790.0 (ΔE Back Donation) + 3510.0 (Nucleophilicity Index) + 13200.0 (Chemical Softness) + 4460.0 (Chemical Hardness) (17)

Here, (Table S7) the intercept (-11491.5) denotes the baseline Absorption maxima when all predictors are zero. Band Gap, with a coefficient of -278.5, suggests a decrease in Absorption maxima with increasing Band Gap, indicating narrower band gaps are favorable for higher absorption maxima. Optical Softness contributes positively, although to a lesser extent (coefficient: 30.0), implying that higher softness slightly increases the absorption capacity. ΔE Back Donation exhibits a significant positive effect (coefficient: 24790.0), the largest among all predictors, highlighting its strong influence on increasing Absorption maxima. Nucleophilicity Index, Chemical Softness, and Chemical Hardness also have a positive influence, indicating higher values in these descriptors are associated with higher Absorption maxima. However, Chemical Potential and Electrophilicity Index could not be evaluated due to computational issues (NaN values), likely reflecting multicollinearity or data insufficiency. The model fit measures (Table S7) R^2 (0.999) and Adjusted R^2 (0.991) indicate an excellent fit, with the model explaining nearly all the variance in Abs Max among molecules.

Thus the analysis reveals that certain molecular descriptors significantly influence the Absorption maxima. The positive coefficients for ΔE Back Donation, Nucleophilicity Index, Chemical Softness, and Chemical Hardness suggest that molecules with higher values of these properties tend to have higher Absorption maxima, which is crucial for materials used in light-absorbing applications. The negative coefficient for Band Gap aligns with theoretical expectations: a smaller band gap typically allows for easier electron excitation, which corresponds to higher light absorption. The negligible impact of Optical Softness and the undefined coefficients for Chemical Potential and Electrophilicity Index call for a reconsideration of the data quality or experimental design, potentially adjusting for multicollinearity or collecting more comprehensive data.

The 17 test molecules (Figure S2), selected to assess the efficiency of the linear regression equation obtained, were optimized using the M062X/aug-cc-pVDZ method (solvent: DCM). The optimized structures of these molecules are provided (Figure S3). Similar to the molecules used in the training set, the absorption maxima of each test molecule were determined using TD-DFT analysis. Additionally, the six descriptors identified as major contributors to absorption maxima (ΔE Back Donation, Nucleophilicity Index, Chemical Softness, Chemical Hardness, Band Gap, and Optical Softness) were obtained through FMO analysis (Table 3). The absorption maxima (modelled absorption maxima) predicted by the linear regression equation, using these descriptors for the test molecules, showed strong correlation with the theoretically calculated absorption maxima.

The agreement between the modelled and calculated λmax values was found to be particularly robust, with differences of approximately ±20 nm. This level of agreement is further supported by the R^2 value of 0.7512, (Figure 5) indicating that the multiple linear regression model derived from known molecular data can predict the absorption maxima of molecules without much variance.

These findings underscore the utility and reliability of the developed regression model in predicting the absorption maxima of molecules beyond those used in the training set. Such predictive capabilities

hold significant promise for guiding the design and optimization of novel molecules for various applications, particularly in fields reliant on precise control over optical properties, such as OLEDs. Moreover, the close agreement between modelled and theoretically calculated absorption maxima highlights the robustness and accuracy of the computational approach employed in this study. Overall, these results contribute to advancing our understanding of the relationships between molecular descriptors and optical properties, facilitating informed molecular design strategies in chemoinformatics and material science.

Table 3. Test molecule dataset.

Test molecule	λ max	Band Gap	Hardness	Softness	ΔE Back Donation	Nucleophilicity Index	Optical Softness	Modelled λ max
Ts1	520.23	4.15	2.074	1.037	-0.518	0.897	-1.929	527.84
Ts2	517.02	4.17	2.084	1.042	-0.521	0.884	-1.919	526.00
Ts3	516.94	4.17	2.085	1.042	-0.521	0.883	-1.919	525.95
Ts4	510.4	4.20	2.101	1.051	-0.525	0.863	-1.904	524.98
Ts5	514.58	4.18	2.091	1.046	-0.523	0.875	-1.913	525.28
Ts6	515.89	4.17	2.086	1.043	-0.522	0.881	-1.917	525.73
Ts7	520.29	4.15	2.074	1.037	-0.519	0.896	-1.928	527.70
Ts8	508.22	4.21	2.105	1.052	-0.526	0.858	-1.900	525.08
Ts9	517.81	4.16	2.080	1.040	-0.520	0.889	-1.923	526.67
Ts10	513.87	4.18	2.091	1.046	-0.523	0.875	-1.913	525.30
Ts11	531.76	4.09	2.046	1.023	-0.511	0.934	-1.955	537.58
Ts12	525.87	4.07	2.037	1.019	-0.509	0.946	-1.963	541.92
Ts13	525.63	4.11	2.055	1.027	-0.514	0.922	-1.947	533.63
Ts14	522.96	4.11	2.056	1.028	-0.514	0.920	-1.945	533.16
Ts15	516.79	4.17	2.086	1.043	-0.521	0.882	-1.918	525.80
Ts16	518.27	4.16	2.081	1.041	-0.520	0.887	-1.922	526.39
Ts17	527.4	4.10	2.049	1.024	-0.512	0.931	-1.953	536.26

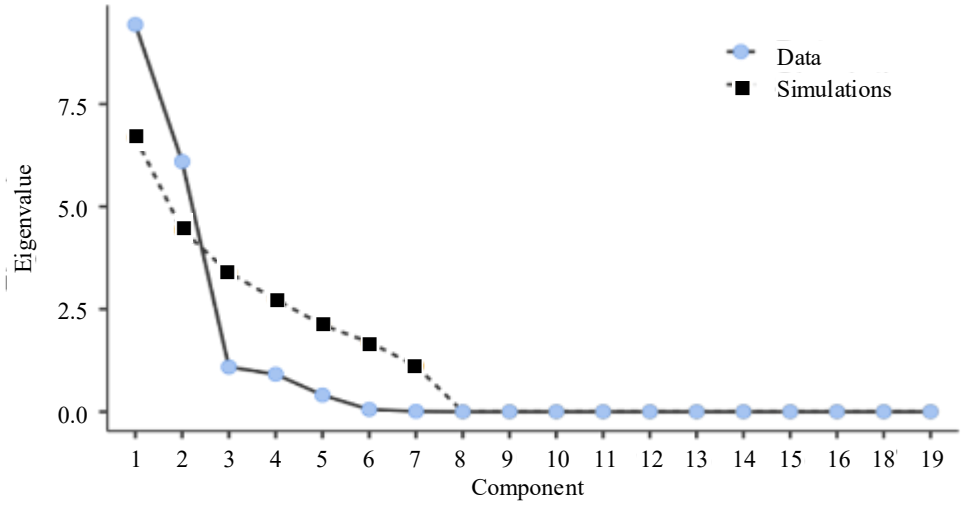


Figure 4. Scree plot.

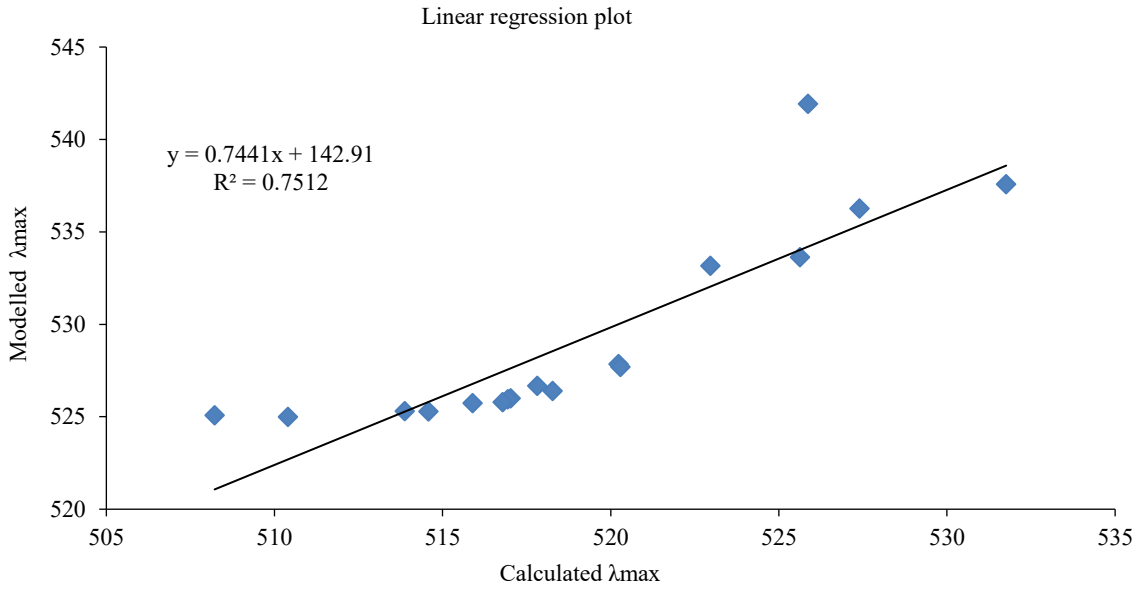


Figure 5. Comparison plot of modelled and calculated abs max.

CONCLUSION

In this interdisciplinary study, we employed descriptive statistics and computational chemistry techniques to delve into the OLED properties of rubrene and its derivatives. Through meticulous analysis and model development, we identified key molecular descriptors influencing absorption maxima, unveiling a clear relationship between various traits and optical behavior. Our developed multiple linear regression model exhibited a robust predictive capability, validated by average agreement with experimental data for newly designed molecules. This model, exhibiting an R^2 value of 0.7512, showcases average efficiency in predicting absorption maxima, indicating its average reliability beyond the training set. These findings not only enhance our comprehension of rubrene's OLED properties but also offer a pathway for the tailored design of novel materials with specific optical characteristics crucial for advancements in organic electronics, photovoltaics, and optoelectronic devices. Future work should focus on expanding the dataset for validation, exploring other computational methods, and addressing the limitations identified. The integration of machine learning and computational chemistry presents a promising route for accelerating materials discovery and development, enabling efficient screening and analysis of molecular structures to drive innovation in materials science and technology.

Acknowledgments

Renjith Thomas thanks the Royal Society of Chemistry, London, for the RSC Researcher Collaborations Grant (C23-1708759250). SEAGrid (http:www.seagrid.org) is acknowledged for computational resources and services for the selected results used in this publication.

Author Credit

Renjith Thomas: Conception, Methods, Simulations, Resources, Supervision, Writing manuscript.

Anjana KS: Simulations, Execution, Data validation, Writing manuscript.

Rencemon T Thoppil: Simulations, Execution, Data validation, Writing manuscript.

Anila Skariah: Descriptive Statistics and Data Analysis, Writing Manuscript

Sneha Anna Sunny: Simulations, Writing manuscript.

Data Availability Statement

The authors declare that the data supporting the findings of this study are available within the paper and its Supplementary Information files. Should any raw data files be needed in another format they are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical statement

No human trials or animal trials are conducted to get data for this manuscript

REFERENCES

1. Huang B, Symonds NO, Von Lilienfeld OA. Quantum Machine Learning in Chemistry and Materials. In: Andreoni W, Yip S, eds. Handbook of Materials Modeling. Springer International Publishing; 2020:1883–1909. doi:10.1007/978-3-319-44677-6_67
2. Badillo S, Banfai B, Birzele F, et al. An Introduction to Machine Learning. Clin Pharma and Therapeutics. 2020;107(4):871-885. doi:10.1002/cpt.1796
3. Rebala G, Ravi A, Churiwala S. Machine Learning Definition and Basics. In: An Introduction to Machine Learning. Springer International Publishing; 2019:1–17. doi:10.1007/978-3-030-15729-6_1
4. Principal Component Analysis. Springer-Verlag; 2002. doi:10.1007/b98835
5. Perros HG. An Introduction to IoT Analytics. First edition. CRC Press, Taylor & Francis Group; 2021.
6. Tang CW, VanSlyke SA. Organic electroluminescent diodes. Applied Physics Letters. 1987;51(12):913-915. doi:10.1063/1.98799
7. Tsujimura T. OLED Displays: Fundamentals and Applications. 1st ed. Wiley; 2012. doi:10.1002/9781118173053
8. Song J, Lee H, Jeong EG, Choi KC, Yoo S. Organic Light-Emitting Diodes: Pushing Toward the Limits and Beyond. Advanced Materials. 2020;32(35):1907539. doi:10.1002/adma.201907539
9. Murawski C, Gather MC. Emerging Biomedical Applications of Organic Light-Emitting Diodes. Advanced Optical Materials. 2021;9(14):2100269. doi:10.1002/adom.202100269
10. Tarnowski T, Kreuzer M, Haidenthaler R, Aichholz M, Pohl M, Pross A. 61-1: Invited Paper: OLED Technology for Automotive Display Applications. Symp Digest of Tech Papers. 2022;53(1):794-797. doi:10.1002/sdtp.15611
11. Klauk H. Organic Electronics: Materials, Manufacturing and Applications. Wiley-VCH; 2006.
12. Petrenko T, Krylova O, Neese F, Sokolowski M. Optical absorption and emission properties of rubrene: insight from a combined experimental and theoretical study. New J Phys. 2009;11(1):015001. doi:10.1088/1367–2630/11/1/015001
13. Hasegawa T, Takeya J. Organic field-effect transistors using single crystals. Science and Technology of Advanced Materials. 2009;10(2):024314. doi:10.1088/1468–6996/10/2/024314
14. Wang Z, Naka S, Okada H. Performance improvement of rubrene-based organic light emitting devices with a mixed single layer. Appl Phys A. 2010;100(4):1103–1108. doi:10.1007/s00339-010-5710-4
15. Wang S, Kirch A, Sawatzki M, et al. Highly Crystalline Rubrene Light-Emitting Diodes with Epitaxial Growth. Adv Funct Materials. 2023;33(14):2213768. doi:10.1002/adfm.202213768
16. Saikia D, Sarma R. Characterization of organic light-emitting diode using a rubrene interlayer between electrode and hole transport layer. Bull Mater Sci. 2020;43(1):35. doi:10.1007/s12034-019-2003-1
17. Keith JA, Vassilev-Galindo V, Cheng B, et al. Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. Chem Rev. 2021;121(16):9816–9872. doi:10.1021/acs.chemrev.1c00107
18. Hwang, Tae-Kyu, Kim, Ju-Young, Lee, Sungyul. Quantum Chemical Study of the OLED Materials Tris[4’-(1"-phenylbenzimidazol-2"-yl)phenyl] Derivatives of Amine and Benzene. Bulletin of the Korean Chemical Society. 2011;32(5):1733–1736.

doi:10.5012/BKCS.2011.32.5.1733
19. Hoffman DM, Johnson PV, Kim JS, Vargas AD, Banks MS. 240 Hz OLED technology properties that can enable improved image quality. J Soc Info Display. 2014;22(7):346–356. doi:10.1002/jsid.258
20. Valencia-Marquez D, Flores-Tlacuahuac A. Improving molecular design through a machine learning approach. Chemical Engineering and Processing - Process Intensification. 2020;158:108173. doi:10.1016/j.cep.2020.108173
21. Liu H, Yao X, Gramatica P. The Applications of Machine Learning Algorithms in the Modeling of Estrogen-Like Chemicals. CCHTS. 2009;12(5):490-496. doi:10.2174/138620709788489037
22. Makjan S, Promkatkaew M, Hannongbua S, Boonsri P. Theoretical Study of the Electronic Structure and Properties of Alternating Donor-Acceptor Couples of Carbazole-Based Compounds for Advanced Organic Light-Emitting Diodes (OLED). KEM. 2019;824:236–244. doi:10.4028/www.scientific.net/KEM.824.236
23. Ullah A, Hossain MdR, Chawdhury N. Theoretical Investigation of Optoelectronic Properties of Aryl Substituted Pentacene Derivatives. In: 2022 International Conference on Recent Progresses in Science, Engineering and Technology (ICRPSET). IEEE; 2022:1–4. doi:10.1109/ICRPSET57982.2022.10188533
24. Jin R, Ahmad I. Theoretical study on photophysical properties of multifunctional star-shaped molecules with 1,8-naphthalimide core for organic light-emitting diode and organic solar cell application. Theor Chem Acc. 2015;134(7):89. doi:10.1007/s00214-015-1693-8
25. Ji LF, Fan JX, Qin GY, Zhang NX, Lin PP, Ren AM. Theoretical Study on the Electronic Structures and Charge Transport Properties of a Series of Rubrene Derivatives. J Phys Chem C. 2018;122(37):21226-21238. doi:10.1021/acs.jpcc.8b07018
26. Zhang M, Hua Z, Liu W, et al. A DFT study on the photoelectric properties of rubrene and its derivatives. J Mol Model. 2020;26(2):32. doi:10.1007/s00894-020-4295-x
27. Zhao L, Yang G, Su Z, Qin C, Yang S. Theoretical studies on one- and two-photon absorption properties of rubrene and its derivatives. Synthetic Metals. 2006;156(18-20):1218-1224. doi:10.1016/j.synthmet.2006.09.006
28. Yang M, Liu X, Luo Y, et al. Machine learning-enabled non-destructive paper chromogenic array detection of multiplexed viable pathogens on food. Nat Food. 2021;2(2):110-117. doi:10.1038/s43016-021-00229-5
29. Liu Y, Zhao T, Ju W, Shi S. Materials discovery and design using machine learning. Journal of Materiomics. 2017;3(3):159-177. doi:10.1016/j.jmat.2017.08.002
30. Artrith N, Butler KT, Coudert FX, et al. Best practices in machine learning for chemistry. Nat Chem. 2021;13(6):505-508. doi:10.1038/s41557-021-00716-z
31. Stefani R. State of the Art and of Outlook of Data Science and Machine Learning in Organic Chemistry. Published online February 13, 2023. doi:10.26434/chemrxiv-2023-p7hdw
32. Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ. Machine Learning in Drug Discovery: A Review. Artif Intell Rev. 2022;55(3):1947-1999. doi:10.1007/s10462-021-10058-4
33. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-477. doi:10.1038/s41573-019-0024-5
34. Burés J, Larrosa I. Organic reaction mechanism classification using machine learning. Nature. 2023;613(7945):689-695. doi:10.1038/s41586-022-05639-4
35. Segler MHS, Waller MP. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. Chemistry A European J. 2017;23(25):5966-5971. doi:10.1002/chem.201605499
36. Saeki A, Kranthiraja K. A high throughput molecular screening for organic electronics via machine learning: present status and perspective. Jpn J Appl Phys. 2020;59(SD):SD0801. doi:10.7567/1347-4065/ab4f39
37. Zhao Y, Fu C, Fu L, Lu Z, Pu X. Data-driven machine learning models for the quick and accurate prediction of thermal stability properties of OLED materials. Published online August 3, 2021. doi:10.26434/chemrxiv-2021-j5pfd-v3
38. Joshi PB. Navigating with chemometrics and machine learning in chemistry. Artif Intell Rev. 2023;56(9):9089-9114. doi:10.1007/s10462-023-10391-w

39. Paraskar AS, Reddy AR, Patra A, et al. Rubrenes: Planar and Twisted. Chemistry A European J. 2008;14(34):10639-10647. doi:10.1002/chem.200800802
40. Xie G. Synthesis and Characterization of Azaacenes and Stable Azaacene Radical Cations. Published online 2020. doi:10.11588/HEIDOK.00028779
41. Marill KA. Advanced Statistics: Linear Regression, Part II: Multiple Linear Regression. Academic Emergency Medicine. 2004;11(1):94-102. doi:10.1197/j.aem.2003.09.006
42. Draper NR, Smith H. Applied Regression Analysis. 1st ed. Wiley; 1998. doi:10.1002/9781118625590
43. Tong Q, Gao P, Liu H, et al. Combining Machine Learning Potential and Structure Prediction for Accelerated Materials Design and Discovery. J Phys Chem Lett. 2020;11(20):8710-8720. doi:10.1021/acs.jpclett.0c02357
44. Huang JJ, Lin YY, Lee CL, Lai CC, Chou CY, Lin CC. P-83: Optimizing Water-proofing Algorithm with Machine Learning for Wearable OLED In-Cell Touch Panel. Symp Digest of Tech Papers. 2023;54(1):1664-1666. doi:10.1002/sdtp.16917
45. Nomura Y, Akashi R. Density functional theory. In: Encyclopedia of Condensed Matter Physics. Elsevier; 2024:867-878. doi:10.1016/B978-0-323-90800-9.00148-7
46. Sholl DS, Steckel JA. Density Functional Theory: A Practical Introduction. 1st ed. Wiley; 2009. doi:10.1002/9780470447710
47. Cramer CJ. Essentials of Computational Chemistry: Theories and Models. 2nd ed. Wiley; 2004.
48. Frisch MJ, Trucks GW, Schlegel HB, et al. Gaussian˜16 Revision C.01. Published online 2016.
49. Dunning TH. Gaussian Basis Functions for Use in Molecular Calculations. I. Contraction of (9s5p) Atomic Basis Sets for the First-Row Atoms. The Journal of Chemical Physics. 1970;53(7):2823-2833. doi:10.1063/1.1674408
50. Lu T, Chen F. Multiwfn: A multifunctional wavefunction analyzer. J Comput Chem. 2012;33(5):580-592. doi:10.1002/jcc.22885
51. Eryılmaz S. The theoretical investigation of global reactivity descriptors, NLO behaviours and bioactivity scores of some norbornadiene derivatives. Sakarya University Journal of Science. 2018;22(6):1638-1647. doi:10.16984/saufenbilder.359837
52. Ejidike IP, Direm A, Parlak C, et al. Spectroscopic characterization, DFT calculations, in vitro pharmacological potentials, and molecular docking studies of N, N, O-Schiff base and its trivalent metal complexes. Chemical Physics Impact. 2024;8:100549. doi:10.1016/j.chphi.2024.100549
53. Danish IA, Kores JJ, Chelliah DA, Sankar TB, Jebaraj JW. In silico analyses of solvent effects, toxicity, NBO, homo-lumo and hole-electron transfer of 7-hydroxy-2-nitrofluoranthene. Journal of the Indian Chemical Society. 2024;101(5):101147. doi:10.1016/j.jics.2024.101147
54. Gázquez JL, Cedillo A, Vela A. Electrodonating and Electroaccepting Powers. J Phys Chem A. 2007;111(10):1966-1970. doi:10.1021/jp065459f
55. Brinzei M, Stefaniu A, Iulian O, Ciocirlan O. Density Functional Theory (DFT) and Thermodynamics Calculations of Amino Acids with Polar Uncharged Side Chains. In: The 24th International Electronic Conference on Synthetic Organic Chemistry. MDPI; 2020:56. doi:10.3390/ecsoc-24-08420
56. Ramirez-Balderrama K, Orrantia-Borunda E, Flores-Holguin N. Calculation of global and local reactivity descriptors of carbodiimides, a DFT study. J Theor Comput Chem. 2017;16(03):1750019. doi:10.1142/S0219633617500195
57. Wehrens R. Principal Component Analysis. In: Chemometrics with R. Use R! Springer Berlin Heidelberg; 2020:45-68. doi:10.1007/978-3-662-62027-4_4
58. Gnecco G, Bacigalupo A, Fantoni F, Selvi D. Principal Component Analysis Applied to Gradient Fields in Band Gap Optimization Problems for Metamaterials. J Phys: Conf Ser. 2021;2015(1):012047. doi:10.1088/1742-6596/2015/1/012047
59. Belsley DA, Kuh E, Welsch RE. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. 1st ed. Wiley; 1980. doi:10.1002/0471725153
60. The jamovi project (2022). https://www.jamovi.org/
61. Gillio Meina E, Niyogi S, Liber K. Multiple Linear Regression Modeling Predicts the Effects of

Surface Water Chemistry on Acute Vanadium Toxicity to Model Freshwater Organisms. Enviro Toxic and Chemistry. 2020;39(9):1737-1745. doi:10.1002/etc.4798
62. Gupta SKr, Agarwal AP. Predicting Total Sugar Production Using Multivariable Linear Regression. In: 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). IEEE; 2021:465–469. doi:10.1109/ICCCIS51004.2021.9397078
63. Hahs-Vaughn DL, Lomax RG. An Introduction to Statistical Concepts. Fourth edition. Routledge, Taylor & Francis Group; 2020.
64. Jihad A, Nuraida I, Wutsqo SU. The prediction analysis model using the simple linear regression methods. In:; 2023:040059. doi:10.1063/5.0139381
65. Sutton C, Tummala NR, Beljonne D, Brédas JL. Singlet Fission in Rubrene Derivatives: Impact of Molecular Packing. Chem Mater. 2017;29(7):2777-2787. doi:10.1021/acs.chemmater.6b04633

