Journal of Molecular Modeling (2025) 31:18
https://doi.org/10.1007/s00894-024-06240-4

ORIGINAL PAPER

Chemical feature-based machine learning model for predicting photophysical properties of BODIPY compounds: density functional theory and quantitative structure–property relationship modeling

Gerardo M. Casanola-Martin[1] · Jing Wang[2] · Jian-ge Zhou[2] · Bakhtiyor Rasulev[1] · Jerzy Leszczynski[2]

Received: 15 August 2024 / Accepted: 28 November 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

Context Boron-dipyrromethene (BODIPY) compounds have unique photophysical properties and have been applied in fluorescence imaging, sensing, optoelectronics, and beyond. In order to design effective BODIPY compounds, it is crucial to acquire a comprehensive understanding of the relationships between the structures of BODIPY and the corresponding photoproperties. Fifteen molecular descriptors were identified to be strongly correlated with the maximum absorption wavelength. The developed ML/QSPR model exhibited good predictive performance, with coefficients of determination (R^2) of 0.945 for the training set and 0.734 for the test set, demonstrating robustness and reliability. A posterior analysis of some of the selected descriptors in the model provided insights into the structural features that influence BODIPY compound properties; meanwhile, it also emphasizes the importance of molecular branching, size, and specific functional groups. This work shows that applied combined cheminformatics and machine learning approach is robust to screen the BODIPY compounds and design novel structures with enhanced performance.

Methods In the present study, all the BODIPY models studied were fully optimized, and the corresponding absorption spectrum was obtained at DFT/TDDFT//B3LYP/6-311G(d,p) level. All the above calculations were executed by the Gaussian 16 program. Based upon the theoretical computational results, the machine learning-based quantitative structure–property relationship (ML/QSPR) model was employed for predicting the maximum absorption wavelength (λ) of BODIPY compounds by combining hand-crafted molecular descriptors (MD) and explainable machine learning (EML) techniques using Scikit-learn python library. A dataset of 131 BODIPY compounds with their experimental photophysical properties was used to generate a diverse set of molecular descriptors capturing information about the size, shape, connectivity, and other structural features of these compounds using Chemaxon and Alvadesc software. A genetic algorithm (GA) variable selection together with the multi-linear regression (MLR) method were applied to develop the best predictive model using the Genetic Selection python library.

Keywords BODIPY · DFT · TDDFT · Explainable machine learning · Absorption wavelength

Introduction

4,4-Difluoro-4-bora-3a,4a-diaza-s-indacene (or boron pyrromethene, or bora-indacene, BODIPY) dyes are small molecules with strong UV absorption that emit relatively sharp fluorescence peaks with high quantum yields [1]. They were first discovered by Treibs and Kreuzer in 1968 [2]. BODIPY dyes have since been applied as labeling reagents [3, 4], fluorescent switches [5–7], chemosensors [8, 9], and laser dyes [10]. Their high quantum yields, photostability, and tunable absorption wavelengths make them invaluable tools for visualizing biological structures and dynamics with high sensitivity and precision. Additionally, BODIPY

This manuscript is dedicated to Professor Alejandro Toro-Labbé, on the occasion of his 70th birthday.

✉ Bakhtiyor Rasulev
bakhtiyor.rasulev@ndsu.edu

✉ Jerzy Leszczynski
jerzy@icnanotox.org

[1] Department of Coatings and Polymeric Materials, North Dakota State University, Fargo, ND 58102, USA

[2] Interdisciplinary Center for Nanotoxicity, Department of Chemistry, Physics, and Atmospheric Sciences, Jackson State University, Jackson, MS 39217, USA

Published online: 12 December 2024

compounds have been employed in the development of fluorescent sensors for detecting various analytes, including metal ions, pH, and reactive oxygen species [11–13]. Their selective response to specific analytes, coupled with their robust photophysical properties, enables the design of highly sensitive and selective sensing platforms for applications in environmental monitoring, medical diagnostics, and biological research [14–16].

In addition to fluorescence imaging and sensing, BODIPY compounds have shown promise in optoelectronic devices, such as organic light-emitting diodes (OLEDs) and organic photovoltaics (OPVs) [17, 18]. Their excellent charge transport properties, high luminescence efficiency, and facile synthetic modification make them attractive candidates for use as active materials in these devices. BODIPY-based OLEDs have demonstrated impressive electroluminescence performance, with high brightness, low operating voltages, and narrow absorption spectra, making them suitable for display and lighting applications. Furthermore, BODIPY derivatives have been incorporated into organic photovoltaic devices as electron-accepting materials, where they contribute to enhanced light absorption, efficient charge generation, and improved device stability [19, 20]. The versatility of BODIPY compounds, combined with their favorable photophysical and electronic properties, continues to drive their exploration and application in diverse areas of science and technology [21, 22].

One of the primary challenges in designing new BODIPY (boron-dipyrromethene) compounds lies in achieving an adequate balance between synthetic accessibility and desired photophysical properties. BODIPY derivatives often demand intricate synthetic routes, necessitating careful consideration of reaction conditions, regioselectivity, and functional group compatibility. Additionally, tuning the photophysical characteristics of BODIPY molecules, such as absorption and emission wavelengths, quantum yields, and fluorescence lifetimes, presents a formidable challenge due to the complex interplay of molecular structure, electronic effects, and environmental factors. Achieving optimal stability under various conditions, expanding functional diversity beyond fluorescence imaging, and predicting biological activity further enhance the challenges in BODIPY compound design, necessitating innovative synthetic strategies and interdisciplinary collaborations.

Machine learning (ML) techniques offer a promising avenue to address the challenges associated with designing BODIPY compounds [23–26]. For example, a recent study carried out by Chebotaev et al. [27] found that data of 70 BODIPY structures in polar (45) and non-polar (39) have R^2 between 0.73 and 91 with non-linear ML algorithms like random forest and support vector machine, but the mechanistic explanation does not show the structure–property relationship in the fragment descriptors. Another interesting study was done by Buglak et al. [28], where three datasets of 48, 45, and 41 were used to develop three models with R^2 values between 0.88 and 0.91, using multi-linear regression (MLR) as the fitting algorithm. However, the mechanistic interpretation does not show the variations in the structure–property relationship. Besides, the above-mentioned works do not use custom-generated libraries based on the fragments extracted from the models.

Based on that, in the present study, we built a large dataset comprising BODIPY structures and their corresponding photophysical properties and then used ML models to learn the complex relationships between molecular features and desired properties. These models can then be used to predict the photophysical characteristics of novel BODIPY compounds, thereby accelerating the compound design process. Moreover, ML algorithms such as neural networks, random forests, and support vector machines can identify those complex hidden patterns and correlations that may escape traditional structure–property relationships [29–32], enabling the discovery of innovative BODIPY structures with tailored properties.

Furthermore, ML-driven virtual screening approaches [33–35] can expedite the identification of lead BODIPY candidates with desired characteristics. By screening vast chemical space and prioritizing compounds with high predicted activity or desirable physicochemical properties, ML algorithms can guide experimental efforts toward the synthesis of promising candidates, thereby reducing time and resource requirements. Moreover, ML models can facilitate the exploration of structure–activity relationships and guide the rational modification of BODIPY scaffolds to optimize specific properties. Through the integration of ML techniques into the BODIPY compound design workflow, researchers can accelerate the discovery of novel compounds with enhanced photophysical properties and diverse applications, unlocking new opportunities in fields such as fluorescence imaging, sensing, and optoelectronics.

Material and methods

Data collection

The experimental values of the λ(nm) of 131 different BODIPY compounds were collected from the literature [36], and the data values are shown in Table S1 in Supplementary Information. The data set of these BODIPY compounds was split into training (105 chemicals) and prediction (26 compound) sets representing approximately 80% and 20% of the total data. The data set was divided by sorting in descending order the data according to the response variable λ(nm) and selecting in a ratio 4:1 for training: prediction sets.

Molecular DFT structure optimization and generation

Density functional theory (DFT) with hybrid exchange correlation functional B3LYP [37–40] was applied in the present study, with a basis set of the standard triple zeta augmented by polarization functions 6-311G(d,p) [41–43]. To simulate the solvated environment of the corresponding experimental conditions, the Barone-Tomasi polarizable continuum model (PCM) with the related dielectric constant of the solvents was applied [44]. The ground state geometries of the studied BODIPY models were fully optimized using the above-mentioned theoretical level. The harmonic vibrational frequencies were analyzed to prove that all considered structures represent minimum energy geometries. Time-dependent density functional theory (TDDFT) [45] has been developed to theoretically study excitation energies, absorption wavelengths, and oscillator strengths. TDDFT has proved to be a useful tool in our previous work to explore optical properties for large and medium size molecules [46]. The combination of DFT and TDDFT can provide efficient and reasonably accurate predictions of excited state properties of BODIPY dyes. The results obtained through the above methods show the consistency of the calculated maximum absorption wavelengths with the experimental data (Table S2 in the Supporting Information). All the calculations were performed by Gaussian 16 [47]. The maximum absorption wavelengths predicted from the above theoretical calculations reveal a very good correlation relationship with the corresponding experimental data (shown in Figure S1 in Supporting Information), with $R^2 = 0.9979$. This indicates that DFT/TDDFT/B3LYP is appropriate and reliable for the present study systems.

Molecular descriptors calculation

A specific set of molecular features/descriptors was calculated for all compounds in this work, where descriptors are mathematical representations of chemical information contained in a molecule. The studied BODIPY models were fully optimized at B3LYP/6-311G(d,p) level. The optimized structures were used in further steps by cheminformatics methods to generate descriptors by alvaDesc software [48]. The alvaDesc software generated 5666 molecular descriptors corresponding to 0D-, 1D-, 2D-, and 3D- indexes, including a total of 33 different classes of descriptors comprising constitutional, topological, walk and path counts, connectivity indices, information indices, 2D autocorrelations, edge adjacency indices, Burden eigenvalues, topological charge indices, eigenvalue-based indices, Randic molecular profiles, geometrical descriptors, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, functional groups, atom-centered fragments, charge descriptors, and molecular property descriptors [49]. After filtering constant, missing values, and near-constant descriptors (> 0.95 similarity), about 790 descriptors were generated per each BODIPY chemical structure. The workflow of the ML model is shown in Fig. 1.

Feature selection and explainable machine learning model

Based upon the generated molecular descriptors, the optimal linear correlations between the descriptors (quantitative features) and the λ(nm) are examined. Various machine learning algorithms are used as a foundation for determining the ideal relationship between the structure and the property [30, 50, 51]. They help to discover the features with greatest effects on the property. This greatest combination of characteristics is coupled in a mathematical equation where the physical property might be predicted for new chemical entities based on the same features retrieved from its structure as those included in the machine learning models [23, 31, 52, 53].

In the current work, the feature selection procedures use a genetic algorithm (GA) wrapper with multi-linear regression as the fitting method [32, 54–56] for variable selection and multiple linear regression analysis (MLRA) for ML model generation using the sklearn-genetic package. The GA started with a population of 2500 random variable combinations and 27,000 iterations for evolution with the mutation probability specified at 40%, where a subset of 15 variables was selected. The MLR model was optimized and used for the development of machine learning models [57, 58].

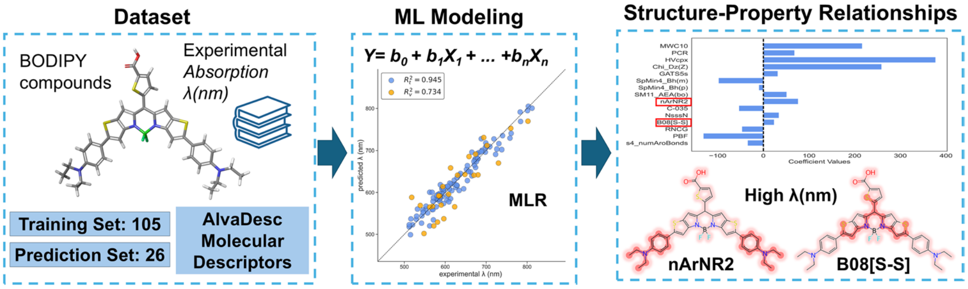


Fig. 1 Workflow for ML model assembled and validation in this study

The quality of models should be assessed to get robust models and hence reliable predictions gathered from them, where the different external and internal validation procedures play a fundamental role at time to check for the robustness and stability of the QSPR models [59–61]. The cross-validation technique "leave-one-out" was used in the internal validation process of the QSPR models obtained from the GA-MLR feature selection. This procedure removes one molecule at each time from the training set and re-runs the selected model against the individual molecules ($Q^2{}_{LOO}$) [55, 62]. Both of the observed response values and the predicted response values calculated by the models are used to obtain the correlation coefficients, R^2 (Eq. 1) and the root mean square errors *RMSE* (Eq. 2), which function as statistical parameters to evaluate the performance of each model. This process is carried out through training, cross-validation data, and the external set $\left(R^2_{\text{training},} Q_{LOO} R^2_{test,}\right)$.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{obs} - y_i^{pred}\right)^2}{\sum_{i=1}^{n}\left(y_i^{obs} - \widetilde{y}^{obs}\right)^2} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i^{obs} - y_i^{pred}\right)^2}{n}} \tag{2}$$

where $y_i{}^{\text{obs}}$ is the λ(nm) from data (observed) value of the property for the i^{th} compound; y_i^{pred} is the predicted value for i^{th} compound. $\widetilde{y}$ is the mean experimental value of the property in the training and validation set, respectively; n is the number of compounds in the training or validation set.

Concurrently, the chemical applicability domain (AD) is calculated for best model by the leverage approach in order to verify the reliability of the predictions [63]. The Williams plot was used to visualize the applicability domain of the QSPR models. The Williams plot of the standardized cross-validated residuals (*RES*) vs. leverage (*Hat* diagonal) values (*HAT*) clearly illustrates both of the response outliers (Y outliers) and structurally influential compounds (X outliers) in a model [64].

Results and discussion

Aromatic [b]-fused BODIPY are promised as near-infrared dyes. Therefore, 131 aromatic [b]-fused BODIPY dyes with diverse structural features were selected as the studied models [36]. All the compounds were optimized, and the compatible absorption spectra were obtained through applied DFT/TDDFT methods. The corresponding experimental data for the maximum absorption wavelength were collected. A QSPR model was therefore developed through finding the relationships between the chemical structure and the maximum absorption wavelength values. The QSPR model is built systematically, starting from a one-variable model to a fifteen-variable model for response value. The regression coefficient of the training and prediction set is the main fitting parameter. Our results reveal that the 15-variable model shows a good combination of R^2 for both training and prediction sets and is the best model of all. Other statistical parameters are listed in Table 1.

The selected QSPR-MLR model with 15 variables predicts the absorption wavelength for the BODIPY dataset according to the following equation (Eq. 3):

$$\begin{aligned}
\lambda(\text{nm}) = {} & 306.02 + 216.523 * \text{MWC10} + 68.255 * \text{PCR} \\
& + 377.08 * \text{HVcpx} + 258.906 * \text{Chi\_Dz(Z)} \\
& + 31.778 * \text{GATS5s} + -98.247 * \text{SpMin4\_Bh(m)} \\
& + -9.962 * \text{SpMin4\_Bh(p)} + 51.199 * \text{SM11\_AEA(bo)} \\
& + 76.248 * \text{nArNR2} + -53.625 * \text{C} - 035 \\
& + 34.163 * \text{NsssN} + 23.698 * \text{B08[S} - \text{S]} \\
& + -47.063 * \text{RNCG} + -131.427 * \text{PBF} \\
& + -34.605 * \text{s4\_numAroBonds}
\end{aligned} \tag{3}$$

The statistical parameters for the models are explained in the "Materials and methods" section and Table 1, and the descriptors related to each model are depicted in Table 2, together with the specific family-type descriptors.

The descriptors involved in the 15-variable model for the λ(nm) are shown in Table 2, including the charge descriptors, 2D autocorrelations, information indices, 2D matrix-based descriptors, 2D atom pairs, atom-type E-state indices, molecular properties, walk and path counts, functional group counts, Burden eigenvalues, Chirality descriptors, atom-centred fragments, and edge adjacency indices.

Moreover, the influence of each descriptor for the 15-variable QSPR model is shown in Fig. 2. The coefficient values are graphically represented in this figure, and their positive or negative effect in the QSPR-MLR model is illustrated. In this model, nine variables have a positive contribution to the property in the following order: HVcpx > Chi_Dz(Z) > MWC10 > nArNR2 > PCR > SM11_AEA(bo) > NsssN > G

Table 1 List of models and statistical parameters in the selection process

Model	Descriptors	$R^2{}_{\text{Train}}$	RMSE$_{tr}$	$Q^2{}_{\text{LOO}}$	RMSE$_{cv}$	*F*-Test	$R^2{}_{\text{Test}}$	*RMSE*$_{\text{Test}}$
Equation 3	MWC10, PCR, HVcpx, Chi_Dz(Z), GATS5s, SpMin4_Bh(m), SpMin4_Bh(p), SM11_AEA(bo), nArNR2, C-035, NsssN, B08[S–S], RNCG, PBF, s4_numAroBonds	0.945	16.442	0.923	15.467	102.3	0.734	35.553

Table 2 Descriptors that were included in the QSPR-MLR models with a short explanation

Descriptor	Descriptor information	Type
MWC10	Molecular walk count of order 10	Walk_and_path_counts
PCR	Ratio of multiple path count over path count	Walk_and_path_counts
HVcpx	Graph vertex complexity index	Information_indices
Chi_Dz(Z)	Randic-like index from Barysz matrix weighted by atomic number	2D_matrix-based_descriptors
GATS5s	Geary autocorrelation of lag 5 weighted by I-state	2D_autocorrelations
SpMin4_Bh(m)	Smallest eigenvalue n. 4 of Burden matrix weighted by mass	Burden_eigenvalues
SpMin4_Bh(p)	Smallest eigenvalue n. 4 of Burden matrix weighted by polarizability	Burden_eigenvalues
SM11_AEA(bo)	Spectral moment of order 11 from augmented edge adjacency mat. weighted by bond order	Edge_adjacency_indices
nArNR2	Number of tertiary amines (aromatic)	Functional_group_counts
C-035	R–CX..X	Atom-centred_fragments
NsssN	Number of atoms of type sssN	Atom-type_E-state_indices
B08[S–S]	Presence/absence of S – S at topological distance 8	2D_Atom_Pairs
RNCG	Relative negative charge	Charge_descriptors
PBF	Plane of best fit	Molecular_properties
s4_numAroBonds	Number of aromatic bonds of the substituent 4	Chirality_descriptors

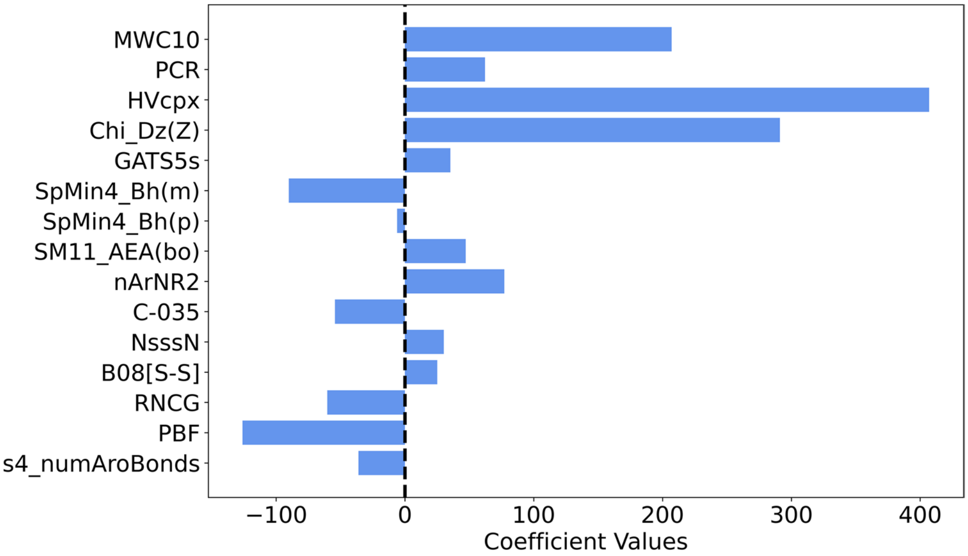


Fig. 2 The magnitude of influence of different descriptors of a 15-variable model on the BODIPY dataset according to Eq. 3

ATS5s > B08[S–S]. On the other hand, six variables show negative contribution towards the property in the following order: SpMin4_Bh(p) > s4_numAroBonds > RNCG > C-035 > SpMin4_Bh(m) > PBF; notice that the ranking of the variables is displayed in the order of the absolute value of the coefficient to reflect the impact of the variables in the model in the opposite direction.

Besides, the predictive ability of the QSPR-MLR model was evaluated, $R^2 = 0.945$ for the training and $R^2 = 0.734$ for the test sets. This proves a good correlation between the predicted and the observed values for our dataset 131 BODIPY compounds as reflected in Fig. 3A.

Williams's plot, shown in Fig. 3B for the ML model, gives a mathematical representation of the chemical space based on the training set. The standardized residuals are plotted on the Y-axis, and the leverage value is plotted on the X-axis, and both values were obtained from the QSPR-MLR model. The outlier response is considered when the data points with standardized residuals are greater than the (-3σ; $+3\sigma$) range. On the other hand, the hat/leverage values account for how much every single compound has the effects on the QSPR-MLR model. As shown in Williams's plot, Fig. 3B, all the data points are located within the 3σ error limit zone (Y-axis) which reveals the absence of

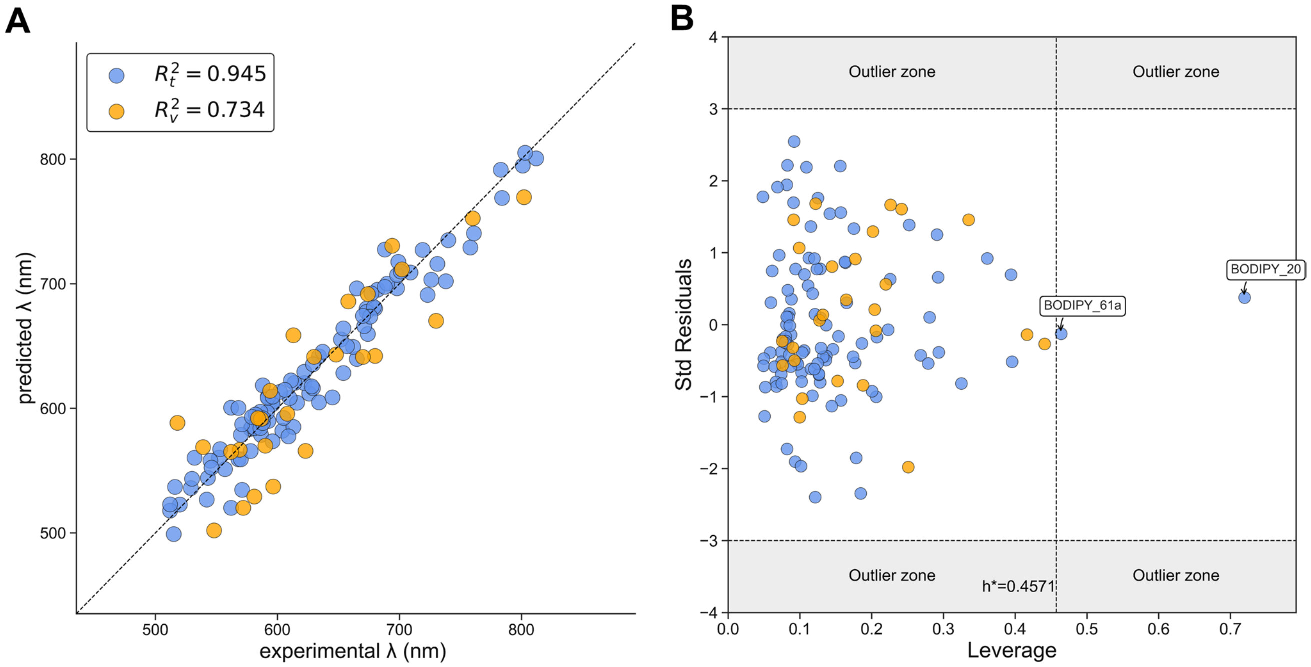


Fig. 3 **A** The correlation plot between the observed and predicted values of photovoltaic dataset. **B** Williams's plot of standardized residual versus leverage of BODIPY compounds dataset. Training set (blue dots), test set (orange dots)

outliers for our maximum absorption wavelength model. It is also noted that all the predicted values are included inside the applicability domain, and this makes the predictions for both test sets reliable for all the compounds. The leverage values suggest that only two compounds for the QSPR model showed values higher than the critical leverage value, which may be caused by some substructural differences in respect to the other compounds in the training set [65]. All the other remaining points have values lower than the critical leverage for both training and prediction set.

A detailed analysis of the descriptors of the 15-variable model was performed to get further understanding of the main features that implicate the tendency of the descriptors regarding the property. A density plot was hence performed for some of the most relevant descriptors in the EML model, where the x-axis are the original descriptors values, and the y-axis corresponds to the experimental maximum absorption wavelength λ(nm) (Fig. 4). The density plot for two of the most relevant descriptors in the model is depicted in Fig. 4. In the case of MWC10 descriptor (Fig. 4A) which is related to the molecular walk count of order 10 [66], the structures with higher values in the descriptor also correspond to the highest values in the wavelength λ(nm). These structures match with the more branched molecular structures, as can be seen in some example compounds in Fig. 4B.

The same phenomenon could be noticed in Fig. 4C for the HVcpx. It is a molecular descriptor associated with the graph vertex complexity index [67]. More vertex for the molecular graph indicates more branches in the chemical structures. High values of the response variables are associated with high HVcpx values. PCR is the ratio of multiple paths count over path count, and it also has a positive impact in the property under study. Both above-mentioned descriptors are molecular descriptors that are related with the branching and size of the molecular structures. Furthermore, the Chi_Dz(Z) descriptor, which is the Randic-like index from Barysz matrix weighted by atomic number, also takes into consideration the size and the atoms in the molecules.

The absorption wavelength data set also demonstrated some interesting findings (shown in Fig. 5) for the nArNR2 and NsssN molecular descriptors that are both related with the number of nitrogens in the molecules. The nArNR2 descriptor is related to the number of tertiary amines bonded to one aromatic group (Ar) and two R groups (R2) where high values ($n = 2$) correspond to high absorption wavelength values. As described in Fig. 5A, the right side of the figure shows two nArNR2 groups of compounds 36b and 36a that are highlighted in red, and λ(nm) equal to 812 and 783, respectively. The left part of Fig. 5A showed some compounds such as 55d and 5i, with nArNR2 highlighted in red and values equal to one but with medium to low values in the absorption wavelength. It suggests that the addition of two nArNR2 group fragment types increases the absorption wavelength significantly, which is in opposite of the case

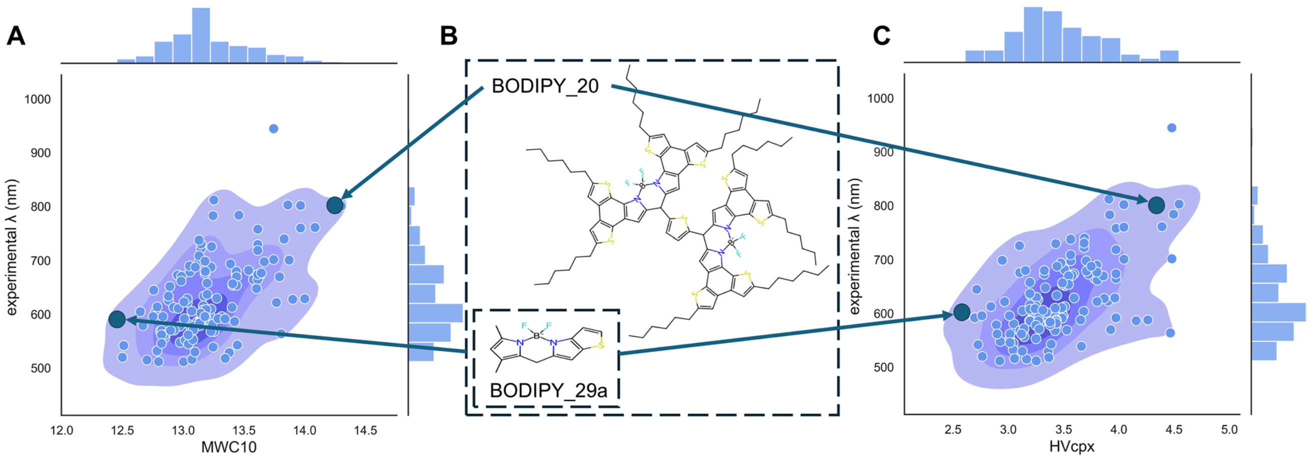


Fig. 4 Density distribution plots. **A** MWC10 molecular descriptor. **B** Representative structures with highest and lowest absorption values. **C** HVcpx molecular descriptor

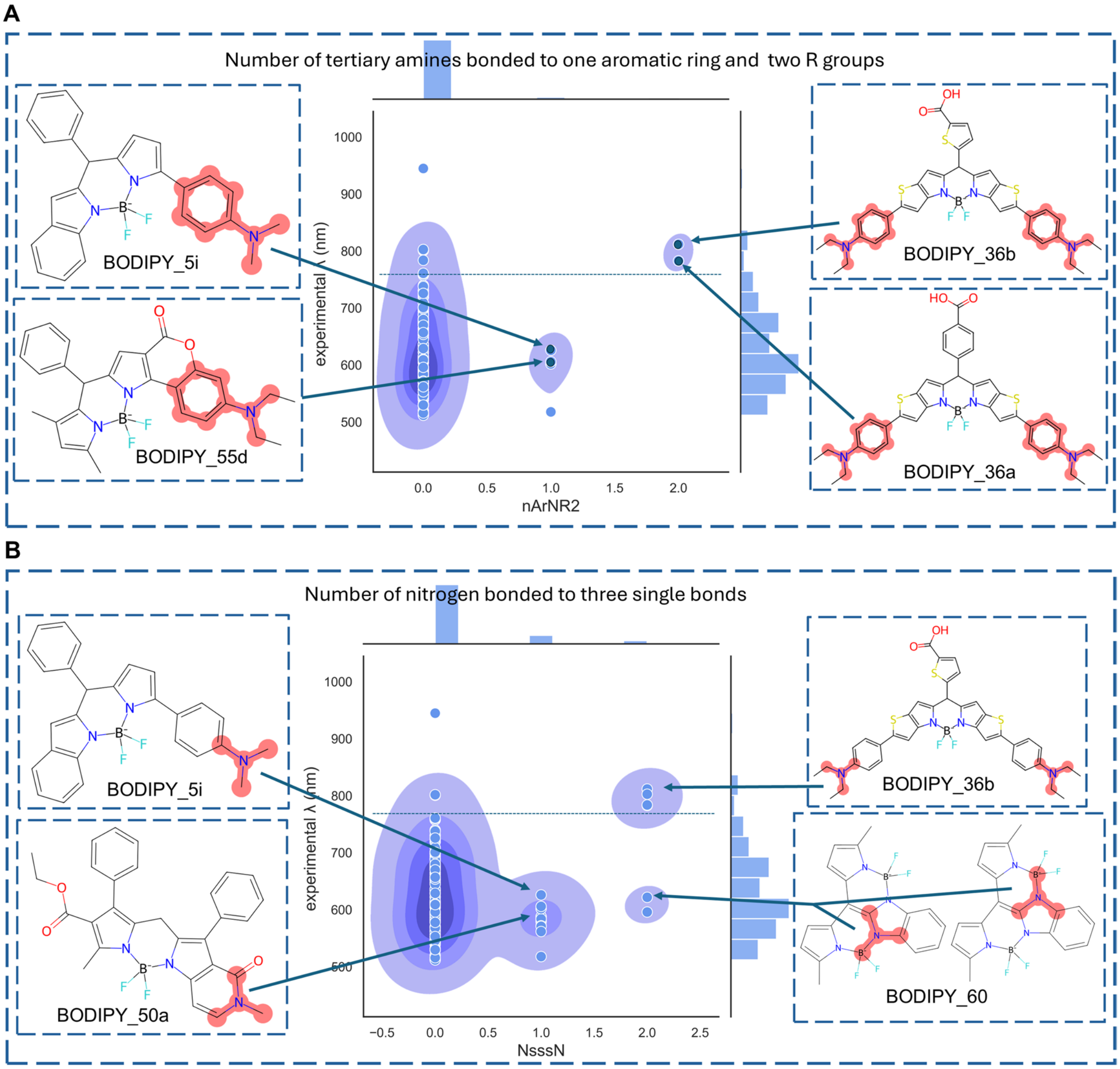


Fig. 5 Density distribution plots. **A** nArNR2 molecular descriptor. **B** NsssN molecular descriptor

of molecule 34a with a relatively large size but without a nARNR2 group showing a medium λ(nm) = 564.

The molecular descriptor NsssN also plays an important role in the same kind of fragments which involve nitrogen atoms (Fig. 5B). It counts for the number of nitrogen atoms connected through 3 single bonds (> N-), where " > " means two single bonds. NsssN has a smaller contribution in the model than the nArNR2 descriptor. The same trend is observed in Fig. 5B. BODIPY-compounds (36b, 61c, 61a, and 36a) show the value of NsssN by 2 and with high absorption wavelength values. It is interesting to notice that for the compounds 9 and 60, where absorption wavelengths fall within the range of 595–625, the > N- fragment is sitting inside the rings of the chemical structure, which can be seen from the structure of BODIPY 60 in the bottom right part of Fig. 5B. This is on the contrary with the previously described 36b where the bonds are not forming parts of ring systems. For the structures with the number of three single bonds connected to nitrogen atoms as one, the absorption wavelengths are in the middle zone of the graph, such as 5i and 50a (depicted in the left part of Fig. 5B). For those molecules with NsssN = 0, the λ(nm) < 761, except for the compounds of 34bR, 58c and 20, but these could be done to other factors/descriptors influencing the property, since this is a multi-factor model accounting for different substructures contributing to the overall value of the absorption wavelengths, as was discussed previously.

In our QSPR model, it is important to mention the B08[S–S] molecular descriptor which has a positive contribution to the absorption wavelength. Although this encoded feature shows a smaller contribution in comparing with the other descriptors in the model, it belongs to the 2D atom pairs family of descriptors that set structural features in the molecules as B08[S–S] descriptor. This descriptor encodes the presence/absence of S–S walk counts at topological distance 8 (separated by eight bonds), where the beginning and ending atoms are both sulfurs. It is encoded as zero where there are no any fragments fitting the pattern and as one when at least one fragment type is found in the structure. Figure 6 depicts the density plot of this descriptor vs the experimental λ(nm) in order to show the trends related with this feature. As shown in the right part of Fig. 6, compounds 34bR and 36b have sulfurs connected at a distance of eight bonds (highlighted in red), and both compounds show high absorption wavelengths.

It is noted that the compounds with B08[S–S] = 1 have higher λ(nm) and are more concentrated in the top part of the density plot 18/24 (76%), while those compounds with B08[S–S] = 0 are more distributed in the lower part of the density plot representing 77 out of 107 compounds (72%), as the case of compounds 55d and 3d. As discussed previously, there are other factors that contribute to the values of the absorption wavelength and also play a role in the shifting of this characteristic.

Finally, a virtual library was designed with the aim to explore the capabilities of the developed model in the findings of new BODIPY compounds with better absorption wavelength values. For this, we used three main cores distributed within three different ranges of absorption values. The first BODIPY compound was 3b ($\lambda = 512$ nm), the second selected core was BODIPY 23c with $\lambda = 662$ nm, and the third one was BODIPY 36b ($\lambda = 821$ nm) as can be seen in Fig. 7. In order to assemble a custom design library, we selected some of the fragments previously discussed

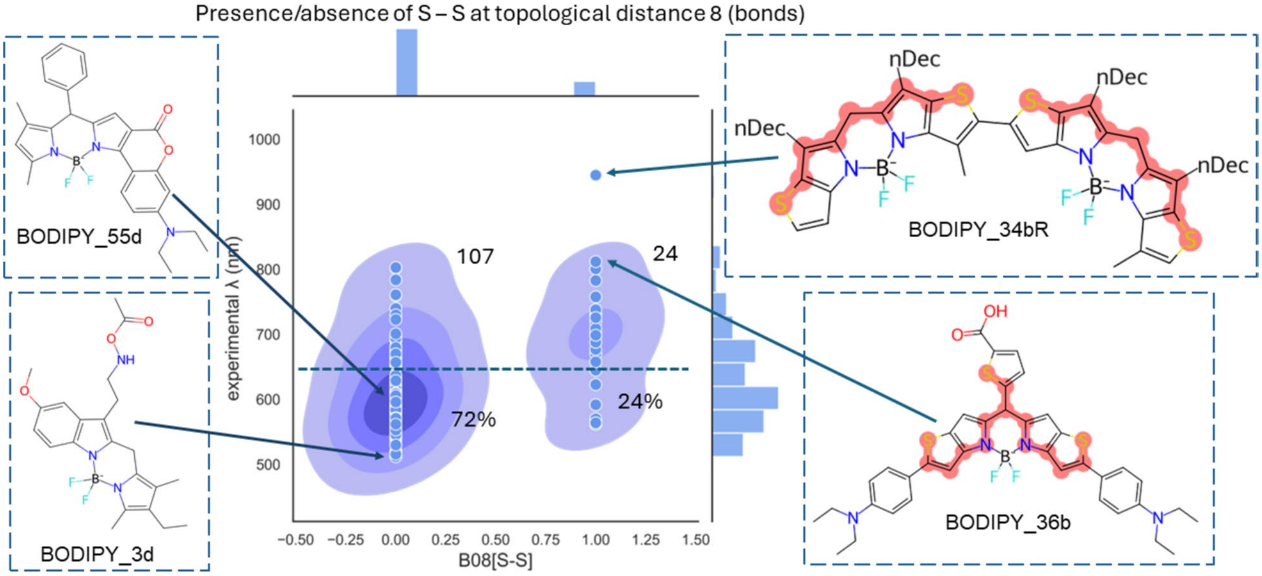


Fig. 6 Density distribution plots and representative chemical structure for B08[S–S] molecular descriptor

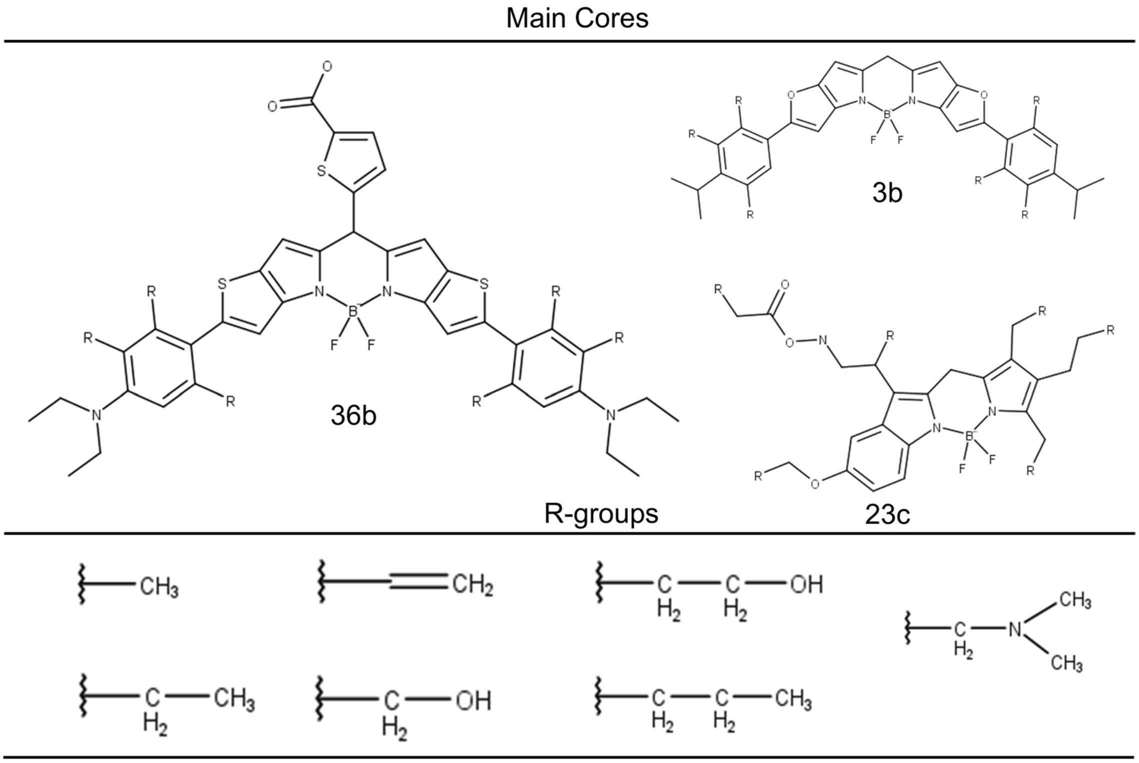


Fig. 7 Main cores and R-groups selected for the virtual library

and responsible for increasing the *λ* values for some of the BODIPY compounds resulting in a total of seven fragments as shown in the bottom part of Fig. 7. Then, the selected fragments were attached to six R-positions of the main cores as shown in the figure. This combination ends up with a total of 117,649 per core with a total of 352,947 BODIPY compounds in the virtual library that were attached in a systematic way using the ChemAxon's MarvinSketch suite [68].

After generating the virtual library, the ML-QSPR model was re-run to generate predictions for this new set of BODIPY compounds. Figure 8 shows the distribution of the three derivative types using a UMAP-based approach where the intensity in each family reflects the absorption wavelength values being the BODIPY 36b derivatives, the ones with the highest absorption values as could be observed from the bars in Fig. 2. Besides, Fig. 9 shows the BODIPY compounds from the virtual library with the highest absorption, all of them belonging to the 36b derivatives.

Conclusions

In this work, we have developed a quantitative structure–property relationship (QSPR) model for predicting the maximum absorption wavelength (λ) of Boron-dipyrromethene (BODIPY) compounds using a combination of molecular descriptors and machine learning techniques to address the challenges associated with hand-crafted design of BODIPY compounds, by leveraging computational methods and datasets of BODIPY structures and their corresponding photophysical properties. Through the systematic optimization of BODIPY compounds using density functional theory (DFT) calculations, we generated a diverse dataset comprising 131 BODIPY compounds with experimentally determined maximum absorption wavelengths. We then employed a wide range of molecular descriptors to represent the structural features of these compounds, capturing information about their size, shape, connectivity, and chemical composition.

Using a genetic algorithm (GA) wrapper with multi-linear regression (MLR) as the fitting method, we identified a subset of 15 molecular descriptors that exhibited the strongest correlations with the maximum absorption wavelength. The developed QSPR-MLR model demonstrated good predictive performance, with adequate coefficients of determination (R^2) of 0.945 and 0.734, for the training and test set, respectively. The model exhibited robustness and reliability, as evidenced by the low root mean square errors (RMSE) and the absence of outliers in both the training and test sets.

Furthermore, our analysis of the selected molecular descriptors provided valuable insights into the structural features influencing the maximum absorption wavelength

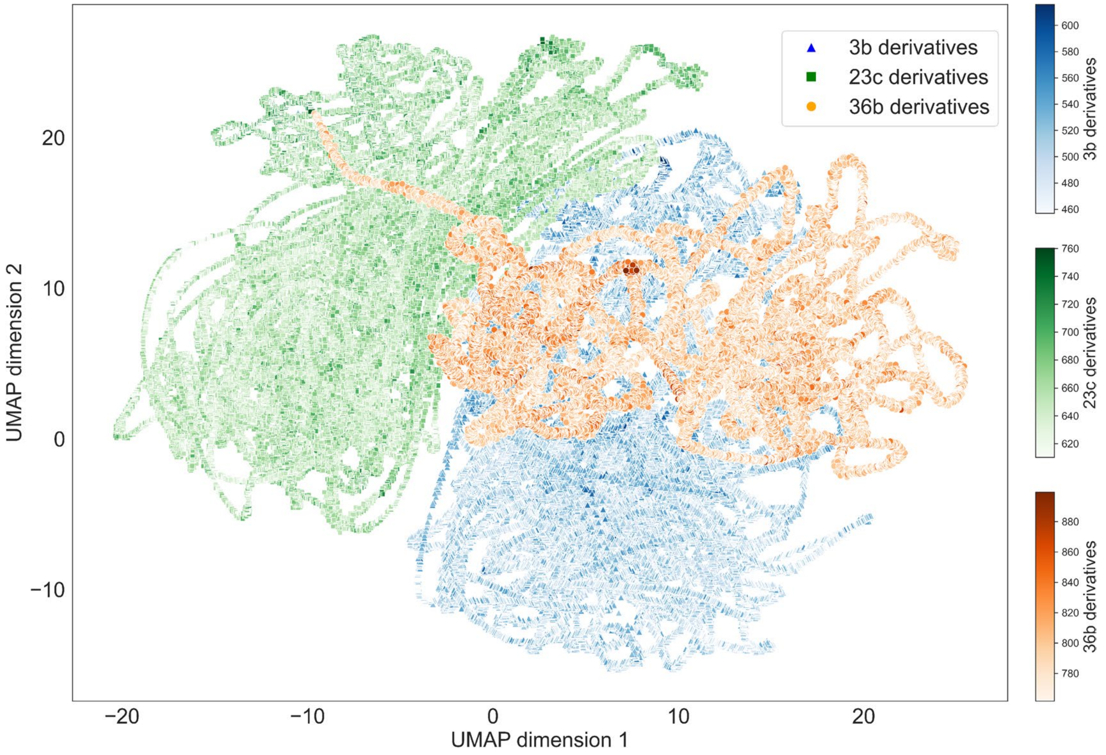
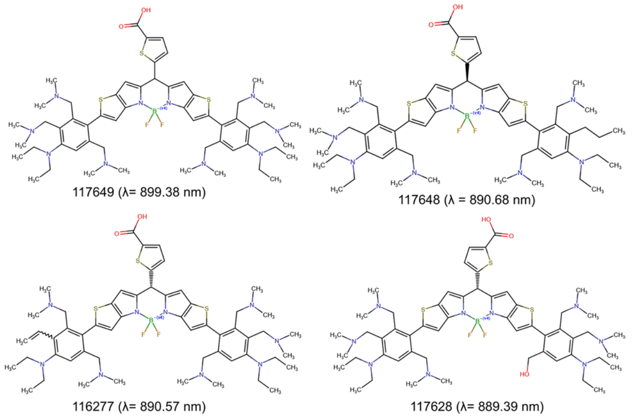


Fig. 8 UMAP distribution for the 3b, 23c, and 36b derivatives


117649 (λ= 899.38 nm)

117648 (λ = 890.68 nm)

116277 (λ= 890.57 nm)

117628 (λ= 889.39 nm)

Fig. 9 New generated BODIPY compounds from the virtual libraries with the highest λ values

of BODIPY compounds. We observed significant contributions from descriptors related to molecular branching, size, and presence of specific functional groups like the presence/absence of S–S fragments separated by eight bonds, and the number of tertiary amines bonded to one aromatic group (Ar) and two R groups (R2), highlighting the importance of these factors in determining photophysical properties.

Overall, our study demonstrates the utility of machine learning approaches in accelerating the design of BODIPY compounds with tailored photophysical property and helping to expedite the discovery of novel BODIPY structures with improved performance for applications in fluorescence imaging, sensing, optoelectronics, and so on. Our findings show the potential of interdisciplinary research between computational chemistry, machine learning, and experimental research to advance the field of molecular design and materials science.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00894-024-06240-4.

Acknowledgements This research was supported by the NSF award (OIA-1757220, EES-2300321, and HRD-2100971) (JSU) and NSF grant no. IIA-1355466 from the North Dakota Established Program to Stimulate Competitive Research (EPSCoR) through the Center for Sustainable Materials Science (CSMS). This work was also supported in part by the Department of Energy (DOE) award (DE-SC0022239) and the National Science Foundation NSF-MRI award OAC-2019077. The authors are grateful for the financial and administrative support provided by the North Dakota State University Department of Coatings and Polymer Materials. The authors acknowledge the North Dakota State University Center for Computationally Assisted Science and Technology (CCAST) for providing computing resources.

Author contribution All authors (G.C., J.W., J.Z., B.R., and J.L.) contributed to the study conception and design. All authors reviewed and agreed to the publication of the manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Loudet A, Burgess K (2007) BODIPY dyes and their derivatives: syntheses and spectroscopic properties. Chem Rev 107(11):4891–4932. https://doi.org/10.1021/cr078381n
2. Treibs A, Kreuzer F-H (1968) Difluorboryl-Komplexe von Di- und Tripyrrylmethenen. Justus Liebigs Ann Chem 718(1):208–223. https://doi.org/10.1002/jlac.19687180119
3. Haugland RP, Spence MTZ, Johnson ID (1996) Handbook of fluorescent probes and research chemicals, 6th edn. Molecular Probes, Eugene
4. Yee MC, Fas SC, Stohlmeyer MM, Wandless TJ, Cimprich KA (2005) A cell-permeable, activity-based probe for protein and lipid kinases. J Biol Chem 280(32):29053–29059. https://doi.org/10.1074/jbc.M504730200
5. Golovkova TA, Kozlov DV, Neckers DC (2005) Synthesis and properties of novel fluorescent switches. J Org Chem 70(14):5545–5549. https://doi.org/10.1021/jo050540k
6. Trieflinger C, Rurack K, Daub J (2005) "Turn ON/OFF your LOV light": Borondipyrromethene–flavin Dyads as biomimetic switches derived from the LOV domain. Angew Chem Int Ed 44(15):2288–2291. https://doi.org/10.1002/anie.200462377
7. Kowada T, Yamaguchi S, Ohe K (2010) Highly fluorescent BODIPY dyes modulated with spirofluorene moieties. Org Lett 12(2):296–299. https://doi.org/10.1021/ol902631d
8. Turfan B, Akkaya EU (2002) Modulation of boradiazaindacene emission by cation-mediated oxidative PET. Org Lett 4(17):2857–2859. https://doi.org/10.1021/ol026245t
9. Gee KR, Rukavishnikov A, Rothe A (2003) New Ca2+ fluoroionophores based on the BODIPY fluorophore. Comb Chem High Throughput Screen 6(4):363–366. https://doi.org/10.2174/138620703106298455
10. Arbeloa TL, Arbeloa FL, Arbeloa IL, García-Moreno I, Costela A, Sastre R, Amat-Guerri F (1999) Correlations between photophysics and lasing properties of dipyrromethene–BF2 dyes in solution. Chem Phys Lett 299(3):315–321. https://doi.org/10.1016/S0009-2614(98)01281-0
11. Glavaš M, Zlatić K, Jadreško D, Ljubić I, Basarić N (2023) Fluorescent pH sensors based on BODIPY structure sensitive in acidic media. Dyes Pigm 220. https://doi.org/10.1016/j.dyepig.2023.111660
12. Li S, Chang X, Kong X, Wang Q, Zhao F, Han J, Liu Y, Wang T (2024) A visible BODIPY-based sensor for 'Naked-Eye' recognition of Ag+ and its application on test paper strips. Spectrochim Acta A Mol Biomol Spectrosc 304. https://doi.org/10.1016/j.saa.2023.123446
13. Kumarasamy K, Devendhiran T, Chien WJ, Lin MC, Ramasamy SK, Yang JJ (2024) Bodipy-based quinoline derivative as a highly Hg2+-selective fluorescent chemosensor and its potential applications. Methods 223:35–44. https://doi.org/10.1016/j.ymeth.2024.01.002
14. Behera KC, Mohanty R, Ravikanth M (2024) An ?-benzithiazolyl 3-pyrrolyl BODIPY probe for ratiometric selective sensing of cyanide ions and bioimaging studies. Phys Chem Chem Phys 26(7):5868–5878. https://doi.org/10.1039/d3cp05230c
15. Parisi C, Pastore A, Stornaiuolo M, Sortino S (2024) A fluorescent probe with an ultra-rapid response to nitric oxide. J Mater Chem B. https://doi.org/10.1039/d4tb00064a
16. Li X, Liu X (2024) A sensitive probe of meso-cyanophenyl substituted BODIPY derivative as fluorescent chemosensor for the detection of multiple heavy metal ions. J Fluoresc. https://doi.org/10.1007/s10895-024-03581-4
17. Rajagopalan R, Shankar SS, Balasubramaniyan N, Sharma GD (2023) Simple and efficient acceptor-donor-acceptor-type non-fullerene acceptors for a BODIPY-thiophene-backboned polymer donor for high-performance indoor photovoltaics. ACS Appl Mater Interfaces 15(10):13405–13414. https://doi.org/10.1021/acsami.2c23048
18. Ceugniet F, Labiod A, Jacquemin D, Heinrich B, Richard F, Lévêque P, Ulrich G, Leclerc N (2023) Non-fused BODIPY-based acceptor molecules for organic photovoltaics. J Mater Chem C 11(31):10492–10501. https://doi.org/10.1039/d3tc02039h
19. Tok M, Say B, Dölek G, Tatar B, Özgür DÖ, Kurukavak ÇK, Kuş M, Dede Y, Çakmak Y (2022) Substitution effects in distyryl BODIPYs for near infrared organic photovoltaics. J Photochem Photobiol A Chem 429. https://doi.org/10.1016/j.jphotochem.2022.113933
20. Aguiar A, Farinhas J, da Silva W, Santos IC, Alcácer L, Brett CMA, Morgado J, Sobral AJFN (2021) New series of BODIPY dyes: synthesis, characterization and applications in photovoltaic cells and light-emitting diodes. Dyes Pigm 193. https://doi.org/10.1016/j.dyepig.2021.109517
21. Wang Y, Miao J, Dou C, Liu J, Wang L (2020) BODIPY bearing alkylthienyl side chains: a new building block to design conjugated polymers with near infrared absorption for organic photovoltaics. Polym Chem 11(36):5750–5756. https://doi.org/10.1039/d0py00868k
22. Feng R, Mori T, Yasuda T, Furuta H, Shimizu S (2023) Panchromatic small-molecule organic solar cells based on a pyrrolopyrrole aza-BODIPY with a small energy loss. Dyes Pigm 210. https://doi.org/10.1016/j.dyepig.2022.111020
23. Zhuravskyi Y, Iduoku K, Erickson ME, Karuth A, Usmanov D, Casanola-Martin G, Sayfiyev MN, Ziyaev DA, Smanova Z, Mikolajczyk A, Rasulev B (2024) Quantitative structure─permittivity relationship study of a series of polymers. ACS Mater Au. https://doi.org/10.1021/acsmaterialsau.3c00079
24. Diéguez-Santana K, Casañola-Martin GM, Green JR, Rasulev B, González-Díaz H (2021) Predicting metabolic reaction networks with Perturbation-Theory Machine Learning (PTML) models. Curr Top Med Chem 21(9):819–827. https://doi.org/10.2174/1568026621666210331161144
25. Diéguez-Santana K, Puris A, Rivera-Borroto OM, Pham-The H, Le-Thi-Thue H, Rasulev B, Casanola GM (2019) Beyond model interpretability using LDA and decision trees for α-amylase and α-glucosidase inhibitor classification studies. Chem Biol Drug Des 94:1414–1421. https://doi.org/10.1111/cbdd.13518
26. Daghighi A, Casanola-Martin GM, Timmerman T, Milenković D, Lučić B, Rasulev B (2022) *In silico* prediction of the toxicity of nitroaromatic compounds: application of ensemble learning QSAR approach. Toxics 10(12):746. https://doi.org/10.3390/toxics10120746
27. Chebotaev PP, Buglak AA, Sheehan A, Filatov MA (2024) Predicting fluorescence to singlet oxygen generation quantum yield ratio for BODIPY dyes using QSPR and machine learning. Phys Chem Chem Phys 26(38):25131–25142. https://doi.org/10.1039/D4CP02471K
28. Buglak AA, Charisiadis A, Sheehan A, Kingsbury CJ, Senge MO, Filatov MA (2021) Quantitative structure-property relationship modelling for the prediction of singlet oxygen generation by heavy-atom-free BODIPY photosensitizers. Chem A Eur J 27(38):9934–9947. https://doi.org/10.1002/chem.202100922
29. Terrones GG, Duan C, Nandy A, Kulik HJ (2023) Low-cost machine learning prediction of excited state properties of iridium-centered phosphors. Chem Sci 14(6):1419–1433. https://doi.org/10.1039/D2SC06150C

30. Chew AK, Sender M, Kaplan Z, Chandrasekaran A, Chief Elk J, Browning AR, Kwak HS, Halls MD, Afzal MAF (2024) Advancing material property prediction: using physics-informed machine learning models for viscosity. J Cheminform 16 (1). https://doi.org/10.1186/s13321-024-00820-5
31. Karuth A, Casanola-Martin GM, Lystrom L, Sun W, Kilin D, Kilina S, Rasulev B (2024) Combined machine learning, computational, and experimental analysis of the iridium(III) complexes with red to near-infrared emission. J Phys Chem Lett 15(2):471–480. https://doi.org/10.1021/acs.jpclett.3c02533
32. Dieguez-Santana K, Pham-The H, Villegas-Aguilar PJ, Le-Thi-Thu H, Castillo-Garit JA, Casañola-Martin GM (2016) Prediction of acute toxicity of phenol derivatives using multiple linear regression approach for Tetrahymena pyriformis contaminant identification in a median-size database. Chemosphere 165:434–441. https://doi.org/10.1016/j.chemosphere.2016.09.041
33. Pham-The H, Casañola-Martin G, Diéguez-Santana K, Nguyen-Hai N, Ngoc NT, Vu-Duc L, Le-Thi-Thu H (2017) Quantitative structure-activity relationship analysis and virtual screening studies for identifying HDAC2 inhibitors from known HDAC bioactive chemical libraries. SAR QSAR Environ Res 28(3):199–220. https://doi.org/10.1080/1062936x.2017.1294198
34. Ahmed L, Rasulev B, Kar S, Krupa P, Mozolewska MA, Leszczynski J (2017) Inhibitors or toxins? Large library target-specific screening of fullerene-based nanoparticles for drug design purpose. Nanoscale 9(29):10263–10276. https://doi.org/10.1039/C7NR00770A
35. Ponce YM, Khan MTH, Martín GMC, Ather A, Sultankhodzhaev MN, Torrens F, Rotondo R, Alvarado YJ (2007) Atom-based 2D quadratic indices in drug discovery of novel tyrosinase inhibitors: results of in silico studies supported by experimental results. QSAR Comb Sci 26(4):469–487. https://doi.org/10.1002/qsar.200610156
36. Wang J, Boens N, Jiao L, Hao E (2020) Aromatic [b]-fused BODIPY dyes as promising near-infrared dyes. Org Biomol Chem 18(22):4135–4156. https://doi.org/10.1039/D0OB00790K
37. Becke AD (1993) Density-functional thermochemistry. III. The role of exact exchange. J Chem Phys 98(7):5648–5652. https://doi.org/10.1063/1.464913
38. Lee C, Yang W, Parr RG (1988) Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. Phys Rev B 37(2):785–789. https://doi.org/10.1103/PhysRevB.37.785
39. Stephens PJ, Devlin FJ, Chabalowski CF, Frisch MJ (1994) Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. J Phys Chem 98(45):11623–11627. https://doi.org/10.1021/j100096a001
40. Vosko SH, Wilk L, Nusair M (1980) Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. Can J Phys 58(8):1200–1211. https://doi.org/10.1139/p80-159
41. McLean AD, Chandler GS (1980) Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, Z=11–18. J Chem Phys 72(10):5639–5648. https://doi.org/10.1063/1.438980
42. Krishnan R, Binkley JS, Seeger R, Pople JA (1980) Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. J Chem Phys 72(1):650–654. https://doi.org/10.1063/1.438955
43. Hehre WJ (1976) Ab initio molecular orbital theory. Acc Chem Res 9(11):399–406. https://doi.org/10.1021/ar50107a003
44. Cossi M, Barone V, Cammi R, Tomasi J (1996) Ab initio study of solvated molecules: a new implementation of the polarizable continuum model. Chem Phys Lett 255(4):327–335. https://doi.org/10.1016/0009-2614(96)00349-1
45. Scalmani G, Frisch MJ, Mennucci B, Tomasi J, Cammi R, Barone V (2006) Geometries and properties of excited states in the gas phase and in solution: theory and application of a time-dependent density functional theory polarizable continuum model. J Chem Phys 124(9):094107. https://doi.org/10.1063/1.2173258
46. Rajapaksha IN, Wang J, Leszczynski J, Scott CN (2023) Investigating the effects of donors and alkyne spacer on the properties of donor-acceptor-donor xanthene-based dyes. Molecules 28 (13). https://doi.org/10.3390/molecules28134929
47. Gaussian 16 Revision C.01, (2016) Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Petersson GA, Nakatsuji H, Li X, Caricato M, Marenich AV, Bloino J, Janesko BG, Gomperts R, Mennucci B, Hratchian HP, Ortiz JV, Izmaylov AF, Sonnenberg JL, Williams, Ding F, Lipparini F, Egidi F, Goings J, Peng B, Petrone A, Henderson T, Ranasinghe D, Zakrzewski VG, Gao J, Rega N, Zheng G, Liang W, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Throssell K, Montgomery Jr. JA, Peralta JE, Ogliaro F, Bearpark MJ, Heyd JJ, Brothers EN, Kudin KN, Staroverov VN, Keith TA, Kobayashi R, Normand J, Raghavachari K, Rendell AP, Burant JC, Iyengar SS, Tomasi J, Cossi M, Millam JM, Klene M, Adamo C, Cammi R, Ochterski JW, Martin RL, Morokuma K, Farkas O, Foresman JB, Fox DJ. Gaussian, Inc, Wallingford, CT
48. Mauri A (2020) alvaDesc: a tool to calculate and analyze molecular descriptors and fingerprints. In: Roy K (ed) Ecotoxicological QSARs. Springer US, New York, pp 801–820. https://doi.org/10.1007/978-1-0716-0150-1_32
49. Todeschini R, Consonni V (2009) Molecular descriptors for chemoinformatics. Methods and Principles in Medicinal Chemistry. Wiley. https://doi.org/10.1002/9783527628766
50. Zhang Z, Chen C, Cao Y, Wen L, He X, Liu Y (2024) Descriptors applicability in machine learning-assisted prediction of thermal decomposition temperatures for energetic materials: insights from model evaluation and outlier analysis. Thermochimica Acta 735. https://doi.org/10.1016/j.tca.2024.179717
51. Guo S, Yu J, Liu X, Wang C, Jiang Q (2019) A predicting model for properties of steel using the industrial big data based on machine learning. Comput Mater Sci 160:95–104. https://doi.org/10.1016/j.commatsci.2018.12.056
52. Liu R, Tang Y, Tian J, Huang J, Zhang C, Wang L, Liu J (2023) QSPR models for sublimation enthalpy of energetic compounds. Chem Eng J 474:145725. https://doi.org/10.1016/j.cej.2023.145725
53. Jung J, Yoon JI, Park HK, Kim JY, Kim HS (2019) An efficient machine learning approach to establish structure-property linkages. Comput Mater Sci 156:17–25. https://doi.org/10.1016/j.commatsci.2018.09.034
54. Devillers J (1996) Genetic algorithms in computer-aided molecular design. Genetic Algorithms in Molecular Modeling. Elsevier. https://doi.org/10.1016/b978-012213810-2/50002-5
55. Sifonte EP, Castro-Smirnov FA, Jimenez AAS, Diez HRG, Martínez FG (2021) Quantum mechanics descriptors in a nano-QSAR model to predict metal oxide nanoparticles toxicity in human keratinous cells. J Nanopart Res 23(8):161. https://doi.org/10.1007/s11051-021-05288-0
56. Hao Y, Sun G, Fan T, Sun X, Liu Y, Zhang N, Zhao L, Zhong R, Peng Y (2019) Prediction on the mutagenicity of nitroaromatic compounds using quantum chemistry descriptors based QSAR and machine learning derived classification methods. Ecotoxicol Environ Saf 186. https://doi.org/10.1016/j.ecoenv.2019.109822
57. Marrero-Ponce Y, Teran JE, Contreras-Torres E, García-Jacas CR, Perez-Castillo Y, Cubillan N, Peréz-Giménez F, Valdés-Martini JR (2020) LEGO-based generalized set of two linear algebraic 3D bio-macro-molecular descriptors: Theory and validation by QSARs. J Theor Biol 485:110039. https://doi.org/10.1016/j.jtbi.2019.110039
58. Dieguez-Santana K, Pham-The H, Rivera-Borroto OM, Puris A, Le-Thi-Thu H, Casanola-Martin GM (2017) A two QSAR way for

antidiabetic agents targeting using α-amylase and α-glucosidase inhibitors: model parameters settings in artificial intelligence techniques. Lett Drug Des Discovery 14(8):862–868. https://doi.org/10.2174/1570180814666161128121142

59. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. Mol Inf 29(6–7):476–488. https://doi.org/10.1002/minf.201000061
60. Golbraikh A, Tropsha A (2002) Beware of q2! J Mol Graph Model 20(4):269–276. https://doi.org/10.1016/s1093-3263(01)00123-1
61. Golbraikh A, Tropsha A (2000) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. Mol Divers 5(4):231–243. https://doi.org/10.1023/A:1021372108686
62. Gramatica P (2014) External evaluation of QSAR models, in addition to cross-validation: verification of predictive capability on totally new chemicals. Mol Inf 33(4):311–314. https://doi.org/10.1002/minf.201400030
63. Gramatica P, Giani E, Papa E (2007) Statistical external validation and consensus modeling: a QSPR case study for Koc prediction. J Mol Graph Model 25(6):755–766. https://doi.org/10.1016/j.jmgm.2006.06.005
64. Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, Jaworska JS, Kahn S, Klopman G, Marchant CA, Myatt G, Nikolova-Jeliazkova N, Patlewicz GY, Perkins R, Roberts DW, Schultz TW, Stanton DT, van de Sandt JJM, Tong W, Veith G, Yang C (2005) Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships: the report and recommendations of ECVAM workshop 521,2. Altern Lab Anim 33(2):155–173. https://doi.org/10.1177/026119290503300209
65. Kowalska D, Sosnowska A, Bulawska N, Stępnik M, Besselink H, Behnisch P, Puzyn T (2023) How the structure of per- and polyfluoroalkyl substances (PFAS) influences their binding potency to the peroxisome proliferator-activated and thyroid hormone receptors—an in silico screening study. Molecules 28 (2). https://doi.org/10.3390/molecules28020479
66. Ruecker G, Ruecker C (1993) Counts of all walks as atomic and molecular descriptors. J Chem Inf Comput Sci 33(5):683–695. https://doi.org/10.1021/ci00015a005
67. Bonchev D, Trinajstić N (1982) Chemical information theory: structural aspects. Int J Quantum Chem 22(S16):463–480. https://doi.org/10.1002/qua.560220845
68. Chemaxon (2016) MarvinView. 16.3.14.0-master-4840 edn. Accessed 26 Apr 2019

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.