RSC Advances

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2025, **15**, 50065

Interpretable machine learning integrated with TD-DFT descriptors and SHAP analysis for predicting the maximum absorption wavelength of azo dyes

Yonghao Fang, †[a] Changqing Cao,†[b] Dong Yin,[c] Gang Luo,[a] Yanmin Cheng[d] and Qing Wang*[a]

Received 5th October 2025
Accepted 9th December 2025

DOI: 10.1039/d5ra07578e

rsc.li/rsc-advances

The maximum absorption wavelength (λ_{max}) represents a key property determining the application performance of azo dyes, and accurate prediction of λ_{max} is of paramount importance for accelerating the rational design of novel dye molecules. Existing prediction models exhibit significant limitations in terms of prediction accuracy and chemical interpretability. In this work, we propose an innovative prediction framework for λ_{max} of azo dyes by integrating Gaussian Process Regression (GPR) with key molecular descriptors derived from time-dependent density functional theory (TD-DFT) calculations. Results indicate that the coefficient of determination (R^2) for leave-one-out cross-validation (LOOCV) was 0.83, and that for the independent test set was 0.74. According to SHAP analysis, the $S_0 \rightarrow S_1$ transition energy exhibits a negative correlation with λ_{max} (maximum absorption wavelength), while the concurrent elevation of HOMO and LUMO energies induces a red-shift in λ_{max}. Notably, the number of sulfur atoms in the *R* substituent shows a positive correlation with λ_{max}. Furthermore, a high-throughput screening strategy was employed to identify 21 azo molecules with relatively large λ_{max} values from 14 376 virtual samples. The predicted λ_{max} of these identified molecules is expected to undergo a red-shift relative to the baseline maximum λ_{max} of 650 nm in the original dataset. This study presents a straightforward approach for the discovery of azo dyes with extended λ_{max}, providing a practical reference for the targeted design of such functional materials.

1. Introduction

Azo dyes constitute a class of organic molecules defined by the presence of one or more azo groups (–N=N–), which link two substituent moieties (denoted as *R* and *R*′, typically aryl or alkyl groups). They are classified into monoazo and polyazo types based on the number of azo groups within their molecular structure.[1] Given their remarkable thermal stability and structural tunability, azo dyes have also achieved significant advancements in various fields such as photothermal agents,[2] liquid crystal materials,[3–5] nonlinear optical materials,[6–9] light-emitting diodes,[10] and dye-sensitized solar cells,[11,12] where they serve as photovoltaic absorbers. However, the limited light-harvesting capability in near-infrared (NIR) regions remains a critical bottleneck for azo dyes, severely restricting their utility in advanced optoelectronic applications. Take Disperse Red 1 (DR1)—a widely studied commercial azo dye commonly used in devices like dye-sensitized solar cells (DSSCs) and nonlinear optical (NLO) materials—as an example: its maximum absorption wavelength peaks at only 479 nm, far below the NIR range required for high-efficiency energy conversion systems. This spectral limitation directly undermines its performance in devices demanding strong NIR responses. Consequently, developing azo-based compounds with both thermal stability and enhanced NIR light absorption remains an urgent challenge in contemporary dye design. The traditional development of new disperse dyes relies on a "synthesis-testing-optimization" cycle, which entails a screening period of 6 to 18 months and faces challenges in precisely regulating color shade.[13] With the progress of computational chemistry, time-dependent density functional theory (TD-DFT) has emerged as an effective tool for predicting dye spectra. For instance, Alshaye *et al.*[14] confirmed through calculations on quinolinone-based dyes that electron-donating groups (EDGs) can induce a red shift. Additionally, Jacquemin *et al.* (2006)[15] employed time-dependent density functional theory (TD-DFT) to model the molecular structures and optical spectra of thioindigo dyes and their derivatives. Muhammad Usman Khan *et al.* (2024)[16] investigated seven novel benzodithiophene (BDT)-based donor molecules (D1–D7) with the core structure of benzo[1,2-*b*:4,5-*b*′]dithiophene *via* density functional theory (DFT) and time-

[a] College of Intelligent Textile and Fabric Electronics, Zhongyuan University of Technology, Zhengzhou, 450007, China. E-mail: 3644@zut.edu.cn

[b] Institute of Chemistry Co., Ltd, Henan Academy of Science, Zhengzhou, 450002, China

[c] Shanghai Qingshi Chemical Co., Ltd, Shanghai, 201615, China

[d] Xinjiang Career Technical College, Yili, 833200, China

† These authors contributed equally to this work.

dependent density functional theory (TD-DFT) methods. To explore the geometric, photovoltaic, and optoelectronic properties of these newly designed sensitizers (D1–D7), the CAM-B3LYP/6-31G(d,p) method was employed for computational analysis. Their findings demonstrated the potential of molecules designed through DFT and TD-DFT calculations in photovoltaic applications. Nora Hamad Al-Shaalan *et al.* (2025)[17] explored donor–acceptor (D–A) fluorescent dyes and nonlinear optical (NLO) phosphate systems based on perfluorobisphenol (PFBP) using DFT calculations. The analysis confirmed strong D–A interactions and charge delocalization, highlighting the potential of these materials in applications such as electro-optical modulation, harmonic generation, and ultrafast optical switching. This further illustrates the utility of DFT calculations in the discovery of nonlinear optical materials.

Currently, the research on time-dependent density functional theory (TD-DFT) is significantly limited by its tendency to encounter high computational cost in the context of large-scale calculations.[18] Recently, the Sarkar group attempt to simulate the interaction between polyester and dyes using molecular dynamics, but the computational cost is excessively high, making it difficult to promote.[19,20] Data-centric methodologies have emerged as the fourth paradigm in materials science, with machine learning (ML) emerging as a transformative design approach celebrated for its efficiency, precision, and rapidity. Z. Shafiq *et al.* (2025)[21] proposed a strategy for designing novel monomeric substances with numerous terminal electron-withdrawing groups, and screened out 30 small molecule acceptors (SMAs) with the highest power conversion efficiency (PCE) values *via* a trained machine learning model. T. Mubashir *et al.* (2024)[22] established a database containing 700 organic semiconductors and employed machine learning (ML) to design a large number of organic molecules, thereby providing an efficient and systematic framework for the design of organic semiconductor (OSC) polymers. A. Mahmood *et al.* (2024)[23] generated an extensive polymer database through an automated method implemented in RDKit, and identified a large number of potential candidate materials for polymer solar cells using machine learning (ML). Machine learning has been employed for predicting the maximum absorption wavelength ($\lambda_{\max}$) of azo dyes. In their 2021 study, Ksenofontov predicted the position of the maximum absorption band of BODIPY derivatives and synthesized 26 external datasets to validate the feasibility of their proposed model.[24] In their 2020 study, Ye and colleagues leveraged aggregation-based descriptors combined with machine learning algorithms to forecast the fluorescence emission maxima of organic fluorescent molecules.[25] In their 2021 study, Ju and colleagues expanded this framework by integrating structural and solvation-related descriptors to model emission wavelengths and quantum yields for over 3000 distinct organic fluorophores.[26] Their approach achieved computational precision comparable to TD-DFT calculations, demonstrating the utility of descriptor-based machine learning in reproducing quantum chemical accuracy for large-scale material property prediction. Furthermore, machine learning coupled with DFT calculations has demonstrated substantial potential in the discovery of organic molecules with tailored excitation energies. Greenman *et al.* (2022)[27] integrated time-dependent density functional theory (TD-DFT) with directed message-passing neural networks (D-MPNN) to predict the molecular absorption peaks in solutions. Nguyen *et al.* (2025)[28] proposed an efficient graph neural network (GNN)-based approach for predicting molecular optical properties, which combines molecular graph representations with molecular fingerprints. This integration enables the model to capture detailed structural and electronic features as well as solvent effects. Although the aforementioned studies have enhanced the interpretability of conventional machine learning models, they still lack an analysis of the impact of model training features on the training outcomes. Deep learning, in particular, has provided substantial support for spectral prediction tasks of dyes with large-scale datasets,[29] but single models lack the physical interpretability of quantum chemical calculation predictions, resulting in suboptimal prediction performance.[30,31] Therefore, it is imperative to develop a calculation-experiment validation framework adapted to actual environments, so as to achieve accurate prediction from molecular structure to applied color.

In this work, a novel strategy integrating computational simulation and experimental validation is proposed.[32–34] We collected the $\lambda_{\max}$ values of a total of 197 azo molecules in different solvents—including those originally available in the laboratory and others retrieved from publicly accessible literature—to construct the dataset. In the experiment, different colors of azo-structured disperse dyes were selected. These samples exhibited a maximum absorption wavelength ($\lambda_{\max}$) range of 380–650 nm, essentially covering the visible light region. This ensures that the computational and experimental samples form a systematic system with sufficient representativeness. To overcome the limitation of high computational cost in time-dependent density functional theory (TD-DFT) calculations, the Gaussian Process Regression (GPR) algorithm in machine learning is introduced to quantify the contribution weights of substituent descriptors to color.[35,36] To address the lack of chemical interpretability in traditional machine learning (ML), we incorporated features such as the S_0 to S_1 transition energy and HOMO energy—obtained *via* time-dependent density functional theory (TD-DFT) calculations—into the training features of the model. When implementing leave-one-out cross-validation (LOOCV), 5-fold, and 10-fold cross-validation protocols, the determination coefficients (R^2) were calculated as 0.837, 0.783, and 0.832, respectively. Corresponding root mean square error (RMSE) values for these validation methods were 18.9, 20.734, and 21.7. These results indicate that the model exhibits good predictive performance. Unlike other machine learning models that primarily target the prediction of maximum absorption wavelength ($\lambda_{\max}$), this study uniquely integrates SHAP (Shapley Additive Explanations) analysis to decode the correlation between training features and corresponding $\lambda_{\max}$ values. This approach enables the targeted design of dye molecules by establishing interpretable structure–property relationships, distinguishing it from conventional predictive models focused solely on numerical estimation. Ultimately, *via* high-throughput computational screening, 21 disperse dye structures with red-shift values exceeding those of the known dataset were identified. These molecules hold promise as nonlinear optical materials.

2. Methods

2.1. GPR predicts substituent-color relationships

Density Functional Theory (DFT) exhibits relatively high accuracy in calculating the maximum ultraviolet absorption wavelength of dye molecules; however, it demands extremely high computational resources. This is especially true when dealing with large molecular structures, which require stronger computing power and longer calculation time. To simplify this process, this study proposes to predict the color of dye molecules based on Gaussian Process Regression (GPR). GPR is a non-parametric Bayesian technique for solving regression problems. Its core idea is to perform probabilistic modeling on the relationship between input data and target values, providing predicted values along with corresponding uncertainty estimates. A Gaussian process is characterized by two feature functions: the mean function $m(x)$ and the covariance function $k(x, x')$.

$$\mathbf{f}(\mathbf{x}) \sim \mathbf{GP}(\mathbf{m}(\mathbf{x}), \mathbf{k}(\mathbf{x}, \mathbf{x}')) \tag{1}$$

In eqn (1), $\mathbf{m}(\mathbf{x})$ denotes the mean function, representing the predicted value of the function $f(x)$, which is generally assumed to have a zero mean. $k(x, x')$ describes the correlation between any two points x and x'. All variables conforming to a Gaussian distribution are characterized as Gaussian process regression. The establishment of GPR involves four steps. The first stage is to establish the prior distribution of the training data, and the Gaussian process assumes that this prior distribution is consistent with the target values of the data.

$$\mathbf{y} \sim \mathbf{N}(\mathbf{m}(\mathbf{x}), \mathbf{K}(\mathbf{X}, \mathbf{X})) \tag{2}$$

In this case, consider a dataset $\{X, y\}$, where $X = \{x_1, x_2, \ldots, x_n\}$ represents the input points and $y = \{y_1, y_2, \ldots, y_n\}$ denotes the associated target values. Both X and y satisfy eqn (2), in which $\mathbf{m}(\mathbf{x})$ is the mean function, generally assumed to be a zero vector, and $\mathbf{K}(\mathbf{X}, \mathbf{X})$ represents the kernel matrix with parameters $k(x_i, x_j)$ describing the covariance between any two points x_i and x_j. The second stage involves observational noise; in practice, the observed values y typically include a noise component:

$$\mathbf{y} = \mathbf{f}(\mathbf{x_i}) + \boldsymbol{\varepsilon} \tag{3}$$

$\varepsilon \sim \mathsf{N}(0, \boldsymbol{\sigma}^2)$ denotes independently and identically distributed Gaussian noise, representing the uncertainty of observations. The third stage involves formulating the joint distribution. For a new test point x^*, the joint distribution of its associated predicted value $\mathbf{f}(\mathbf{x}^*)$ and the training target values y is:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}(\mathbf{x}^*) \end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix} \mathbf{m}(\mathbf{X}) \\ \mathbf{m}(\mathbf{x}^*) \end{bmatrix}, \begin{bmatrix} \mathbf{K}(\mathbf{X}, \mathbf{X}) + \boldsymbol{\sigma}^2\mathbf{I} & \mathbf{K}(\mathbf{X}, \mathbf{x}^*) \\ \mathbf{K}(\mathbf{x}^*, \mathbf{X}) & \mathbf{K}(\mathbf{x}^*, \mathbf{x}^*) \end{bmatrix}\right) \tag{4}$$

$\mathbf{X}$ represents the known points, $\mathbf{x}^*$ denotes the prediction position, $\mathbf{K}(\mathbf{X}, \mathbf{X})$ is the covariance between known points, $\mathbf{K}(\mathbf{X}, \mathbf{x}^*)$ stands for the covariance between known points and the prediction point, and $\mathbf{K}(\mathbf{x}^*, \mathbf{x}^*)$ indicates the covariance between prediction points. The fourth stage involves deriving the conditional distribution from the joint distribution to obtain the expected distribution of the test points:

$$\mathbf{f}(\mathbf{x}^*)|\mathbf{X}, \mathbf{y}, \mathbf{x}^* \sim \mathbf{N}(\boldsymbol{\mu}(\mathbf{x}^*), \boldsymbol{\sigma}^2(\mathbf{x}^*)) \tag{5}$$

where the predicted mean value $\boldsymbol{\mu}(\mathbf{x}^*)$ is calculated by the following formula:

$$\boldsymbol{\mu}(\mathbf{x}^*) = \mathbf{K}(\mathbf{x}^*, \mathbf{X})[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \boldsymbol{\sigma}^2\mathbf{I}]^{-1}\mathbf{y} \tag{6}$$

where the predicted variance $\boldsymbol{\sigma}^2(\mathbf{x}^*)$ is calculated by the following formula:

$$\boldsymbol{\sigma}^2(\mathbf{x}^*) = \mathbf{K}(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{K}(\mathbf{x}^*, \mathbf{X})[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \boldsymbol{\sigma}^2\mathbf{I}]^{-1} \cdot \mathbf{K}(\mathbf{X}, \mathbf{x}^*) \tag{7}$$

In eqn (5) and (6) mentioned above, the expected mean value characterizes the final prediction result, while the predicted variance represents the uncertainty of the prediction.

To establish the model in this study, it is necessary to quantify the substituent effects of dye molecules, construct a structure–activity relationship between substituent characteristics and their maximum absorption wavelength ($\lambda_{\max}$), realize rapid color prediction for substituent combinations, and demonstrate the prediction reliability through confidence intervals. Firstly, the physicochemical descriptors of each substituent need to be extracted: Hammett constants for electronic effects, van der Waals volume, *etc.* Secondly, the overall molecular descriptors are obtained *via* Gaussian calculations, including HOMO/LUMO energy levels and dipole moments. For modeling with Gaussian Process Regression (GPR), the kernel function (Matérn kernel) + noise (WhiteKernel) are set to capture nonlinear relationships and quantify uncertainties. The Matérn kernel served as the main component for modeling smooth nonlinear correlations between input descriptors and the target variable ($\lambda_{\max}$). A Matérn kernel ($\nu = 3/2$) was selected over a squared exponential kernel (RBF) to balance the ability to capture nonlinear relationships and resistance to overfitting. It was initialized with two key hyperparameters: a signal variance of 1.0 (reflecting the overall variability of the target variable, $\lambda_{\max}$) and a length scale of 1.0 (regulating the sensitivity of predictions to changes in input descriptors). These parameters were not fixed; instead, they were optimized *via* maximum likelihood estimation (MLE) to maximize the marginal likelihood of the training data—a critical step in GPR that balances model fit and complexity to avoid overfitting. To account for irreducible noise in the dataset (*e.g.*, experimental measurement errors and unmodeled minor descriptor effects), a WhiteKernel (white noise kernel) was integrated into the model. It was defined by a noise level parameter, initialized to 1×10^{-5} (a small value to prevent initial overestimation of noise), and this parameter was co-optimized alongside the Matérn kernel's signal variance and length scale to ensure accurate noise quantification. Kernel parameter optimization relied on the L-

BFGS optimizer (Limited-memory Broyden–Fletcher–Goldfarb–Shanno), the default solver for scikit-learn's GPR module. This optimizer was chosen for its efficiency in optimizing smooth, gradient-based objective functions. Strict convergence criteria were imposed: a maximum of 1000 iterations to allow sufficient exploration of the parameter space, and a function value tolerance of 1 × 10^{-5} to terminate optimization once updates to the marginal likelihood no longer yielded a significant improvement (defined as a relative change <1 × 10^{-5}). Existing dye molecular structures are selected as the training set, and Leave-One-Out Cross-Validation (LOOCV) cross-validation is adopted. For Leave-One-Out Cross-Validation (LOOCV)—adopted to assess the model's generalization ability given the dataset scale—each iteration excluded one dye molecule as the test sample and trained the GPR model on the remaining samples. Notably, the optimal kernel parameters (determined from the grid search) were fixed during all LOOCV iterations to avoid data leakage and ensure an unbiased evaluation of predictive stability. Ultimately, the SHAP methodology was employed to rank feature importance, with all findings visualized in subsequent sections through dedicated graphs. Python was employed to conduct model training, hyperparameter optimization (through grid search coupled with cross-validation), and interpretability analysis. Comprehensive model validation necessitates the use of diverse evaluation metrics. Significantly, model fitting performance was quantified using three key metrics: root-mean-square error (RMSE), determination coefficient (R^2), and mean relative error (MRE), as defined in eqn (8)–(10). These metrics enabled a multi-faceted assessment of predictive accuracy, encompassing both absolute deviation (RMSE) and relative error (MRE), alongside goodness-of-fit (R^2).

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{y}_{\mathrm{ei}} - \mathbf{y}_{\mathrm{pi}}\right)^2} \tag{8}$$

$$R^2 = 1 - \frac{\sum\limits_{i=1}^{N}\left(\mathbf{y}_{\mathrm{ei}} - \mathbf{y}_{\mathrm{pi}}\right)^2}{\sum\limits_{i=1}^{N}\left(\mathbf{y}_{\mathrm{ei}} - \bar{\mathbf{y}}_{\mathrm{e}}\right)^2} \tag{9}$$

$$\mathrm{MRE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\mathbf{y}_{\mathrm{pi}} - \mathbf{y}_{\mathrm{ei}}}{\mathbf{y}_{\mathrm{ei}}}\right| \times 100\% \tag{10}$$

Herein, y_{ei} and y_{pi} designate the experimental and predicted $\lambda_{\max}$ values for individual azo dye molecules, respectively. The symbols $\bar{\mathbf{y}}_{\mathrm{e}}$ correspond to the mean values of experimental and predicted $\lambda_{\max}$ across the entire dataset, whereas N denotes the total sample size.

2.2. TD-DFT theoretical calculations

To validate the prediction accuracy of the Gaussian process regression (GPR) model, we employed Time-Dependent Density Functional Theory (TD-DFT) to calculate the maximum absorption wavelength ($\lambda_{\max}$) of dye molecules. Initial molecular structures of the dyes were built using Avogadro 1.2.0, an open-source molecular modeling tool, while all electronic structure calculations—including geometric optimization and TD-DFT spectral analysis—were carried out with the open-source ORCA 5.0.3 package. The B3LYP hybrid functional, characterized by its balanced description of electron exchange-correlation and computational efficiency, has become a standard choice for dye molecule ground-state geometry optimization and electronic structure calculation.[37] The 6-31G(d,p) basis set, which includes polarization functions on heavy atoms and hydrogen atoms, can accurately capture the local electronic effects of functional groups (*e.g.*, azo groups, electron-donating/accepting substituents) that dominate the color properties of dyes, while avoiding excessive computational costs.[38] For all computations, we selected the B3LYP hybrid functional paired with the 6-31G(d,p) basis set, which employs a spherical harmonic basis set (the default configuration for this basis set in ORCA 5.0.3). The SMD (Solvation Model Based on Density) solvation model was used to replicate the *N*,*N*-dimethylformamide (DMF)[39] solvent environment. The constructed dye structures were submitted to ORCA 5.0.3 for geometric optimization, where internal coordinates (*z*-matrix) were adopted to enhance efficiency. Strict convergence criteria were employed to ensure reliability: total energy convergence set to 10^{-6} Hartree, electron density convergence at 10^{-4} e Bohr^{-3}, orbital gradient limits of 0.001 Hartree/Bohr (maximum) and 0.0005 Hartree/Bohr (root-mean-square, RMS), and nuclear force thresholds of 0.005 Hartree/Bohr (maximum) and 0.003 Hartree/Bohr (RMS). These settings aimed to secure the most stable conformations of the dye molecules. In terms of integral parameters for both optimization and spectral calculations, ORCA 5.0.3 was configured with a Schwarz screening tolerance of 10^{-8} Hartree for Coulomb integrals and electron density. Single-value decomposition (SVD) settings followed the program's default precision controls, while the exchange-correlation potential integration utilized the "Grid5" option—ORCA's standard high-precision grid—featuring 99 radial and 302 angular grid points per shell, with a grid tolerance of 10^{-5}. Integral quadrature relied on the Lebedev–Laikov scheme and the program's pre-defined grid weights, and no level shifting was applied to the empty orbital diagonal elements of the Fock matrix. Post-optimization, we adjusted the calculation parameters to "%td nstates 10" (ORCA's syntax for defining 10 excited states) to compute UV-visible absorption spectra, maintaining consistency with the integral and SVD settings used during optimization. This workflow yielded the UV-visible absorption spectra of each dye molecule in DMF.[40] All calculations—both optimization and spectral analysis—were validated *via* frequency calculations to confirm the absence of imaginary frequencies, thereby verifying the stability of the results and the reliability of the optimized conformations.

3. Result and discussion

3.1. Dataset construction and collection methodology

This dataset comprises 197 pure azo dye molecular entities, with structural diversity meticulously designed based on the key features governing the optical properties of azo dyes. It encompasses three subcategories classified by the number of azo groups (–N═N–): monazo, diazo, and triazo/polyazo derivatives. These dyes exhibit systematic variations in core structural parameters: conjugated chain length; substituent type and substitution position, with substituents distributed at the *ortho*-, *meta*-, and *para*-positions of the aromatic rings; and aromatic skeletons (phenyl, naphthyl, and heterocyclic cores containing N/O/S atoms such as pyridine or thiophene). The screened dataset was randomly partitioned into training and test subsets at a 4 : 1 ratio. Overall, 160 samples (80%) constituted the training cohort for model development, parameter optimization, and 5-fold cross-validation, while the remaining 37 samples (20%) served as an independent test set for evaluating predictive performance. To further validate the generalizability of the model, nine newly synthesized azo dyes from our laboratory were incorporated into an extended validation cohort. The λ_{max} values of all samples range from 380 to 650 nm, effectively covering the entire visible light spectrum.

3.2. Feature construction and transformation

After the structures of azo dyes were transformed to SMILES format, RDkit-derived molecular objects yielded a total of 5188 structural descriptors.[41] To mitigate the risk of overfitting, feature screening was conducted to enhance the model's predictive performance. This study employed a two-stage feature selection strategy, taking into account the high dimensionality of original features. First, the variable autocorrelation screening method was implemented using a threshold of 0.9 to remove collinear features and mitigate potential redundancy. Subsequently, Maximum Relevance and Minimum Redundancy (MRMR) was integrated with GPR to determine the optimal feature subset for model construction. Model performance was evaluated using the determination coefficient (R^2) and root-mean-square error (RMSE), with these metrics serving as key indicators of predictive accuracy.[42,43] The determination coefficient (R^2) and root-mean-square error (RMSE) for leave-one-out cross-validation (LOOCV) are presented in Fig. 1.[44] In the GPR model, excessively high dimensionality will lead to low computational efficiency, overfitting, and kernel function failure; excessively low dimensionality will result in underfitting due to information loss.[45] To achieve a balance between feature dimensionality and model performance, iterative validation revealed that 12 features yielded optimal R^2 and RMSE metrics. As depicted in Fig. 2, Pearson correlation coefficients below 0.90 for all feature pairs indicate negligible collinearity between selected descriptors. Table 1 summarizes the key molecular descriptors curated through systematic feature engineering, including both quantum chemical parameters derived from TD-DFT calculations and topological indices characterizing molecular conjugation.

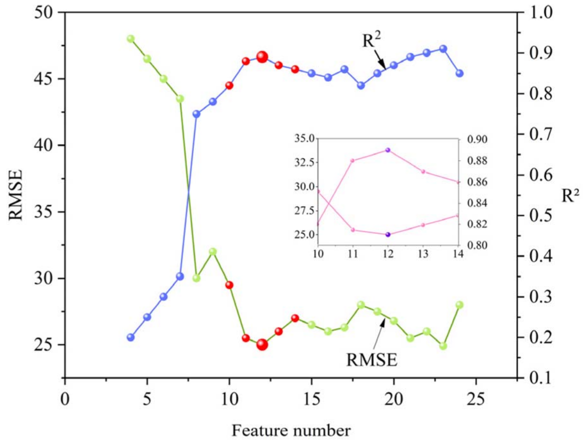

Fig. 1 LOOCV-derived RMSE and R^2 during feature engineering. The violet marker indicates the optimal R^2/RMSE combination, with blue/green lines representing RMSE/R^2 trends.

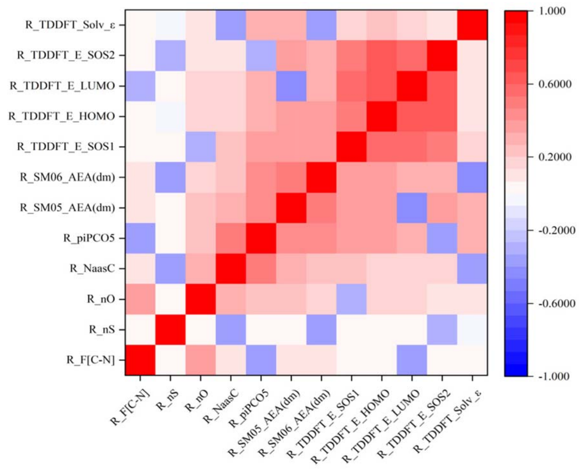

Fig. 2 Pearson correlation matrix depicting inter-feature relationships, with color intensity encoding the magnitude of correlation coefficients.

3.3. Model selection

Selection of a suitable machine learning (ML) algorithm facilitates the accurate prediction of the λ_{max} of azo dyes. In this work, five commonly used ML algorithms were employed, including Gaussian Process Regression (GPR), Extreme Gradient Boosting (XGBoost), Random Forest Regression (RFR), Decision Tree Regression (DTR), and Support Vector Regression with radial basis function kernel (SVR-RBF). Table 2 presents various evaluation metrics of the regression models based on different algorithms in the Leave-One-Out Cross-Validation (LOOCV) framework. Our results demonstrate that the GPR algorithm exhibits the highest coefficient of determination (R^2) value as well as the lowest root mean square error (RMSE) and relative error, enabling it to achieve superior predictive performance for the λ_{max} of azo dyes.

3.4. Tuning of hyper-parameters

Hyperparameter tuning constitutes a key procedure in machine learning. With the aim of further boosting the model's efficiency and performance, a grid search approach was utilized to identify the optimal hyperparameters for GPR. This method centered on hyperparameters such as kernel function type and

Table 1 Selected critical descriptors from two-stage feature engineering

Feature	Description
R_F[C–N]	Frequency of C–N within topological frameworks
R_nS	Sulfur atom count
R_nO	Oxygen atoms count
R_NaasC	Count of aasC-type atoms[46,47]
R_piPC05	5th-order molecular multiple-path count
R_SM05_AEA (dm)	5th-order spectral moment derived from the dipole moment-weighted augmented edge adjacency matrix[48]
R_SM06_AEA (dm)	6th-order spectral moment derived from the dipole moment-weighted augmented edge adjacency matrix
R_TDDFT_E_S0S1	TD–DFT calculated $S_0 \rightarrow S_1$ transition energy (eV)
R_TDDFT_E_HOMO	TD–DFT calculated HOMO energy (eV)
R_TDDFT_E_LUMO	TD–DFT calculated LUMO energy (eV)
R_TDDFT_E_S0S2	TD–DFT calculated $S_0 \rightarrow S_2$ transition energy (eV, for multi – state coupling)
R_TDDFT_Solv_ε	Solvent dielectric constant used in TD–DFT (for PCM model)

Table 2 Evaluation metrics of regression models with different algorithms in LOOCV[a]

Algorithm model	R^2	RMSE	MRE
GPR	0.837	18.9	0.043
XGBoost	0.823	23.42	0.047
RFR	0.809	25.28	0.051
SVR	0.638	39.81	0.099
DTR	0.773	31.64	0.052

[a] Note: the hyper-parameters of all algorithms were adopted as default values prior to parameter optimization.

noise standard deviation ratio. Following optimization, the RBF kernel was selected as the kernel function, characterized by a length scale of 1, an amplitude of 0.5, and a noise standard deviation set at 0.03. The RMSE derived from LOOCV served as the criterion throughout the hyperparameter optimization process.

3.5. Evaluation of model performance

In the present study, RMSE, R^2, and MRE were employed as metrics to objectively assess model performance; specifically, superior performance is indicated by smaller RMSE and MRE values coupled with a larger R^2.[49] Given the small dataset size, LOOCV results offer a more objective evaluation of the model, while 5-fold and 10-fold CV results act as supplementary assessments. Fig. 3 depicts the experimental $\lambda_{\max}$ values and GPR-predicted results under LOOCV, test set, 5-fold CV, and 10-fold CV scenarios. The dataset includes model performance metrics across various cross-validation protocols to assess prediction robustness.

In a study by Gopala Krishna and co-workers, QSPR models were developed to model $\lambda_{\max}$ for triphenylamine, phenothiazine, and indoline dye systems. The approach yielded test set R^2 values of 0.606, 0.624, and 0.759 for the respective dye classes, leveraging molecular descriptors to establish structure–activity relationships. In contrast, the current research leverages a GPR model for azo dye $\lambda_{\max}$ prediction, yielding a test set R^2 of 0.736, which stands as a favorable outcome. As depicted in Fig. 3(a), the LOOCV set achieves $R^2 = 0.837$, RMSE = 18.9, and MRE = 0.043, presenting a relatively concentrated distribution of prediction errors where the maximum absolute wavelength difference is less than 15 nm. Analysis of the molecular structures in this set reveals that most molecules possess canonical structural motifs of azo dyes, and such structures occupy a high proportion in the training dataset. As depicted in Fig. 3(b), the test set achieves $R^2 = 0.736$, RMSE = 31.518, and MRE = 0.051, the test set contains outlier data with substantial deviations, where wavelength differences exceed 35 nm. Analysis of the deviant dye molecular structures reveals the presence of heterocycles in all such molecules. As depicted in Fig. 3(c), the 5-fold CV set achieves $R^2 = 0.783$, RMSE = 20.734, and MRE = 0.043, exhibiting an error distribution between that of LOOCV and the test set where the average absolute wavelength difference is 15.2 nm. Analysis of the molecular structures in this set indicates that the structural differences between the training and validation subsets are relatively small, resulting in good model stability depicted in Fig. 3(d), the 10-fold CV set achieves $R^2 = 0.832$, RMSE = 21.7, and MRE = 0.049, having a relatively compact error distribution where the maximum absolute wavelength difference is 28 nm. Analysis of the molecular structures in this set shows that the model's generalization ability is close to that of LOOCV, further verifying the reliability of the GPR model with a small dataset. Given the low prevalence of molecules with analogous molecular structures in the training dataset, accurately predicting $\lambda_{\max}$ for these test set compounds presents a substantial challenge. In contrast, wavelength predictions for azo dyes with canonical structural motifs align closely with the expected accuracy benchmarks, as corroborated by the model's performance metric.

3.6. Model reliability analysis

To assess the reliability of the model predictions, uncertainty quantification analysis was performed by leveraging the probabilistic output characteristic of Gaussian Process Regression (GPR). For samples in both the training and test sets, the 95% confidence intervals (corresponding to predicted values ±2 times the standard deviation) and prediction standard deviations were calculated. The results demonstrate that the average prediction standard deviation of the training set is 11.2 nm, with 95% confidence intervals generally less than ±25 nm. For

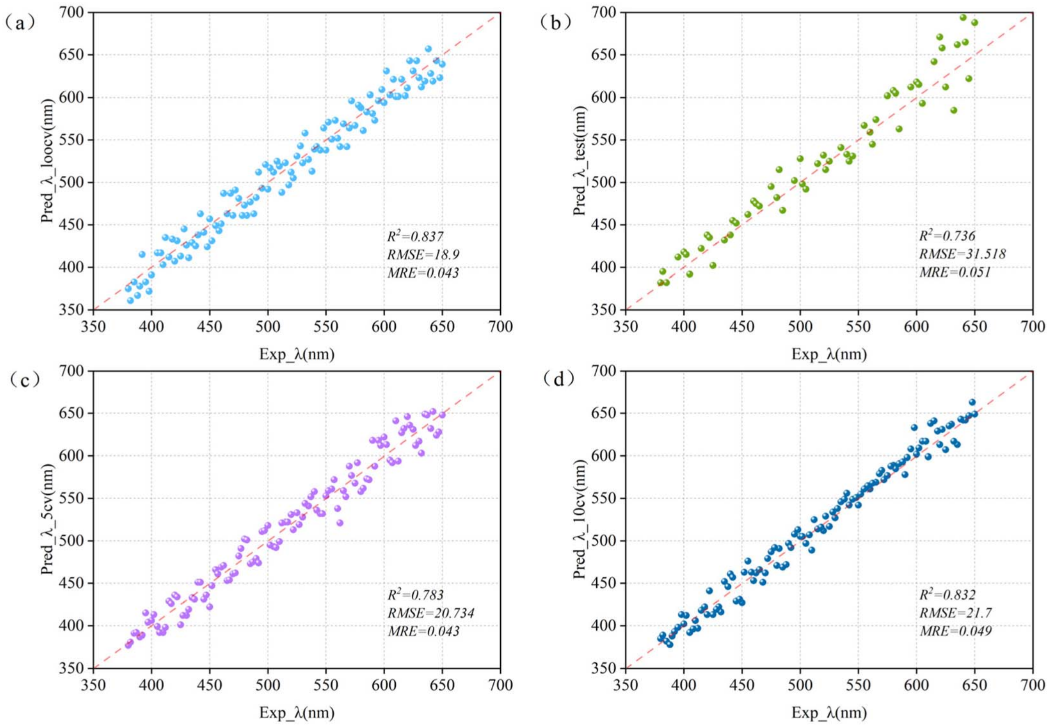

Fig. 3 Comparison of experimental λ_{max} values against GPR-predicted results in (a) LOOCV, (b) independent test set, (c) 5-fold CV, and (d) 10-fold CV frameworks.

the test set, the average prediction standard deviation is 16 nm, where azo dye molecules containing heterocyclic skeletons (*e.g.*, pyridine derivatives) exhibit a significant increase in prediction standard deviation (average of 20.7 nm), which is directly associated with the insufficient representation of such structures in the training set. Statistical methods were employed to verify the significance and stability of the model performance. Bootstrap resampling (1000 iterations) was employed to calculate the 95% confidence intervals (CIs) of the coefficient of determination (R^2): $R^2 = 0.837$ for the training set (95% CI [0.80, 0.87]) and $R^2 = 0.736$ for the test set (95% CI [0.68, 0.79]), indicating the statistical stability of the model performance. A paired *t*-test was conducted between the machine learning predictions and TD-DFT calculated values on the test set, yielding $p = 0.25$ ($p > 0.05$), with no significant difference observed between the two.

Collectively, the uncertainty quantification and statistical validation results confirm that the GPR model constructed in this study not only achieves practical-level numerical prediction accuracy but also possesses statistical significance and categorical generalization. It provides reliable theoretical guidance for the prediction of spectral properties of azo dyes.

3.7. Model explanation

The SHAP (SHapley Additive exPlanations) framework was utilized to interpret the descriptors and target variables within the GPR model.[50–52] As an additive feature attribution approach grounded in cooperative game theory, SHAP allows for the quantification of feature significance and delivers both local and global model interpretations. In this model, each feature acts as a "contributor," where the SHAP value associated with a feature in a given sample quantifies its marginal contribution to the target variable for that specific observation. Visualizing SHAP values for all features across the dataset enables a clear demonstration of each descriptor's relative importance to the model's predictive logic, highlighting how individual features influence the target variable at both the global and local levels.

In the visualization results of Fig. 4, features are vertically ordered in descending order of their impact on the prediction target. Each data point corresponds to a sample in the model

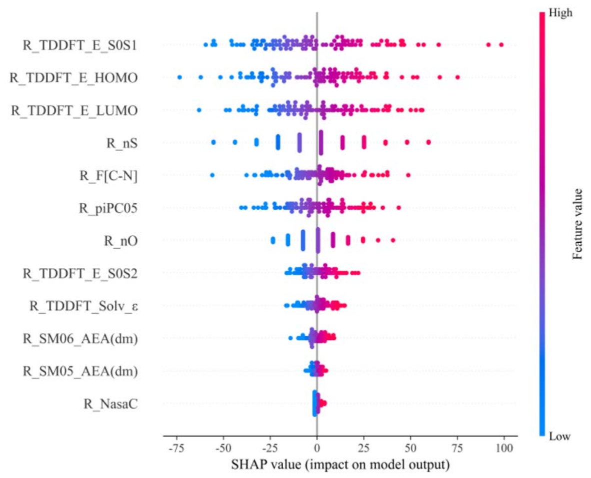

Fig. 4 SHAP value (impact on model output).

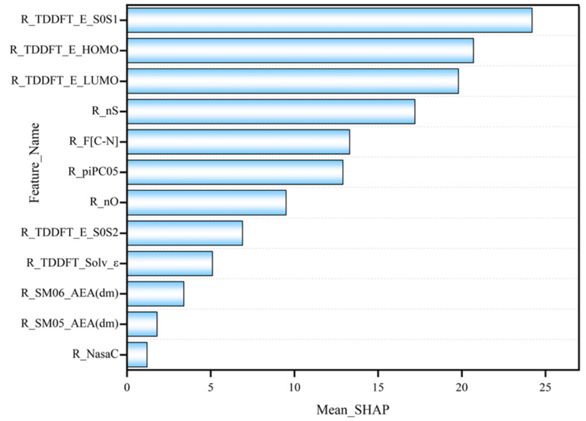

Fig. 5 |SHAP value|(average impact on model output magnitude).

training set, where the color intensity indicates the magnitude of the feature value (with red representing higher values and blue representing lower values). The horizontal axis represents the distribution range of SHAP values, ranging from negative to positive values from left to right. Fig. 5 depicts the comprehensive correlation landscape between predictor variables and the target property, with feature importance quantified *via* the mean absolute SHAP values. This metric serves as a proxy for each feature's relative influence: higher values denote stronger associations with variations in the target variable.

The subsequent analysis focuses on the most impactful descriptors identified from this ranking. Below, we dissect the top-performing features, elucidating their structural or physicochemical meanings and their mechanistic roles in governing the target property.

3.7.1. R_TDDFT_E_S0S1. R_TDDFT_E_S0S1 denotes the S_0 to S_1 transition energy calculated *via* time-dependent density functional theory (TD-DFT) (eV). As revealed by the SHAP univariate plot, among all molecular structural features (Fig. 6(a)), the S_0 to S_1 transition energy exerts the most prominent influence on λ_{max} and exhibits a negative correlation with λ_{max} This indicates that reducing the S_0 to S_1 transition energy is conducive to the red-shift of λ_{max} in azo molecules.

3.7.2. R_TDDFT_E_HOMO and R_TDDFT_E_LUMO. R_TDDFT_E_HOMO represents the energy of the highest occupied molecular orbital (HOMO) calculated calculated *via* density functional theory (DFT) (eV), while R_TDDFT_E_LUMO denotes the energy of the lowest unoccupied molecular orbital (LUMO) also obtained through DFT calculations (eV). The SHAP values of these two features are shown in Fig. 6(b) and (c). Both features exhibit a positive correlation with λ_{max}, with quantitative trends: for R_TDDFT_E_HOMO, every 0.5 eV increase in energy correlates with a 15 nm red-shift in λ_{max} (95% CI: [12, 18] nm); for R_TDDFT_E_LUMO, a 0.5 eV increase correlates with a 12 nm red-shift (95% CI: [9, 15] nm) They synergistically influence electron transition characteristics by regulating the "orbital energy gap (HOMO–LUMO)". Specifically, a higher HOMO energy implies a lower binding energy of electrons in the highest occupied orbital, making it easier for electrons to be excited from this orbital. In contrast, a higher LUMO energy elevates the energy level of the lowest unoccupied orbital, bringing it closer to that of the HOMO. Together, these two effects ultimately narrow the

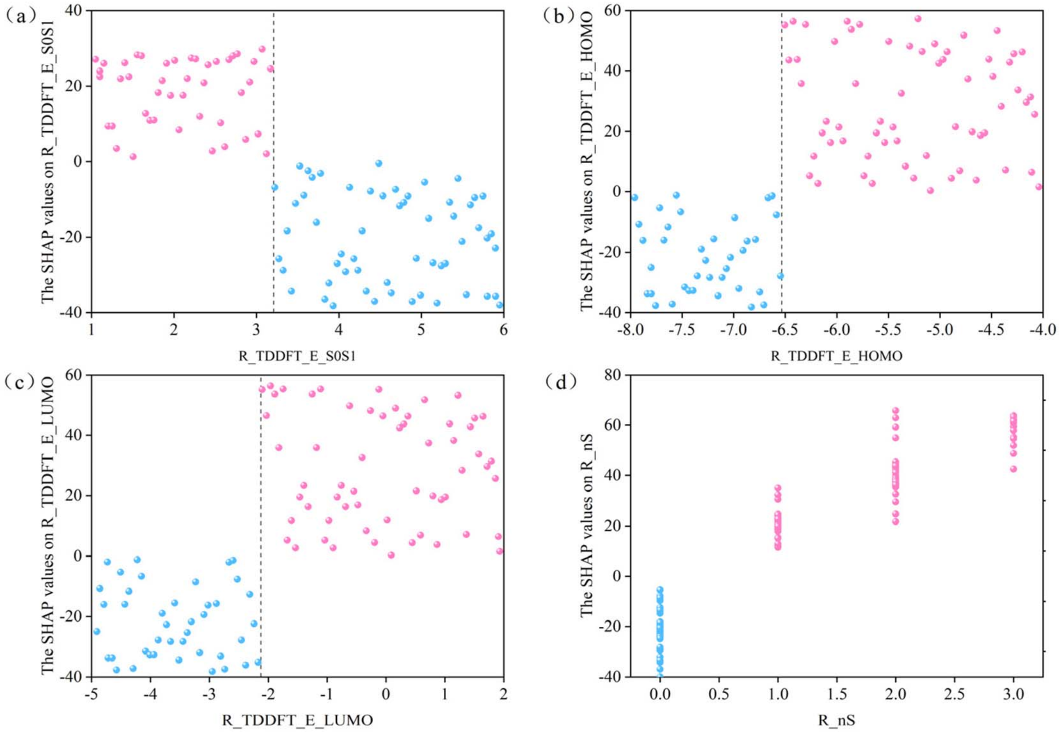

Fig. 6 SHAP values for descriptors exerting substantial influence on the model, including: (a) R_TDDFT_E_S0S1; (b) R_TDDFT_E_HOMO; (c) R_TDDFT_E_LUMO; and (d) R_nS.

Table 3 Maximum absorption wavelength under different method[a]

Method	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
ML	651.04	651.73	651.92	652.1	652.59	652.67	652.83	653.16	653.39	653.52
TD-DFT	702.35	642.04	661.28	668.56	489.65	611.43	671.74	591.62	622.97	637.25

[a] Note: the unit of maximum absorption wavelength (λ_{max}) is nm.

electron transition energy gap between the ground state and the excited state. Additionally, SHAP interaction analysis revealed a nonlinear synergistic effect between R_TDDFT_E_HOMO and R_TDDFT_E_LUMO: when HOMO energy > −5.0 eV and LUMO energy > −1.0 eV, the combined SHAP contribution increased exponentially, far exceeding the additive effect of individual descriptors. This nonlinearity underscores the complex orbital-interaction mechanisms in azo dye electron transitions, consistent with theoretical frameworks in organic photochemistry. In the dataset, among molecules that simultaneously satisfy "HOMO energy > −5.0 eV" and "LUMO energy > −1.0 eV", over 80% of the samples exhibit a λ_{max} exceeding 500 nm. In sharp contrast, for molecules with "HOMO energy < −5.5 eV" and "LUMO energy < −1.5 eV", only approximately 10% reach this wavelength range. This result confirms that the synergistic elevation of HOMO and LUMO energies can effectively narrow the orbital energy gap, reduce the electron transition energy, and thereby shift the λ_{max} of azo dye molecules toward longer wavelengths (*i.e.*, induce a red-shift).

3.7.3. R-nS. In this context, the number of sulfur atoms in the R-substituent (R-nS) exerts the second most substantial influence on λ_{max}. The SHAP univariate analysis (Fig. 6(d)) highlights that among all molecular descriptors, sulfur atom stoichiometry exerts the most substantial influence on the target variable, exhibiting a pronounced positive correlation: each additional sulfur atom correlates with a 20 nm red-shift in λ_{max} (95% confidence interval: [15, 25] nm). Specifically, 72 compounds in the dataset harbor sulfur atoms, of which 51 (71% of the total) display λ_{max} values exceeding 500 nm. Although molecules lacking sulfur atoms can occasionally feature large λ_{max} values, statistical analysis revealed this scenario occurs in only 11.2% of cases. Moreover, SHAP interaction analysis between R-nS and R_TDDFT_E_HOMO revealed a nonlinear interaction: when R-nS ≥ 2 and HOMO energy > −5.0 eV, the combined SHAP contribution exhibited a superlinear increase, demonstrating that sulfur atom incorporation and HOMO energy elevation synergistically amplify the red-shift effect beyond linear expectations. These findings align with experimental studies on sulfur-containing azo dyes, where sulfur-induced conjugation and orbital modulation were identified as key red-shift mechanisms. Structural modification studies indicate that introducing one or two sulfur atoms into azo molecular frameworks facilitates a significant red-shift in λ_{max}, consistent with the model's interpretive outcomes.

3.8. Model implementation and interpretive analysis

High-throughput screening is one of the key approaches for implementing machine learning (ML) models. In this study, high-throughput screening was performed on the collected molecular groups of azo dyes, yielding 14 376 virtual molecules. Notably, 21 of these virtual samples showed a red-shift in their predicted λ_{max} (maximum absorption wavelength) compared to the baseline maximum of 650 nm in the dataset. In previous analogous studies, time-dependent density functional theory (TD-DFT) has been employed to calculate the spectral properties of azo dyes.[53] The top ten virtual molecular structures were selected for structural optimization; subsequently, their absorption spectra were calculated using time-dependent density functional theory (TD-DFT) at the B3LYP/6-31G(d) level As shown in Table 3, the obtained results were compared with those derived from TD-DFT calculations and ML predictions. These findings indicate that predicting the λ_{max} of dye molecules *via* the Gaussian process regression (GPR) algorithm is practically valuable for experimental researchers, and validate the model is utility in guiding spectral property estimation. A critical analysis of the λ_{max} values between ML predictions and TD-DFT calculations in Table 3 reveals that the predicted results are in good agreement with the calculated results for most molecules. However, one specific azo dye molecule exhibits a significant deviation exceeding 150 nm. This substantial deviation is attributed to the molecule's unique heterocyclic structure with a highly extended conjugated system—such a structure has an extremely low occurrence frequency in the original training dataset. The prediction of the GPR model relies on structure–property correlation patterns learned from the training data. When confronted with such novel structures beyond the training distribution, the model suffers from an inherent limitation in extrapolation capability, leading to overestimation of the red-shift effect and ultimately resulting in a significant deviation.

Furthermore, to fully strengthen the conclusion's influence, future work will involve the synthesis of representative candidate molecules and experimental measurement of their absorption spectra. Using machine learning (ML) techniques to predict the λ_{max} of novel azo molecules is a feasible approach. This strategy enables prescreening of molecules with extended λ_{max}, thereby accelerating the material discovery process and minimizing costs associated with TD-DFT calculations and molecular synthesis. Azo dyes are a prominent subclass of nonlinear optical (NLO) materials. High-quality NLO materials must meet the criteria of red-shifted maximum absorption, larger dipole moments, and thermal stability. Absorption in the low-energy spectral region (red region) is particularly beneficial for NLO materials, as it helps enhance their hyperpolarizability.

4. Conclusions

In this study, a machine learning (ML) model for predicting the maximum absorption wavelength (λ_{max}) of azo dyes was constructed based on Gaussian Process Regression (GPR) combined with Time-Dependent Density Functional Theory (TD-DFT) calculations. This model takes the molecular orbital energies obtained from TD-DFT calculations and the polarity parameters of the environment where the dye is located as input features of the GPR model, which effectively enhances the chemical interpretability of the model. Study reveals that the S_0 to S_1 transition energy exerts the most significant influence on the maximum absorption wavelength (λ_{max}), exhibiting a negative correlation. The synergistic elevation of HOMO and LUMO orbital energies effectively induces a red-shift in λ_{max} of azo dye molecules, while the number of sulfur atoms in the functional group makes a positive contribution to (λ_{max}). Through high-throughput screening of 14 376 virtual azo dye molecules generated *via* structural derivation, we identified 21 compounds with predicted λ_{max} red-shifted beyond the dataset's baseline maximum. As a proof-of-concept, the top 10 candidates were validated *via* TD-DFT calculations, and the results demonstrated good agreement between ML-predicted and TD-DFT-calculated λ_{max} for 8 molecules These findings highlight the potential of the proposed ML-TDDFT framework to aid in the screening of azo dye candidates and provide a proof-of-concept for interpretable integrated ML-TDDFT modeling in optical property prediction. It should be noted that the current study has limitations, including the modest size of the dataset and the lack of experimental validation for the identified virtual candidates. Therefore, the framework requires further verification through experimental synthesis and characterization of the predicted compounds, as well as optimization with larger and more diverse datasets to improve its robustness and generalizability for practical materials discovery applications.

Author contributions

Yonghao Fang: formal analysis, writing – original draft. Changqing Cao: writing – original draft, funding acquisition. Dong Yin: writing – review & editing, resources. Gang Luo: writing – review & editing. Yanmin Cheng: writing – review & editing. Qing Wang: writing – review & editing, funding acquisition, resources.

Conflicts of interest

The authors declare no conflicts of interest.

Data availability

All data presented in this study are available in the main article and its supplementary information (SI), with detailed datasets provided in the accompanying SI. Supplementary information is available. See DOI: https://doi.org/10.1039/d5ra07578e.

Acknowledgements

This research was funded by the Central Guidance on Local Science and Technology Development Fund of Henan Province (Z20241471138), Institute-level project of the Henan Academy of Sciences (20251003003, 230503005)

Notes and references

1 K.-T. Chung, *J. Environ. Sci. Health, Part C*, 2016, **34**, 233–261.

2 S. Fuse, T. Oishi, K. Matsumura, Y. Hayashi, S. Kawauchi and H. Nakamura, *Org. Biomol. Chem.*, 2020, **18**, 93–101.

3 P. Zhou, Y. Li, X. Li, S. Liu and Y. Su, *Liq. Cryst. Rev.*, 2016, **4**, 83–100.

4 J. Lu, Y. Sun, Q. Ge, H. Teng and Q. Jiang, *BMC Musculoskeletal Disord.*, 2014, **15**, 438.

5 A. G. Chen and D. J. Brady, *Opt. Lett.*, 1992, **17**, 441–443.

6 A. Özarslan, D. Çakmaz, F. Erol, H. Şenöz, N. Seferoğlu, A. Barsella and Z. Seferoğlu, *J. Mol. Struct.*, 2021, **1229**, 129583.

7 A. Matei, C. Constantinescu, B. Mitu, M. Filipescu, V. Ion, I. Ionita, S. Brajnicov, A.-P. Alloncle, P. Delaporte, A. Emandi and M. Dinescu, *Appl. Surf. Sci.*, 2015, **336**, 200–205.

8 C. W. Ghanavatkar, V. R. Mishra, N. Sekar, E. Mathew, S. S. Thomas and I. H. Joe, *J. Mol. Struct.*, 2020, **1203**, 127401.

9 N. N. Ayare, V. K. Shukla and N. Sekar, *Comput. Theor. Chem.*, 2020, **1174**, 112712.

10 R. Gester, A. Torres, C. Bistafa, R. S. Araújo, T. A. da Silva and V. Manzoni, *Mater. Lett.*, 2020, **280**, 128535.

11 N. N. Ayare, S. Sharma, K. K. Sonigara, J. Prasad, S. S. Soni and N. Sekar, *J. Photochem. Photobiol., A*, 2020, **394**, 112466.

12 B. Derkowska-Zielinska, E. Gondek, M. Pokladko-Kowar, A. Kaczmarek-Kedziera, A. Kysil, G. Lakshminarayana and O. Krupka, *Sol. Energy*, 2020, **203**, 19–24.

13 S. U. Mestry, U. R. Mahajan, A. M and S. T. Mhaske, *Pigm. Resin Technol.*, 2021, **50**, 231–240.

14 N. A. Alshaye, A. Z. Omar, M. Elhag, E. A. Hamed, H. A. Ahmed, N. S. Alharbi, M. A. El-Atawy, R. O. El-Zawawy and M. A. El-Rahman, *J. Mol. Struct.*, 2025, **1319**, 139582.

15 D. Jacquemin, J. Preat, V. Wathelet, M. Fontaine and E. A. Perpète, *J. Am. Chem. Soc.*, 2006, **128**, 2072–2083.

16 M. Usman Khan, F. Shafiq, M. Ramzan Saeed Ashraf Janjua, M. Khalid, J. Yaqoob, M. Arshad, S. M. Alshehri and R. Ahmad Khan, *J. Photochem. Photobiol., A*, 2024, **446**, 115115.

17 N. H. Al-Shaalan, M. U. Khan, M. R. S. A. Janjua, A. Anwar, J. Yaqoob, S. Nadeem and M. A. Amin, *J. Comput. Biophys. Chem.*, 2025, **25**, 1069–1092.

18 T. Deng, P. Hong, C. Hao and Y. Fu, *Chem. Phys.*, 2019, **523**, 70–74.

19 R. Sarkar, M. Boggio-Pasqua, P.-F. Loos and D. Jacquemin, *J. Chem. Theory Comput.*, 2021, **17**, 1117–1132.

20 Y. An, J. Miao, L. Wang, J. Fan, M. Li, M. g. Hu, M. Shao and J. Shao, *J. Text. Inst.*, 2023, **114**, 1423–1430.

21 Z. Shafiq, M. H. Tahir, S. S. A. Shah, K. M. Elhindi, M. S. Din, N. Akram and M. R. S. A. Janjua, *J. Solid State Chem.*, 2025, **345**, 125240.

22 T. Mubashir, M. Hussain Tahir, Z. Shafiq, A. Z. Dewidar, H. O. El-ansary and M. Ramzan Saeed Ashraf Janjua, *J. Photochem. Photobiol., A*, 2024, **447**, 115285.

23 A. Mahmood, S. Naeem, A. Javed, Z. Shafiq, M. A. El-Sheikh, H. O. Elansary and M. R. Saeed Ashraf Janjua, *Mater. Today Commun.*, 2024, **38**, 108403.

24 A. A. Ksenofontov, M. M. Lukanov, P. S. Bocharov, M. B. Berezin and I. V. Tetko, *Spectrochim. Acta, Part A*, 2022, **267**, 120577.

25 Z.-R. Ye, I. S. Huang, Y.-T. Chan, Z.-J. Li, C.-C. Liao, H.-R. Tsai, M.-C. Hsieh, C.-C. Chang and M.-K. Tsai, *RSC Adv.*, 2020, **10**, 23834–23841.

26 C.-W. Ju, H. Bai, B. Li and R. Liu, *J. Chem. Inf. Model.*, 2021, **61**, 1053–1065.

27 K. P. Greenman, W. H. Green and R. Gómez-Bombarelli, *Chem. Sci.*, 2022, **13**, 1152–1162.

28 D. P. Nguyen, Q. M. Le, H. T. P. Tran, P. T. Le and T. V. T. Nguyen, *ACS Omega*, 2025, **10**, 50643–50651.

29 J. F. Joung, M. Han, J. Hwang, M. Jeong, D. H. Choi and S. Park, *JACS Au*, 2021, **1**, 427–438.

30 J. Shao, Y. Liu, J. Yan, Z.-Y. Yan, Y. Wu, Z. Ru, J.-Y. Liao, X. Miao and L. Qian, *J. Chem. Inf. Model.*, 2022, **62**, 1368–1375.

31 J. Mai, T. Lu, P. Xu, Z. Lian, M. Li and W. Lu, *Dyes Pigm.*, 2022, **206**, 110647.

32 M. S. Zakerhamidi, M. Keshavarz, H. Tajalli, A. Ghanadzadeh, S. Ahmadi, M. Moghadam, S. H. Hosseini and V. Hooshangi, *J. Mol. Liq.*, 2010, **154**, 94–101.

33 A. Ghanadzadeh Gilani, V. Taghvaei, E. Moradi Rufchahi and M. Mirzaei, *J. Mol. Liq.*, 2019, **273**, 392–407.

34 U. S. Ameuru, M. K. Yakubu, K. A. Bello, P. O. Nkeonye and A. Z. Halimehjani, *Dyes Pigm.*, 2018, **157**, 190–197.

35 P. G. G. Lagrazon, A. C. Lagman, M. V. Abante, J. H. J. C. Ortega, R. A. Calderon, P. J. L. d. Castro, R. C. Maaño, M. B. Garcia, *International Conference on Information Technology Research and Innovation (ICITRI)*, 2023, pp. 118–122.

36 M. Sumita, X. Yang, S. Ishihara, R. Tamura and K. Tsuda, *ACS Cent. Sci.*, 2018, **4**, 1126–1133.

37 T. Haque, M. A. Shompa and K. Khayer, *J. Phys. Chem. A*, 2025, **129**, 1252–1279.

38 D. Jacquemin, A. Planchat, C. Adamo and B. Mennucci, *J. Chem. Theory Comput.*, 2012, **8**, 2359–2372.

39 K. Meguellati, S. Ladame and M. Spichty, *Dyes Pigm.*, 2011, **90**, 114–118.

40 K. Fu, J. Li, D. Qin, L. Shi, X. Ni, K. Zhao, D. Xu, A. Yuan and C. Zheng, *J. Text. Sci. Technol.*, 2022, **8**, 89–106.

41 P. Kumar and A. Kumar, *SAR QSAR Environ. Res.*, 2020, **31**, 697–715.

42 O. Boursalie, R. Samavi and T. E. Doyle, in *AI for Disease Surveillance and Pandemic Intelligence: Intelligent Disease Detection in Action*, ed. A. Shaban-Nejad, M. Michalowski and S. Bianco, Springer International Publishing, Cham, 2022, pp. 309–322, DOI: 10.1007/978-3-030-93080-6_22.

43 T. Schlagenhauf, Y. Lin and B. Noack, *Mach. Vis. Appl.*, 2023, **34**, 25.

44 T. T. Wong and P. Y. Yeh, *IEEE Trans. Knowl. Data Eng.*, 2020, **32**, 1586–1594.

45 T. Zhou and Y. Peng, *Struct. Multidiscip. Optim.*, 2023, **66**, 131.

46 L. H. Hall and L. B. Kier, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 1039–1045.

47 L. H. Hall, L. B. Kier and B. B. Brown, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 1074–1080.

48 E. Estrada, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 844–849.

49 A. Wale, M. Dessie and H. Kendie, *Air, Soil Water Res.*, 2022, **15**, 11786221221108216.

50 S. Lee, S. Lundberg and G. Erion, *arXiv*, 2018, preprint, arXiv:1802.03888, DOI: 10.48550/arXiv.1802.03888.

51 S. M. Lundberg and S.-I. Lee, *Adv. Neural Inf. Process. Syst.*, 2017, **30**, 4768–4777.

52 S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee, *Nat. Mach. Intell.*, 2020, **2**, 56–67.

53 V. Venkatraman, S. Abburu and B. K. Alsberg, *Chemom. Intell. Lab. Syst.*, 2015, **142**, 87–94.