UNIVERSIDADE FEDERAL DE GOIÁS (UFG)
INSTITUTO DE QUÍMICA (IQ)
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

RAFAEL FERREIRA VERÍSSIMO

Integrating Machine Learning and SHAP Analysis to Advance the Rational Design of Benzothiadiazole Derivatives with Tailored Photophysical Properties
Integração de Aprendizado de Máquina e Análise SHAP para Avançar no Projeto Racional de Derivados de Benzotiadiazola com Propriedades Fotofísicas Personalizadas

GOIÂNIA - GO
2026

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE QUÍMICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[X] Dissertação [ ] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

Rafael Ferreira Veríssimo

3. Título do trabalho

"Integrating Machine Learning and SHAP Analysis to Advance the Rational Design of Benzothiadiazole Derivatives with Tailored Photophysical Properties"

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:

a) consulta ao(à) autor(a) e ao(à) orientador(a);

b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.

O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Heibbe Cristhian Benedito De Oliveira**, **Professor do Magistério Superior**, em 16/03/2026, às 20:54, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rafael Ferreira Veríssimo**, **Discente**, em 09/04/2026, às 11:43, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6046544** e o código CRC **6483084E**.

Referência: Processo nº 23070.002803/2026-18 SEI nº 6046544

RAFAEL FERREIRA VERÍSSIMO

Integrating Machine Learning and SHAP Analysis to Advance the Rational Design of Benzothiadiazole Derivatives with Tailored Photophysical Properties
Integração de Aprendizado de Máquina e Análise SHAP para Avançar no Projeto Racional de Derivados de Benzotiadiazola com Propriedades Fotofísicas Personalizadas

Dissertação apresentada ao Programa de Pós-Graduação em Química, do Instituto de Química (IQ) da Universidade Federal de Goiás (UFG), como requisito para obtenção do título de Mestre em Química.

Área de concentração: Química Teórica, Química Computacional, Aprendizado de Máquina

Orientador(a): Dr. Heibbe Cristhian Benedito de Oliveira
Coorientador(a): Dr. Flávio Olimpio Sanches Neto

GOIÂNIA
2026

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Ferreira Veríssimo, Rafael

Integrating Machine Learning and SHAP Analysis to Advance the Rational Design of Benzothiadiazole Derivatives with Tailored Photophysical Properties [Manuscrito] = Integração de Aprendizado de Máquina e Análise SHAP para Avançar no Projeto Racional de Derivados de Benzotiadiazola com Propriedades Fotofísicas Personalizadas / Rafael Ferreira Veríssimo. - 2026.

LXXII, 72 f.: il.

Orientador: Prof. Dr. Heibbe Cristhian Benedito de Oliveira ; co-orientador: Dr. Flávio Olimpio Sanches Neto

Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Química (IQ), Programa de Pós-Graduação em Química, Goiânia, 2026.

Ilustrações.

Anexo.

Apêndice.

Bibliografia.

Inclui: tabelas, grafico, lista de figuras, lista de tabelas.

1. Aplicação Web. 2. QSPR. 3. Impressão Digital Molecular. 4. SHAP. 5. Benzotiadiazol.
I. Benedito de Oliveira, Heibbe Cristhian , orient. II. Olimpio Sanches Neto, Flávio , co-orient. III. Título.

CDU 54

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE QUÍMICA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº 01 da sessão de **Defesa de Mestrado de Rafael Ferreira Veríssimo** estudante de **Mestrado** no Programa de Pós-Graduação Em Química da Universidade Federal de Goiás - UFG.

Aos **26 (vinte e seis dias) do mês de fevereiro de 2026 (dois mil e vinte e seis)**, a partir das **14h00m**, via **Anfiteatro do Bloco 02 - Instituto de Química,** realizou-se a sessão pública de defesa de **Mestrado** intitulada ***"Integrating Machine Learning and SHAP Analysis to Advance the Rational Design of Benzothiadiazole Derivatives with Tailored Photophysical Properties".*** Os trabalhos foram instalados pelo Orientador, **Prof. Dr. Heibbe Cristhian Benedito de Oliveira (IQ-UFG)**, com a participação dos demais membros da **Banca Examinadora: Prof. Dr. Luciano Ribeiro (UEG - Anápolis)**, **Profª. Drª. Aline Silva Muniz (Lund University, Sweden)**. Durante a arguição, os membros da banca **não fizeram** sugestão de alteração do título do trabalho. Após a arguição do candidato, a Banca Examinadora se reuniu em sessão secreta, a fim de concluir o julgamento da Defesa de Mestrado em andamento, tendo sido o candidato **aprovado** em seu Exame de Qualificação. Proclamados os resultados pelo **Prof. Dr. Heibbe Cristhian Benedito de Oliveira**, **Presidente da Banca Examinadora,** foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, **26 (vinte e seis dias) do mês de fevereiro de 2026 (dois mil e vinte e seis).**

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Heibbe Cristhian Benedito De Oliveira**, **Professor do Magistério Superior**, em 23/07/2026, às 11:23, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **ALINE SILVA MUNIZ**, **Usuário Externo**, em 30/07/2026, às 16:05, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Luciano Ribeiro**, **Usuário Externo**, em 04/08/2026, às 15:36, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6358021** e o código CRC **FF57B451**.

Referência: Processo nº 23070.002803/2026-18 SEI nº 6358021

"Why, still, should the lust for ascension
Seem, in itself, so close to madness?"

Yukio Mishima, Sun & Steel

In memoriam to my beloved Grandfather,

Celso da Costa Milagre.

Acknowledgements

Gostaria de expressar minha sincera gratidão aos meus pais, Meire Cristina Ferreira e Wesley Veríssimo Milagre, pelo apoio incondicional e por sempre acreditarem em meu potencial. Agradeço também aos meus familiares que constantemente me auxiliaram financeiramente e emocionalmente, em especial meus avós Antonio Luis Ferreira e Darvina Maria, minha tia Mariene e minha madrinha Ivonete.

Sou grato também aos meus amigos e colegas de laboratório, com um agradecimento especial a Pedro, Mateus, Ana Gabriela e Carol, que sempre me apoiaram e demonstraram tolerância desde o início da minha jornada no laboratório. Estendo minha gratidão ao meu orientador, Heibbe Cristhian Benedito de Oliveira, por ter me acolhido e acreditado em minhas capacidades, e ao meu co-orientador, Flávio Olimpio Sanches Neto, pelos valiosos momentos de aprendizado e descontração. Agradeço à CAPES pelo suporte financeiro concedido através da bolsa de estudos, que tornou possível a conclusão deste trabalho. Meu reconhecimento se estende a todos que, de alguma forma, contribuíram para a realização deste projeto.

Por fim, dedico este trabalho à memória do meu avô, Celso da Costa Milagre.

Resumo

O desenvolvimento de fluoróforos orgânicos com propriedades ópticas ajustáveis é essencial para o avanço de tecnologias em optoeletrônica e bioimagem. Entre eles, os derivados da 2,1,3-benzotiadiazol (BTD) se destacam por sua forte conjugação π, alta eficiência de fluorescência e capacidade de formar estruturas ordenadas. No entanto, prever racionalmente suas propriedades fotofísicas ainda é um desafio, devido à interação complexa entre efeitos de substituintes, polaridade do solvente e transições eletrônicas. Métodos tradicionais, baseados em química quântica, são precisos, mas computacionalmente dispendiosos, dificultando sua aplicação em triagens de compostos em larga escala. Para preencher essa lacuna, este trabalho propõe uma estratégia baseada em dados que combina *machine learning* (ML) e inteligência artificial explicativa para acelerar a previsão e a interpretação das propriedades fotofísicas de derivados de BTD. O principal objetivo foi desenvolver modelos preditivos robustos para os comprimentos de onda de absorção e emissão máximos, além de obter conhecimento mecanísticos por meio da técnica *SHapley Additive exPlanations* (SHAP). A metodologia envolveu a construção de um conjunto de dados com derivados de BTD, geração de descritores moleculares por meio de *morgan fingerprints* (MF) e o treinamento de modelos usando os algoritmos *Random Forest*, *LightGBM* e *XGBoost*. A validação foi realizada com validação cruzada (*10-fold*) e *y-randomization*, avaliando o desempenho com métricas como R^2, Q^2, MAE e RMSE. Os resultados mostraram que os modelos de ML apresentaram alta acurácia preditiva, com destaque para o *Random Forest*, que superou os demais em ambas as propriedades estudadas. A análise SHAP identificou grupos doadores de elétrons, especialmente aminas terciárias, e a polaridade do solvente como determinantes chave no comportamento óptico. A ferramenta *web* desenvolvida permite a previsão em tempo real e a visualização das contribuições estruturais, promovendo a integração entre química computacional e design racional de compostos. Esses achados reforçam o potencial do ML interpretável para guiar o desenvolvimento racional de materiais orgânicos, reduzindo a dependência de cálculos custosos e ampliando o acesso a essas ferramentas.

Palavras-chave: Aplicação Web, QSPR, Impressão Digital Molecular, SHAP, Engenharia de Compostos, Benzotiadiazola.

Abstract

The design of organic fluorophores with tailored optical properties is critical for advancing technologies in optoelectronics and bioimaging. Among them, 2,1,3-benzothiadiazole (BTD) derivatives stand out due to their strong π-conjugation, high fluorescence efficiency, and ability to form ordered structures. However, rationally predicting their photophysical properties remains challenging due to the complex interplay between substituent effects, solvent polarity, and electronic transitions. Traditional methods relying on quantum chemistry are accurate but computationally expensive, limiting their scalability for high-throughput compound screening. To address this gap, this work proposes a data-driven strategy combining machine learning (ML) and explainable artificial intelligence to accelerate the prediction and interpretation of photophysical properties in BTD derivatives. The primary objective was to develop robust ML models for predicting maximum absorption and emission wavelengths and to derive mechanistic insights via SHapley Additive exPlanations (SHAP). The methodology involved curating a dataset of BTD derivatives, generating molecular descriptors through Morgan fingerprints, and training models using Random Forest, LightGBM, and XGBoost algorithms. Model validation employed 10-fold cross-validation and y-randomization, with performance assessed via R^2, Q^2, MAE, and RMSE metrics. The results demonstrated that ML models achieved high predictive accuracy, with the Random Forest model outperforming others for both photophysical endpoints. SHAP analysis identified electron-donating groups, particularly tertiary amines, and solvent polarity as key determinants of optical behavior. The deployed web tool enables real-time property prediction and substructure attribution, thus bridging computational chemistry and practical compound design. These findings underscore the potential of interpretable ML to guide the rational design of organic materials, reducing reliance on expensive calculations and fostering accessibility for chemists and material scientists.

Keywords: Web Application, QSPR, Molecular Fingerprint, SHAP, Compound Engineering, Benzothiadiazole

Contents

List of Figures

List of Tables

List of Abbreviations

XGBoost	**X**treme **G**radient **Boost**ing
RF	**R**andom **F**orest
LightGBM	**L**ight **G**radient **B**oosting **M**achine
SHAP	**SH**apley **A**dditive ex**P**lanations
SMILES	**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem
MF	**M**organ **F**ingerprint
DFT	**D**ensity **F**unctional **T**heory
TD-DFT	**T**ime-**D**ependent **D**ensity **F**unctional **T**heory
HF	**H**artree **F**ock
HK	**H**ohenberg-**K**ohn
KS	**K**ohn-**S**ham
QSPR	**Q**uantitative **S**tructure **P**roperty **R**elationships
ML	**M**achine **L**earning
BTD	2,1,3 **B**enzo**th**ia**d**iazole
ICT	**I**ntramolecular **C**harge **T**ransfer
$E_{\mathbf{T}}^{\mathbf{N}}$	**N**ormalized Molar **E**lectronic **T**ransition Energies

Chapter 1

Introduction

2,1,3 Benzothiadiazoles (BTD) and its derivatives are highly efficient fluorophores[1] known for forming well-ordered crystalline structures[2]. Their pronounced molecular polarity facilitates strong intermolecular interactions, including heteroatom contacts and π-π stacking[3]. These features render BTD scaffolds effective electron-accepting units in the design of organic semiconducting materials,[4] making them particularly valuable for applications in optoelectronic devices[5,6] and organic light-emitting technologies[7].

Such adaptability has led to their incorporation in devices requiring finely controlled emission spectra, as well as their application in advanced fluorescence imaging, including the selective staining of mitochondria in cancer cell lines.[8] Identifying optimal BTD derivatives with targeted photophysical properties remains a significant challenge, as traditional approaches rely on iterative synthesis, extensive experimental screening, and high-level computational chemistry.[9] While insightful, quantum chemical methods in particular can be too computationally expensive and time-consuming for the demands of rapid compound discovery, highlighting the need for more efficient predictive strategies.[10–12]

This predictive challenge is compounded by the subtle interplay between substituents, solvent polarity, and molecular conformation on the electronic transitions that govern absorption and emission.[13] Consequently, a clear need exists for more accessible, efficient, and systematic strategies to predict these photophysical properties without relying exclusively on expensive computational methods.

In recent years, machine learning (ML) methodologies have emerged as robust alternatives to traditional electronic structure methods, offering significant reductions in computational overhead. While Time-Dependent Density Functional Theory (TD-DFT) remains the gold standard for investigating excited-state dynamics, it is characterized by substantial resource demands; for instance, calculating ground- and excited-state geometries can average 236 minutes per molecule. In contrast, deep learning architectures, such as the model proposed by Joung et al., can predict these photophysical properties in approximately 20 seconds[14]. This orders-of-magnitude acceleration facilitates the high-throughput screening of extensive chemical libraries that would otherwise be computationally prohibitive.

A particularly promising approach is the use of molecular fingerprints, which encode the

presence or absence of specific substructures as binary vectors.[15] These fingerprints provide an effective way to capture chemical information in a format suitable for machine learning algorithms, enabling the prediction of molecular properties based on structural features without the need for extensive quantum chemical calculations.

Recent studies have demonstrated that ML models combined with molecular fingerprints can accurately predict a number of chemical properties, including kinetic parameters and other physicochemical traits.[15–17] The accuracy of these models is often comparable to, or even surpasses, conventional Quantitative Structure-Activity Relationship (QSAR) and Quantitative Structure-Property Relationship (QSPR) methods.

For example, Ju et al. compiled a dataset of over 3.000 organic fluorescent dyes and used it to train ML models that achieved mean absolute errors of 0.13 eV for quantum yield and 0.080 eV for emission wavelengths, providing a fast and efficient alternative to quantum chemical calculations.[18] Building on the success of initial studies, Mai et al. introduced an interpretable ML strategy for azo dyes. Their model pinpointed structural patterns as key contributors to red-shifts in the maximum absorption wavelength and successfully identified 26 new azo molecules with enhanced λ_{max} values through high-throughput screening.[19] Surpassing previous state-of-the-art methods, a ML model developed by Mahato and Kumar predicted the photophysical properties of 3.066 organic dyes. The model demonstrated high predictive accuracy for absorption wavelength, achieving an R^2 value of 0.961, and also predicted emission wavelengths and quantum yields.[17]

Interpretable machine learning techniques, such as SHapley Additive exPlanations (SHAP), can shed light on how specific structural motifs influence photophysical behavior. This mechanistic understanding is invaluable as it enables researchers to rationalize structure-property relationships and to guide the design of new compounds with desired properties. Moreover, the development of user-friendly online platforms can broaden access to these predictive tools for researchers across various fields, facilitating collaborations and expediting advancements in photophysical material design.[20]

In this context this work establishes an integrated, interpretable ML framework for the rational design of BTD derivatives. A ML model was developed to estimate absorption and emission wavelengths with high accuracy. Beyond predictive performance, this study employs SHAP-based interpretability to map specific structural motifs and substituents to their corresponding photophysical effects. The resulting workflow provides both a mechanistic understanding of structure-property relationships and a user-friendly interface to accelerate the discovery of BTD derivatives with tailored optical properties.

Chapter 2

Objectives

- Develop and Validate Predictive Models: The primary objective was to develop machine learning models, including Random Forest, LightGBM, and XGBoost, for the accurate prediction of the photophysical properties of benzothiadiazole (BTD) derivatives. A key part of this goal was to validate the model's performance using robust statistical metrics such as R^2, Q^2, MAE, and RMSE.

- Gain Mechanistic Insight Using SHAP: A central aim of this research was to employ SHAP (SHapley Additive exPlanations) to interpret the complex models. This was proposed to move beyond simple predictive accuracy and uncover clear, mechanistic insights into how specific molecular features, like tertiary amine substituents, and environmental factors, such as solvent polarity, govern the photophysical outcomes.

- Establish a Framework for Rational Design: The overarching goal was to leverage the synergy between predictive modeling and SHAP interpretation to establish a data-driven strategy for the rational design of new BTD fluorophores. This framework was intended to guide targeted structural modifications to achieve desired optical properties more efficiently.

- Create a Easy to Use Web Application: Another objective was to develop a user-friendly web application that allows researchers to input molecular structures and receive predictions of their photophysical properties. This tool was designed to democratize access to the predictive models and facilitate broader use in the scientific community.

Chapter 3

Theoretical Foundations

This chapter presents a conceptual and theoretical foundation essential for understanding the methodology and results discussed throughout this work. It begins with a review of key principles in photochemistry, as these underpin the photophysical phenomena investigated. Subsequently, a section is dedicated to Density Functional Theory, with particular emphasis on the time-dependent extension employed to derive electron density-based descriptors that provide insight into excited-state behavior. The discussion then turns to an overview of supervised machine learning, highlighting the decision tree-based algorithms used herein—Random Forest, XGBoost, and LightGBM—along with a distinction between regression and classification models, and supervised versus unsupervised learning paradigms. Finally, a section introduces SHAP, focusing on its capacity to quantify the relative importance of specific molecular substructures and solvent parameters in predicting maximum absorption and emission wavelengths.

3.1 Photochemistry

Photochemistry is the study of chemical reactions initiated by the absorption of light, typically in the visible and near-ultraviolet regions of the electromagnetic spectrum. When a molecule absorbs a photon, it transitions to an electronically excited state. This excited state possesses excess energy, which can be dissipated through various processes, including chemical reactions. These light-induced reactions are valuable because they can access pathways and products that are not attainable through thermal processes alone.[21]

3.1.1 Photophysical Principles: Luminescence

Luminescence, generally defined as the emission of photons by a substance, originates from electronically excited states. This process is broadly categorized into two primary types: fluorescence and phosphorescence, distinguished by the spin multiplicity of the excited state from which emission occurs.[22]

In fluorescence, emission proceeds from an excited singlet state (S_n, typically S_1). In such a state, the electron occupying the excited molecular orbital maintains an opposite spin orientation relative to its paired counterpart in the ground state orbital. This spin-paired

configuration ensures that the transition back to the ground electronic state (S_0) is spin-allowed, resulting in a rapid radiative decay via photon emission. Consequently, fluorescence typically manifests on timescales ranging from nanoseconds to microseconds.[23]

In contrast to fluorescence, phosphorescence is characterized by photon emission from an excited triplet state (T_n, typically T_1). In this spin-forbidden transition, the electron in the excited molecular orbital possesses the same spin orientation (parallel spin) as its paired counterpart in the ground-state orbital.[23] Consequently, the direct radiative transition from a triplet state to the singlet ground state ($T_1 \rightarrow S_0$) is spin-forbidden according to selection rules. This spin-forbidden nature results in significantly slower emission rates compared to fluorescence, with typical phosphorescence lifetimes ranging from microseconds to seconds, or even longer.[22] A critical characteristic of phosphorescence is its usual absence in fluid solutions at ambient temperature. This is primarily due to the high efficiency of various non-radiative deactivation processes that effectively compete with the slow phosphorescence emission.[22–24]

3.1.1.1 Jablonski Diagrams

Jablonski diagrams are graphical representations that illustrate the electronic states of a molecule and the transitions between them.[25] The singlet ground and subsequent excited electronic states are typically depicted by S_0, S_1, S_2, ..., S_n. At each of these electronic energy levels, the fluorophore can exist in a number of vibrational energy levels, which are depicted by 0, 1, 2, etc and energy losses between excitation and emission are observed for fluorescent molecules in solution. Furthermore, fluorophores generally decay to higher vibrational levels of the ground electronic state (S_0), resulting in a further loss of excitation energy through thermalization of the excess vibrational energy as pointed in figure 3.1.[23]

Transitions between electronic states due to light absorption are depicted as vertical lines in a Jablonski diagram. This vertical representation illustrates the instantaneous nature of light absorption, which occurs on a timescale of approximately 10^{-15} seconds. This timescale is too short for significant displacement of the atomic nuclei, a principle rigorously described by the Franck-Condon principle.[23] A example of a Jablonski diagram is shown in Figure3.1.

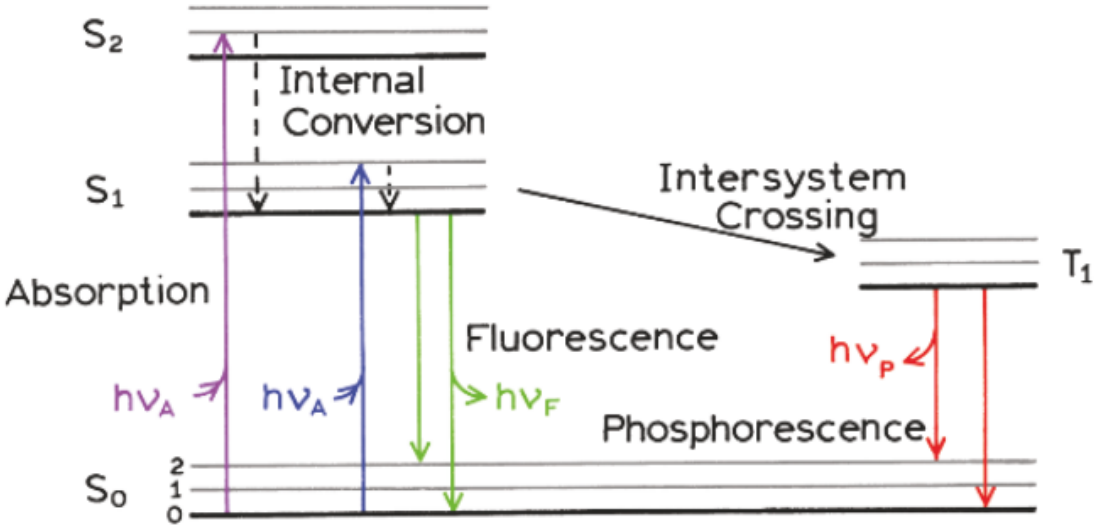


Figure 3.1: Example of a Jablonski diagram.

3.1.2 Franck-Condon Principle and Solvation Effects

The Franck-Condon principle[26] is a foundational concept in photophysics, asserting that electronic transitions occur instantaneously relative to the timescale of nuclear motion or solvent reorientation. Consequently, upon light absorption, the molecule transitions to an initial excited state, termed the Franck-Condon excited state, which retains the nuclear geometry and solvation shell configuration of the preceding equilibrium ground state.

If the excited state lifetime is sufficiently long, the altered charge distribution of the excited molecule induces a dynamic reorientation of the surrounding solvent molecules. This relaxation process leads to the formation of a relaxed excited state, characterized by a solvent shell that has achieved equilibrium with the excited fluorophore. Critically, it is from this relaxed excited state that fluorescence emission predominantly occurs. This solvent-induced stabilization often results in a bathochromic shift (red-shift) of the emission spectrum relative to the absorption spectrum.[23]

Analogously, following photon emission, the molecule initially populates a Franck-Condon ground state. This transient state momentarily possesses the nuclear geometry and solvation pattern of the relaxed excited state from which emission occurred. However, this non-equilibrium configuration rapidly relaxes as solvent molecules reorganize, allowing the system to return to its thermodynamically stable equilibrium ground state.[23]

The different solvation of these two states, specifically the equilibrium ground state and the relaxed excited state, is fundamentally responsible for the influence of the solvent environment on emission or fluorescence spectra. This phenomenon, where the spectral position of absorption or emission bands changes with solvent properties, is broadly termed solvatochromism[27].[24]

While terms such as solvatofluorochromism[28] or fluorosolvatochromism[29] have occasionally been employed to specifically describe the solvent dependence of fluorescence spectra, it is generally accepted within the field that the overarching term solvatochromism sufficiently encompasses both absorption and emission spectral shifts. This is due to the inherent spectroscopic connection and mechanistic similarities between the absorption and emission processes of a given fluorophore, rendering distinct nomenclature for fluorescence-based solvent effects largely unnecessary.[30]

The observed solvatochromism is intricately dependent on the chemical structure and intrinsic physical properties of the chromophore, alongside the characteristics of the solvent molecules. These factors collectively determine the strength and nature of intermolecular solute-solvent interactions in both the equilibrium ground state and the Franck-Condon excited state.

Generally, chromophores exhibiting a substantial change in their permanent dipole moment upon electronic excitation display pronounced solvatochromism. If the solute's dipole moment increases during the electronic transition ($\mu_g < \mu_e$), a positive solvatochromism is typically observed, manifesting as a bathochromic shift (red-shift) of the emission band with increasing solvent polarity, as more polar solvents preferentially stabilize the more dipolar excited state.[30]

Conversely, if the solute's dipole moment decreases upon excitation ($\mu_g > \mu_e$), a negative solvatochromism usually results, wherein increasing solvent polarity leads to a hypsochromic shift (blue-shift) of the emission band, as the more polar ground state is stabilized to a greater extent than the less polar excited state. Compounds exhibiting this particular solvatochromic behavior are commonly found among systems characterized by significant inter or intramolecular charge transfer (ICT) absorptions,[24,31–33] where the electronic transition involves a substantial redistribution of electron density.

In addition to the change in permanent dipole moment upon excitation, the capacity of a solute to engage in hydrogen bonding (donating or accepting) with surrounding solvent molecules, in both its ground and Franck-Condon excited states, further dictates the extent and sign of its solvatochromism.[34–38] Specific hydrogen-bonding interactions can lead to

additional stabilization or destabilization of either the ground or excited state, thereby modulating the observed spectral shifts beyond what is explained by generalized polarity scales alone.

It is also possible to observe inverted solvatochromism, a less common but significant phenomenon. In this case, the absorption or emission band, typically a long-wavelength charge transfer band, initially exhibits a bathochromic shift with increasing solvent polarity, but then undergoes a hypsochromic shift as solvent polarity continues to increase. This behavior is a direct consequence of a solvent-induced change in the electronic ground-state structure of the chromophore, often involving a transition from a less dipolar to a more dipolar ground state with increasing solvent polarity, or a change in the dominant solute-solvent interaction type.[39,40]

3.1.3 $E_T(30)$ and (E_T^N)

$E_T(30)$ values are widely utilized empirical solvent polarity scale, which are based on the negatively solvatochromic pyridinium *N*-phenolate betaine dye as the probe molecule. These values are simply defined, in analogy to Kosower's *Z* values[41], as the molar electronic transition energies (E_T) of the dissolved probe molecule 30[42], measured in normal conditions of temperature and pressure[30,42]. The $E_T(30)$ value is determined according to Equation3.2, where ν_{max} is the frequency of the longest-wavelength absorption band of the betaine dye 30 in the respective solvent. The $E_T(30)$ values, expressed in kcal mol^{-1}, are calculated using the following relationships:

$$E_T(30) = hc\nu_{max} = (2.8591 \times 10^{-3}) \cdot \nu_{max}(\text{cm}^{-1}) = \frac{28591}{\lambda_{max}(\text{nm})}. \tag{3.1}$$

Where h is Planck's constant, c is the speed of light, ν_{max} is the frequency in cm^{-1} of the longest-wavelength, ICT ($\pi \rightarrow \pi^*$) absorption band of the probe dye 30, and λ_{max} is the corresponding wavelength in nanometers (nm). The numerical constants incorporate the necessary unit conversions to yield $E_T(30)$ in kcal mol^{-1}.

$$E_T(30) = h\nu_{max} = \frac{hc}{\lambda_{max}}. \tag{3.2}$$

In the initial publication defining this scale,[42] the betaine dye, by chance, was assigned the formula number 30. Therefore, the number (30) was subsequently appended to the E_T designation to prevent confusion with E_T, which is commonly used in photochemistry as an abbreviation for triplet energy.

Normalized molar electronic transition energies (E_T^N) are dimensionless parameters that provide a quantitative measure of solvent polarity. It is defined using water and tetramethylsilane (TMS) as reference solvents, representing extreme polar and nonpolar environments, respectively. Consequently, the E_T^N scale ranges from 0.000 for TMS, the least polar solvent, to 1.000 for water, the most polar solvent. The E_T^N value of a solvent is calculated using the following equation:[43]

$$E_T^N = \frac{E_T(solvent) - E_T(TMS)}{E_T(water) - E_T(TMS)} = \frac{E_T(solvent) - 30.7}{32.4}. \tag{3.3}$$

where $E_T(\text{solvent})$ is the electronic transition energy of the solvent in question, $E_T(\text{TMS})$ is the electronic transition energy of tetramethylsilane (TMS), and $E_T(\text{water})$ is the electronic transition energy of water.

3.1.4 2,1,3 Benzothiadiazoles

BTD and its derivatives are π-conjugated heterocycles comprising a fused benzene ring and a 1,2,3-thiadiazole unit. Due to the electronegativity of the nitrogen atoms and the vacant d-orbitals of sulfur, the BTD core acts as a potent electron-withdrawing group (EWG). This inherent electron deficiency renders BTD derivatives excellent candidates for electron-transport materials and active redox centers. In donor-acceptor (D-A) architectures, the BTD moiety facilitates efficient intramolecular charge transfer, a property that can be systematically tuned via substituent effects to modulate the frontier molecular orbital (FMO) energy levels and, consequently, the optoelectronic response.[44]

They typically exhibit high fluorescence quantum yields and form well-ordered crystalline architectures, a consequence of their highly polarized nature which promotes robust intermolecular interactions, including heteroatom-heteroatom contacts and π–π stacking. Furthermore, their relatively high electron affinity (EA) and reduction potential make them suitable candidates for electron-injection layers in light-emitting diodes (LEDs). The frontier molecular orbital (FMO) energy levels—the Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbital (LUMO)—of these π-extended systems are intrinsically linked to their ionization potential (IP) and EA. These parameters correlate directly with electrochemical oxidation and reduction potentials, as well as the optical bandgap (E_g) derived from the absorption onset in the solid state.[44,45]

3.2 Machine Learning Models

Instead of deriving physical laws from first principles, machine learning (ML) establishes numerical relationships between easily accessible input variables and target properties. These target properties are often too complex to solve analytically or predict directly from fundamental theory. Generally, ML algorithms are categorized into three main paradigms: supervised learning, which is trained on labeled datasets; unsupervised learning, aimed at classifying unlabeled data; and reinforcement learning, which utilizes reward mechanisms to guide data learning. Among these, supervised learning is the predominant approach in chemical applications, owing to its superior numerical predictability for specific chemical and physical properties.[46]

3.2.1 Decision Trees

Decision trees represent one of the most popular and enduring techniques for discriminatory learning, having been developed independently within both the statistical[47,48] and machine learning[49,50] communities. A decision tree is a tree-structured classification model that is notably easy to understand, even for non-expert users, and can be efficiently induced from data.[51]

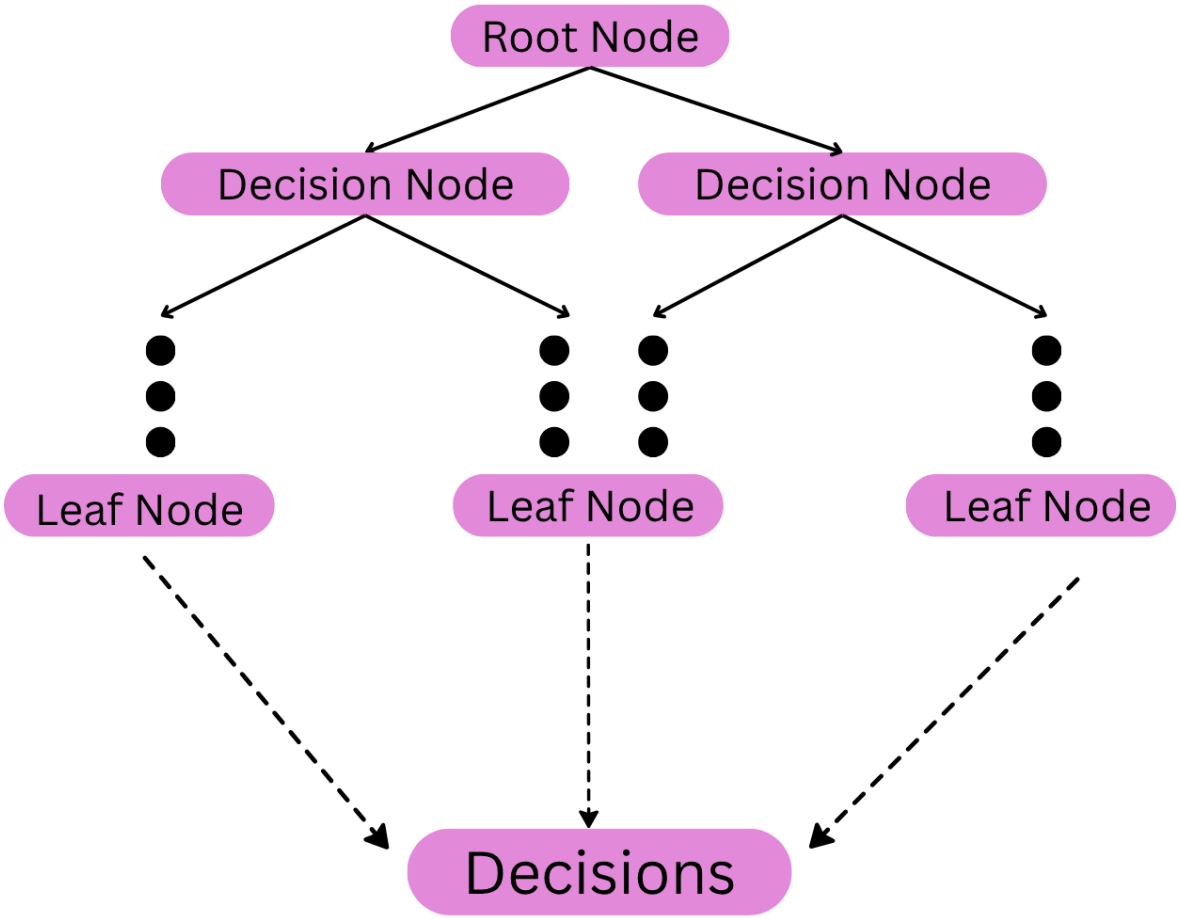


Figure 3.2: Schematic representation of a decision tree model. The tree consists of a Root Node, internal Decision Nodes, and terminal Leaf Nodes. Data flows from the Root Node through Decision Nodes, where rules based on input features guide the path, eventually leading to a final classification or regression Decision at a Leaf Node.

The decision process within a decision tree in Figure 3.2 commences at the Root Node, representing the initial point where the entire dataset begins to be partitioned. From the Root Node, the data proceeds through a series of internal Decision Nodes. At each Decision Node, a specific rule or condition, based on one of the input features, is applied to the data. This rule splits the data into subsets, directing them down different branches of the tree. This branching continues until the data reaches a Leaf Node, each Leaf Node represents a final outcome, which could be a class prediction in classification tasks or a numerical value in regression tasks. The path from the Root Node to a specific Leaf Node represents a unique sequence of decisions or rules, ultimately leading to the final output.[52]

Decision trees are typically learned in a top-down, recursive manner, commonly referred

to as a top-down induction of decision trees (TDIDT) algorithm,[53] recursive partitioning, or a divide-and-conquer approach. The algorithm begins by selecting the optimal feature to serve as the root of the tree, it then splits the initial set of examples into subsets based on the values of this chosen feature and adds corresponding branches to the tree for each subset. The process is then recursively applied to each new subset. For discrete attributes, the simplest splitting criterion involves a test of the form $(A = v)$, where A is the selected attribute and v is one of its possible discrete values.

3.2.1.1 Regression Trees

Regression trees are a specific type of decision tree tailored for regression tasks. As supervised learning methods, they provide a tree-based approximation $\hat{f}$ of an unknown regression function $Y = f(x) + \varepsilon$, where $Y \in \mathbb{R}$ and ε is typically assumed to be normally distributed ($\mathcal{N}(0, \sigma^2)$). This approximation is constructed based on a set of training data $D = \{(x_i, y_i)\}_{i=1}^{n}$. The resulting tree structure consists of a hierarchy of logical tests applied to the values of any of the p predictor variables. In this context, the leaf nodes from Figure 3.2 contain the numerical predictions for the target variable Y.[52]

Binary trees with logical tests are the most common type of regression trees. These tests frequently take the form $X_i < \alpha$, where $\alpha \in \mathbb{R}$, for numerical variables. Tests on nominal variables, conversely, typically have the form $X_j \in \{v_1, \ldots, v_m\}$, where $v_1, \ldots, v_m$ are a subset of the possible values for attribute X_j. Each unique path from the root node to a leaf node effectively represents a logical assertion that defines a specific region within the predictor space. Consequently, any regression tree provides a complete and mutually exclusive partition of the predictor space into L regions. These regions are characterized by boundaries that are parallel to the predictor axes, directly resulting from the axis-parallel nature of the tests.

The approximation provided by a regression tree is defined as follows:

$$Y = \sum_{l \in \mathscr{L}} k_l \cdot I(P_l), \tag{3.4}$$

where Y represents the predicted output or target variable for a given observation, as determined by the regression tree. The summation is performed over $\mathscr{L}$, which denotes the set of all terminal or leaf nodes in the tree. For each leaf node l within this set, k_l is the constant numerical prediction assigned to that specific leaf; this is typically the average or median of the target values from the training data that fall into the region defined by leaf node l. The term P_l refers to the distinct region in the predictor space that corresponds to leaf node l. Finally, $I(P_l)$ is an indicator function that takes a value of 1 if the input observation's

features place it within the region P_l defined by leaf node l, and 0 otherwise. Since the regions defined by the leaf nodes form a mutually exclusive partition of the predictor space, for any given input, only one indicator function $I(P_l)$ will be active (equal to 1), meaning the predicted value Y will correspond directly to the k_l value of the relevant leaf node.

Binary decision trees are constructed using a recursive partitioning algorithm, as previously mentioned. At each node, the algorithm selects the optimal logical test to divide the data into two subsets, with the objective of minimizing the variance of the target variable within each resulting subset. This splitting process continues iteratively until a predefined stopping criterion is satisfied. Such criteria commonly include reaching a maximum allowed tree depth or ensuring a minimum number of samples in a leaf node. The outcome is a hierarchical tree structure where each distinct path from the root to a leaf node represents a unique sequence of decisions based on the input features, ultimately culminating in a prediction for the target variable.

Algorithm 1, as typically presented, encapsulates three main components that define the specific characteristics of the regression tree: (i) the termination criterion, (ii) the method for determining the constant k_l value assigned to each leaf node, and (iii) the procedure for identifying the optimal test for splitting data at each internal node. The selection of these components is directly influenced by the preference criteria employed during the tree building process. The choices for these components are related to the preference criteria that are used to buold the trees. The selection of these components is directly influenced by the preference criteria employed during the tree building process. A widely adopted criterion for this purpose is the minimization of the sum of squared errors, commonly known as the least squares (LS) criterion[54].[52]

Algorithm 1 RecursivePartitioning(D)

Require: D, a sample of cases, $\{(x_i^1, \cdots, x_i^p, y_i)\}$
Ensure: t, a tree node
1: **if** <*TERMINATION CRITERION*> **then**
2: **return** a leaf node with <*CONSTANT* k_l>
3: **else**
4: $t \leftarrow$ new tree node
5: t.split $\leftarrow$ *FIND THE BEST TEST ON ONE OF THE VARIABLES*
6: t.leftNode $\leftarrow$ RECURSIVEPARTITIONING($\{x \in D : x \models t.\text{split}\}$)
7: t.rightNode $\leftarrow$ RECURSIVEPARTITIONING($\{x \in D : x \not\models t.\text{split}\}$)
8: **return** the node t
9: **end if**

The termination criterion for tree growth is generally set to relaxed settings, allowing

an overly large tree to be initially grown. The rationale behind this strategy is that these extensive trees will be pruned afterward, with the primary goal of overcoming the problem of overfitting to the training data. According to the LS criterion, the error at a given node is calculated as:

$$\mathrm{Err}(t) = \frac{1}{n_t} \sum_{(x_i, y_i) \in D_t} (y_i - k_t)^2, \tag{3.5}$$

where $\mathrm{Err}(t)$ quantifies the error at a specific tree node t. The term n_t represents the total number of data samples that fall into node t. The summation runs over all data points (x_i, y_i) belonging to the dataset D_t associated with node t, where x_i denotes the feature vector for the i-th sample and y_i is its corresponding true target value. Finally, k_t is the constant predicted value for the target variable y at node t, which is typically the average of all y_i values in D_t for regression tasks.[52]

Any logical test s applied at a node t divides the cases in its associated dataset D_t into two partitions, D_{tL} (left subset) and D_{tR} (right subset). The resulting pooled error from this split is given by:

$$\mathrm{Err}(t,s) = \frac{n_{tL}}{n_t} \times \mathrm{Err}(t_L) + \frac{n_{tR}}{n_t} \times \mathrm{Err}(t_R), \tag{3.6}$$

where $\mathrm{Err}(t,s)$ denotes the resulting pooled error at node t after applying a specific splitting test s. The terms n_{tL} and n_{tR} represent the number of data samples in the left and right child nodes resulting from the split, respectively, while n_t is the total number of samples in the parent node t before the split. $\mathrm{Err}(t_L)$ is the error calculated for the left child node, and $\mathrm{Err}(t_R)$ is the error calculated for the right child node, each determined using the sum of squared errors criterion. The equation effectively computes a weighted average of the errors in the two resulting child nodes, with the weights being their respective proportions of the data from the parent node.[52]

And it is possible to estimate the value of the split s by its respective error reduction. This error reduction can then be used to evaluate all candidate splitting tests for a node:

$$\Delta(s,t) = \mathrm{Err}(t) - \mathrm{Err}(t,s) \tag{3.7}$$

Finding the optimal splitting test for a given node t involves evaluating all possible candidate tests for that node using Equation 3.7. This process necessitates assessing every potential split point for each predictor variable in the dataset.[52]

3.2.2 Random Forests

Random Forest[55] is an ensemble predictor composed of a collection of M randomized regression trees. For the j-th tree within this ensemble, the predicted value at a query point X is denoted by $m_n(X;\Theta_j,\mathscr{D}_n)$. Here, $\Theta_1,\ldots,\Theta_M$ are independent random variables, all drawn from the same distribution as a generic random variable Θ, and critically, they are independent of the training dataset $\mathscr{D}_n$. The variable Θ serves to introduce randomness, primarily by resampling the training set prior to the growth of individual trees (bootstrapping) and by randomly selecting features for successive splits at each node. In mathematical terms, the estimate from the j-th tree takes the following form:[56]

$$m_n(X;\Theta_j,\mathscr{D}_n) = \sum_{i\in\mathscr{D}_n^*(\Theta_j)} \frac{1_{X_i\in A_n(X;\Theta_j,\mathscr{D}_n)}Y_i}{N_n(X;\Theta_j,\mathscr{D}_n)}, \tag{3.8}$$

where, $m_n(\vec{X};\Theta_j,\mathscr{D}_n)$ represents the predicted value for a given query point $\vec{X}$ generated by the j-th regression tree within the Random Forest ensemble. The summation runs over all data points i belonging to the bootstrapped training subset $\mathscr{D}_n^*(\Theta_j)$, which is the specific resampled dataset used to train the j-th tree, with Θ_j encapsulating the random choices made during its construction. The term $1_{\vec{X}_i\in A_n(\vec{X};\Theta_j,\mathscr{D}_n)}$ is an indicator function that equals 1 if the query point $\vec{X}$ falls into the same terminal leaf node region, denoted as $A_n(\vec{X};\Theta_j,\mathscr{D}_n)$, as the training data point $\vec{X}_i$, and is 0 otherwise. Y_i is the true target value for the i-th training data point. Finally, $N_n(\vec{X};\Theta_j,\mathscr{D}_n)$ represents the total number of training data points from the bootstrapped sample that reside within that same leaf node region A_n. Essentially, the prediction of an individual tree for a new input is the average of the target values of the training samples that reside in the same final leaf node.[56]

And the trees are combined to form the finite forest estimate:

$$m_{M,n}(x;\Theta_1,\ldots,\Theta_M,\mathscr{D}_n) = \frac{1}{M}\sum_{j=1}^{M} m_n(x;\Theta_j,\mathscr{D}_n). \tag{3.9}$$

Since M may be chosen arbitraily large, limited only by the available computational resources, it makes sense, from a moddling point of view, to let M tend to infinity, and consider instead of Equation 3.9 the infinite forest estimate:

$$m_{\infty,n}(x;\mathscr{D}_n) = \mathbb{E}_\Theta[m_n(x;\Theta,\mathscr{D}_n)]. \tag{3.10}$$

Equation 3.10 introduces $m_{\infty,n}(\vec{x};\mathscr{D}_n)$, which signifies the theoretical prediction of an infinitely large Random Forest ensemble for the query point $\vec{x}$, given the training dataset $\mathscr{D}_n$.

This ideal prediction is formally expressed as $\mathbb{E}_{\Theta}[m_n(\vec{x};\Theta,\mathscr{D}_n)]$, representing the expectation (or average) of the predictions of all possible individual trees that could be generated under the randomization mechanism defined by Θ.

3.2.3 eXtreme Gradient Boosting

3.2.3.1 Regulazied Learning Objective

For a given dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$ comprising n examples, where each input $x_i \in \mathbb{R}^m$ has m features and $y_i \in \mathbb{R}$ is the corresponding target output, a tree ensemble model utilizes K additive functions to predict the output y, as shown in the previous section:

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i), \quad f_k \in \mathscr{F}, \tag{3.11}$$

where $\hat{y}_i$ represents the predicted output value for the i-th input example. This prediction is generated by the ensemble model's overall function $\phi(\vec{x}_i)$, where $\vec{x}_i$ is the feature vector for that i-th input. The prediction is constructed as a sum over K individual functions, with K being the total number of trees in the ensemble. Each $f_k(\vec{x}_i)$ denotes the prediction provided by the k-th individual tree for the input $\vec{x}_i$. The notation $f_k \in \mathscr{F}$ signifies that each f_k is a function belonging to the predefined space of regression trees or base functions.

With $\mathscr{F} = \left\{f(\mathbf{x}) = w_{q(\mathbf{x})}\right\} \left(q : \mathbb{R}^m \to T, w \in \mathbb{R}^T\right)$ being the space of regression trees, also known as CART[55], here q represents the structure of each individual tree, which maps an input example $\vec{x}$ to its corresponding leaf index. T is the total number of leaves in that particular tree. Each function f_k in the ensemble corresponds to an independent tree structure q and a set of associated leaf weights w.

Unlike a common decision tree, each regression tree in an ensemble model yields a continuous score at each of its leaves. Using w_i to represent the score on the i-th leaf, for a given example, the decision tree applies rules defined by q to classify it into a specific leaf and then calculates the final prediction by summing up the scores in the corresponding leaves, given by w. The function used to minimize the error during the training process is given by the following regularized objective:[57]

$$\mathscr{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \tag{3.12}$$

$$\text{with } \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2. \tag{3.13}$$

In this formulation, $\mathscr{L}(\phi)$ denotes the overall regularized objective function that the ensemble model aims to minimize during its training process. The first summation, $\sum_i l(\hat{y}_i, y_i)$, represents the training loss, which quantifies the discrepancy between the model's predicted output $\hat{y}_i$ for each training example i and its corresponding true target value y_i. The second summation, $\sum_k \Omega(f_k)$, constitutes the regularization component, where $\Omega(f_k)$ is the regularization term applied to each individual tree or base learner f_k in the ensemble.[52]

The regularization term for a single tree, $\Omega(f)$, is further defined by two parts. γ is a regularization parameter that penalizes the complexity related to the number of leaves in the tree, with T representing the actual number of leaf nodes in that tree. The term $\frac{1}{2}\lambda\|w\|^2$ is an L2 regularization on the leaf weights, where λ is a regularization parameter controlling the magnitude of this penalty, and $\|w\|^2$ is the squared L2 norm of the vector of leaf weights w, effectively penalizing large scores in the leaves.[52]

3.2.3.2 Gradient Boosting

Equation 3.12 includes functions as parameters, and thus cannot be optimized using traditional optimization methods in Euclidean space.[57] Instead, the model is trained by an additive strategy, this approach iteratively builds the ensemble, where $\hat{y}_i^{(t)}$ represents the prediction for the i-th instance at the t-th iteration. To refine this prediction, a new tree, f_t, is added to the ensemble:

$$\mathscr{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)\right) + \Omega(f_t), \tag{3.14}$$

where $\mathscr{L}^{(t)}$ represents the objective function that is being optimized at the current t-th iteration of the additive training process. The summation $\sum_{i=1}^{n}$ is performed over all n training instances. The term $l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\vec{x}_i)\right)$ is the loss function, which quantifies the error for the i-th instance, where y_i is its true target value and $\hat{y}_i^{(t-1)} + f_t(\vec{x}_i)$ is the current total prediction. This total prediction is composed of $\hat{y}_i^{(t-1)}$, the cumulative prediction for the i-th instance from all trees trained in previous iterations, and $f_t(\vec{x}_i)$, the prediction contributed by the new tree being added at the current iteration for the i-th instance with feature vector $\vec{x}_i$. Finally, $\Omega(f_t)$ is the regularization term applied to this new tree f_t, penalizing its complexity to prevent overfitting.

This approach greedily adds the new tree f_t that most significantly improves the model's performance, as evaluated by Equation 3.12. To optimize this objective function efficiently, a second-order Taylor approximation can be employed, which is given by:[58]

$$\mathscr{L}^{(t)} \approx \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t) \tag{3.15}$$

where $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ are the first and second-order gradient statistics, respectively, of the loss function with respect to the prediction from the previous iteration. By removing the constants terms, it is possible to rewrite the objective function at step t as:

$$\tilde{\mathscr{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t) \tag{3.16}$$

Defining $I_j = \{i | q(\mathbf{x}_i) = j\}$ as the instance set that fall into leaf j, Equation 3.16 can be rewritten by expanding the regularization term $\Omega(f_t)$:

$$\begin{aligned} \tilde{\mathscr{L}}^{(t)} &= \sum_{i=1}^{n} \left[g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \end{aligned} \tag{3.17}$$

In these equations, $\tilde{\mathscr{L}}^{(t)}$ represents the approximated objective function at the current iteration t, which is minimized to determine the optimal new tree f_t. The initial summation $\sum_{i=1}^{n}$ is over all n training instances. Within this sum, g_i denotes the first-order gradient of the loss function with respect to the previous prediction for instance i, and h_i is the corresponding second-order gradient (Hessian). $f_t(\vec{x}_i)$ is the prediction (or score) that the new tree f_t provides for the i-th instance. The terms γT and $\frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ constitute the regularization, where γ and λ are regularization parameters, T is the total number of leaf nodes in the tree f_t, and w_j is the weight (score) assigned to leaf node j.

The second form of the equation is a re-arrangement of the first, where terms are re-grouped by their respective leaf nodes. Here, I_j represents the set of all instance indices i that fall into leaf node j of the tree. Consequently, $\left(\sum_{i \in I_j} g_i\right)$ is the sum of the first-order gradients for all instances in leaf j, and $\left(\sum_{i \in I_j} h_i\right)$ is the sum of the second-order gradients for those same instances.[57]

Then, for a fixed tree structure $q(\vec{x})$, the optimal weight w_j^* for leaf j can be computed by:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}, \tag{3.18}$$

where, w_j^* denotes the optimal weight or score that is assigned to leaf node j of the tree. The numerator, $\sum_{i \in I_j} g_i$, represents the sum of the first-order gradients of the loss function for all training instances i that belong to the instance set I_j for leaf j. The denominator, $\sum_{i \in I_j} h_i + \lambda$, consists of the sum of the second-order gradients for all instances in leaf j, augmented by λ, which is a regularization parameter. This λ term stabilizes the denominator and helps to prevent overfitting by penalizing large leaf weights.[57]

And consequently, the optimal value of the objective function for a fixed tree structure can be expressed as:

$$\tilde{\mathscr{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T. \tag{3.19}$$

Equation 3.19 provides a scoring function to measure the quality of a proposed tree structure q within the XGBoost algorithm. Since it is generally impractical to enumerate all possible tree structures, a greedy algorithm is employed instead. This algorithm starts from a single leaf node and iteratively adds branches to the tree. Assuming I_L and I_R are the instance sets of the left and right nodes, respectively, after a split, and letting $I = I_L \cup I_R$ be the instance set of the parent node, the reduction in loss after this split is given by:[57]

$$\mathscr{L}_{\text{split}} = \frac{1}{2} \left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \tag{3.20}$$

3.2.4 Light Gradient Boosting Machine

Gradient Boosting Decision Trees (GBDTs), while effective, suffer from inefficiencies in terms of scalability and perform unsatisfactorily when faced with high-dimensional feature spaces and large datasets.[59] To address these limitations, the Light Gradient Boosting Machine (LightGBM) was developed, which applies two principal techniques: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).[59]

3.2.4.1 Gradient-based One-Side Sampling

Given a training set of n instances, $\{(x_1, \ldots, x_n)\}$, where each x_i is a vector of dimension s in feature space $\mathscr{X}^s$, in each iteration of gradient boosting, the negative gradients of the loss

function with respect to the current output of the model are denoted as $\{g_1, \ldots, g_n\}$. Decision tree models typically split each node based on the most informative feature. For GBDT, the information gain is usually measured by the variance after splitting, which is defined as:[59]

$$V_{j|\mathscr{O}}(d) = \frac{1}{n_{\mathscr{O}}} \left(\frac{\left(\sum_{\{x_i \in \mathscr{O} : x_{ij} \leq d\}} g_i \right)^2}{n^j_{l|\mathscr{O}}(d)} + \frac{\left(\sum_{\{x_i \in \mathscr{O} : x_{ij} > d\}} g_i \right)^2}{n^j_{r|\mathscr{O}}(d)} \right), \tag{3.21}$$

GOSS method proceeds in a multi-step manner. First, it ranks all training instances according to the absolute values of their negative gradients in descending order. Second, it retains the top $a \times 100\%$ instances that possess the largest gradients, forming an instance subset denoted as A. Subsequently, from the remaining set A^c, which comprises $(1-a) \times 100\%$ instances with smaller gradients, a subset B of size $b \times |A^c|$ is randomly sampled. Finally, the decision tree is split based on the estimated variance gain $V_j(d)$ over the combined subset $A \cup B$, which is defined as:[59]

$$\tilde{V}_j(d) = \frac{1}{n} \left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b} \sum_{x_i \in B_l} g_i \right)^2}{n^j_l(d)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b} \sum_{x_i \in B_r} g_i \right)^2}{n^j_r(d)} \right), \tag{3.22}$$

where, $\tilde{V}_j(d)$ quantifies the estimated variance gain for splitting on feature j at a threshold d, utilizing the GOSS method. The term n represents the total number of instances in the original training set, g_i denotes the negative gradient associated with instance x_i, the parameters a and b are hyperparameters of the GOSS algorithm: a defines the proportion of instances with large gradients included in subset A, and b defines the sampling ratio for instances with small gradients included in subset B.

A_l and A_r represent the subsets of instances from A that go to the left and right child nodes, respectively, after the split, similarly, B_l and B_r are the subsets of instances from B that are assigned to the left and right child nodes. $n^j_l(d)$ and $n^j_r(d)$ are the number of instances falling into the left and right child nodes, respectively, for the given split on feature j at threshold d the factor $\frac{1-a}{b}$ is a scaling weight applied to the gradients of instances in subset B to compensate for their undersampling, ensuring their contributions are appropriately represented in the gain calculation.

3.2.4.2 Exclusive Feature Bundling

High-dimensional datasets are frequently characterized by significant sparsity, a property that enables the design of a nearly lossless approach to dimensionality reduction. In such datasets, many features are often mutually exclusive, meaning they rarely, if ever, take on non-zero values simultaneously for the same instance. This mutual exclusivity allows for the safe bundling of these exclusive features into a single, composite feature, referred to as an exclusive feature bundle. A carefully designed feature scanning algorithm can be employed to systematically identify these exclusive features and subsequently group them into such bundles. The underlying theorems and their proofs supporting this methodology are further discussed by Guolin et al.[59]

3.3 Molecular Fingerprints

Molecular fingerprints encompass various types, each based on a specific method of transforming a molecular representation into a bit string or a vector. While most methods primarily utilize only a 2D molecular graph, others are capable of encoding 3D structural information. The most notable categories of fingerprints include substructure keys-based fingerprints, topological or path-based fingerprints, and circular fingerprints.[15]

3.3.1 Substructure Keys-Based Fingerprints

Substructure keys-based fingerprints operate by setting specific bits within a bit string to indicate the presence or absence of predefined substructures or features from a given list of structural keys within a compound. These types of fingerprints are particularly useful when applied to molecules whose structural features are extensively covered by the established set of keys. Conversely, their utility diminishes for molecules containing novel substructures not represented in the key list. The length of the resulting bit string is determined directly by the total number of keys in the predefined list, with each bit corresponding to the presence (1) or absence (0) of its respective structural key in the molecule. Prominent examples of such fingerprints include MACCS keys[60] and PubChem fingerprints[61].[15]

3.3.2 Topological or Path-Based Fingerprints

Topological fingerprints operate by systematically enumerating all molecular fragments that follow a path up to a predetermined number of bonds. Each of these unique paths is then hashed to generate the final fingerprint, this methodology ensures that any given molecule

can produce a meaningful fingerprint, and the length of the resulting bit string can be adjusted as needed.

These fingerprints are also highly effective for rapid substructure searching and filtering applications, a characteristic consequence of using hashed fingerprints, however, is that a single 'on' bit in the string cannot be uniquely traced back to a single specific molecular feature; rather, it may correspond to multiple different paths or fragments due to collisions in the hashing process.

3.3.3 Circular Fingerprints

Circular fingerprints are a class of hashed topological fingerprints that distinguish themselves by recording the environment of each atom up to a determined radius, rather than enumerating linear paths. A prominent example is Extended-Connectivity Fingerprints (ECFPs), which are based on the Morgan algorithm[62] and specifically designed for applications in structure-activity modeling[63]. These fingerprints represent circular atom neighborhoods and can produce bit strings of variable length. They are most commonly employed with a diameter of 4 bonds (meaning a radius of 2), in which case they are referred to as ECFP4, or with a diameter of 6 bonds (radius 3), referred to as ECFP6. It is notable that software packages providing these, such as RDKit[64], commonly refer to them simply as Morgan fingerprints (MF). Figure 3.3 illustrates the process of generating a circular fingerprint, specifically a Morgan Fingerprint (MF), for the 2,1,3-benzothiadiazole molecule.

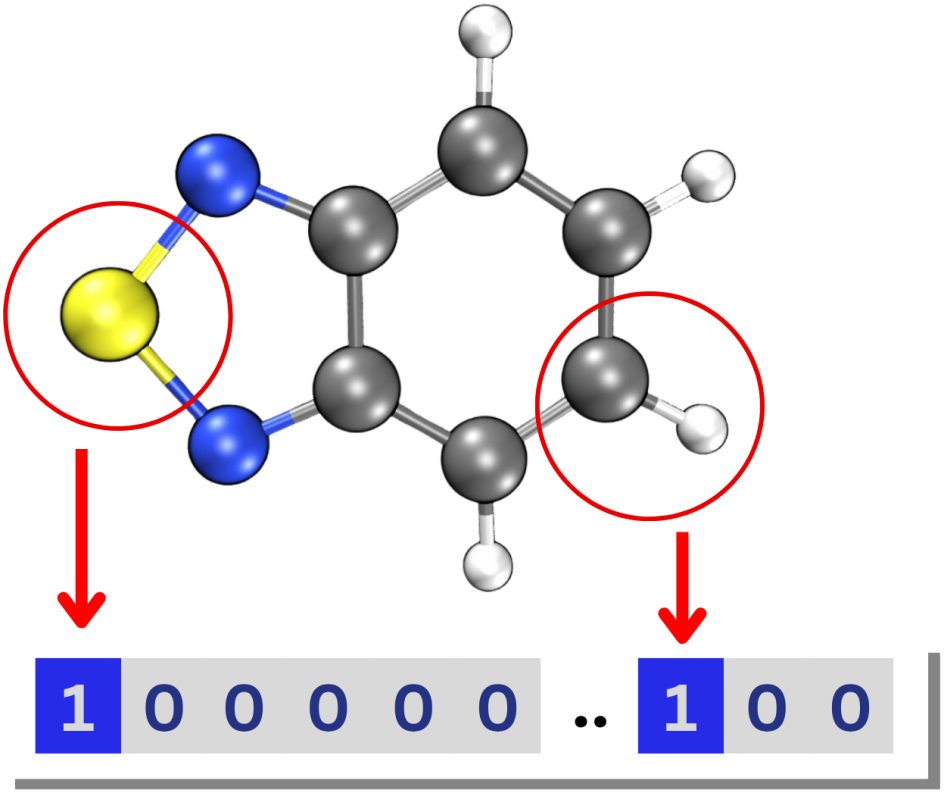


Figure 3.3: A representation of a hypothetical 2,1,3-benzothiadiazole molecule (BTD) and its corresponding Morgan Fingerprint (MF) with a radius of 2.

3.4 SHapley Additive exPlanations

Many current methods for interpreting individual machine learning models fall into the class of additive feature attribution methods.[65] This class encompasses techniques that explain a model's output as a sum of real values attributed to each input feature, as shown by the following general form:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{3.23}$$

where $g(z')$ represents the explanation model, which provides a simplified approximation of the original complex machine learning model's output for a given simplified input z'. The term ϕ_0 is a base value, often interpreted as the expected output of the model in a baseline or feature-absent scenario. The summation runs over M simplified input features, where z'_i denotes the simplified representation of the i-th feature, indicating its presence or absence. ϕ_i

is the attribution value assigned to the i-th simplified feature z'_i, quantifying its contribution to the model's prediction relative to the base value ϕ_0.

An important property of this class of additive feature attribution methods is that there exists a single, unique solution that satisfies three highly desirable properties: local accuracy, missingness, and consistency. Local accuracy dictates that the sum of the feature attributions must precisely equal the output of the function being explained for a particular instance. Missingness ensures that features already absent from the input (i.e., $z'_i = 0$ in the simplified input) are attributed no importance or contribution. Lastly, consistency states that if a model is altered such that a feature's impact on the model's output increases or stays the same, the attribution assigned to that specific feature will never decrease.[65]

To evaluate the effect of missing features on a model f, it is necessary to define a mapping h_x that translates a binary pattern representing missing or present features (denoted by z') to the original function's input space. This mapping enables the evaluation of $f(h_x(z'))$, thereby allowing for the calculation of the effect of either observing or not observing a specific feature (by setting $z'_i = 1$ or $z'_i = 0$, respectively).[66]

By defining $f_x(S) = f(h_x(z')) = \mathbb{E}[f(x|x_S)]$, where S represents the set of non-zero indices in the simplified binary input z', and $\mathbb{E}[f(x|x_S)]$ denotes the expected value of the function f conditioned on the subset x_S of observed input features, SHAP values combine these conditional expectations with the classic Shapley values from cooperative game theory to attribute ϕ_i values to each feature:[66]

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)], \tag{3.24}$$

where ϕ_i represents the SHAP value attributed to feature i, quantifying its contribution to the model's prediction for a specific instance. The summation is performed over all possible subsets S of features that do not include feature i (denoted as $N \setminus \{i\}$). The term $\frac{|S|!(M-|S|-1)!}{M!}$ is a weighting factor that accounts for the number of permutations in which features in S appear before feature i, and the remaining features (not in S and not i) appear after feature i, normalized by the total number of permutations of all M simplified features. The core of the contribution is captured by $[f_x(S \cup \{i\}) - f_x(S)]$, which represents the marginal impact of adding feature i to an existing subset of features S. $f_x(S \cup \{i\})$ is the model's output when both features in S and feature i are considered present, while $f_x(S)$ is the model's output when only features in S are present.

In a practical implementation, such as in the present work, an analogy can be drawn to a game of baseball: a "reward" or contribution value is assigned to each "player" (representing

a molecular substructure) based on their individual impact on the "result" of their "games" (i.e., the observed photophysical properties).

Chapter 4

Methodology

This work proposes the prediction of photophysical properties for 2,1,3-Benzothiadiazole derivatives, utilizing an experimentally curated dataset with decision tree-based machine learning models and SHAP-based interpretation of predictions. The first section details the construction of the dataset. The second section describes the techniques employed for training and validating the models. The final section outlines the methodology for model interpretation and deployment.

4.1 Dataset and Molecular Descriptors

A comprehensive dataset is crucial for the development of predictive models in cheminformatics, as it provides the necessary information for training and validating machine learning algorithms. A dataset comprising 701 BTD derivatives was assembled, with data obtained from the Elsevier database, using its API in conjunction with web scraping techniques implemented in a Python 3.x environment. The dataset includes experimental values reported in the literature, specifically the maximum absorption wavelength (λ_{max}^{abs}) and maximum emission wavelength (λ_{max}^{em}), along with the solvents used in the measurements. Furthermore, for each solvent, the respective normalized molar electronic transition energies (E_T^N) were collected from the literature and incorporated into the dataset.[30]

It is noteworthy that the assembled dataset exhibited an imbalance in the distribution of solvents used, which may have introduced biases into the results. This imbalance could potentially affect the generalizability of the findings, particularly when predicting molecular properties in diverse solvent environments. Furthermore, variations in solvent polarity, hydrogen bonding capabilities, and normalized molar electronic transition energies could contribute to inconsistencies in the model's performance.

To numerically represent the chemical structures of the BTD derivatives, each compound was associated with its Simplified Molecular Input Line Entry System (SMILES)[67] representation, which serves as a standardized format for encoding molecular structures. The SMILES strings were then converted into circular descriptors using the RDKit library. As previously noted, within the RDKit[64] library, these descriptors utilize the Morgan algorithm and so will be referred to as Morgan fingerprints (MF) from this point forward. MF were

generated as binary vectors to represent the presence or absence of specific substructures within the BTD derivatives.[62]

A scan of the fingerprint length and radius was performed with 512, 1024, 2048, 3072, 4096, and 8192 bits, and 1 through 6 radii to determine the optimal parameters for the Morgan fingerprints. The optimal parameters were selected based on the performance of the models trained with these fingerprints, ensuring that the descriptors effectively captured the relevant chemical information necessary for accurate predictions of the molecular properties of interest.

4.2 Models and Validation

In light of the recent success of decision tree-based machine learning algorithms applied in predictive chemistry[16,68], decision tree-based models were selected for the development of predictive models in this study. Specifically, three tree ensemble models were chosen: XGBoost[57], Random Forest[56], and LightGBM[59]. The models were developed using the scikit-learn, XGBoost, and LightGBM libraries in Python 3.x. Hyperparameter optimizations were conducted using the Optuna library[69] to fine-tune the model parameters.

To compare and estimate the performance of the chosen algorithms, the coefficient of determination (R^2), explained variance (Q^2), mean absolute error (MAE), and root mean squared error (RMSE) were used as metrics. A 10-fold cross-validation was performed using K-Fold. The data was split equally into subsets to maintain consistency, with nine subsets (630 BTD derivatives) used for training and one subset (71 BTD derivatives) reserved for testing. This process was repeated 10 times, with the final result being the mean of the outcomes from all ten tests.

External testing was also performed using recent experimental data from 20 BTD derivatives not included in the training set. These new data points provided an independent validation of the model's predictive capabilities. The predicted λ_{max}^{abs} and λ_{max}^{em} values were compared with the experimental values, and the relative errors were calculated for each molecule to assess prediction accuracy.

To further assess the robustness of the models and ensure that their predictive performance was not due to overfitting, a y-randomization test was conducted. This involved randomly shuffling the target variable (λ_{max}^{abs} and λ_{max}^{em}) while keeping the features intact, and then re-evaluating the model performance. The results of this test were compared to the original model performance to confirm that the models were not simply memorizing the training data.

Residual plots were generated to analyze the differences between experimental and predicted values against predicted values. A random scatter of residuals around zero indicates that the model assumptions are valid and that the model provides an appropriate fit to the data.

4.3 Model Interpretation and Deployment

An essential requirement in developing predictive models is the mechanistic interpretation of the model's predictions. Understanding how the model arrives at its predictions provides valuable chemical insights and improves confidence in its reliability. To achieve this, SHapley Additive exPlanations (SHAP) was employed as a unified approach for interpreting model predictions. Based on Section 3.4, SHAP computes the average marginal contribution of each feature i across all possible permutations of the feature set. This involves evaluating the change in the model's output when feature i is added independently to various subsets of features. The significance of each feature is directly reflected in its impact on the variation of the model's output.

Using MF as molecular descriptors, SHAP values provide a solution to assign a quantitative measure of importance to each molecular substructure in predicting λ_{max}^{abs} and λ_{max}^{em} wavelengths. This approach connects the need to interpret the model's predictions with understanding the photophysical properties of BTD derivatives, as it helps identify which molecular features most significantly influence these properties.

A web application was deployed on a cloud-based platform using the Streamlit environment to provide a responsive and user-friendly interface. Machine learning predictions are integrated into the workflow, allowing the rapid evaluation of BTD derivatives. The application uses the Ketcher molecular editor, an open-source web-based tool for drawing and editing chemical structures, to facilitate the input of new BTD derivatives. Prediction results are displayed on the screen for immediate review and can also be exported as spreadsheet files for further analysis algonside with the SHAP analysis of the given molecule showing wich MF are impacting the prediction. The application is designed to be intuitive, enabling users to easily input molecular structures and obtain predictions without requiring extensive technical knowledge. Figure 4.1 illustrates the whole workflow process.

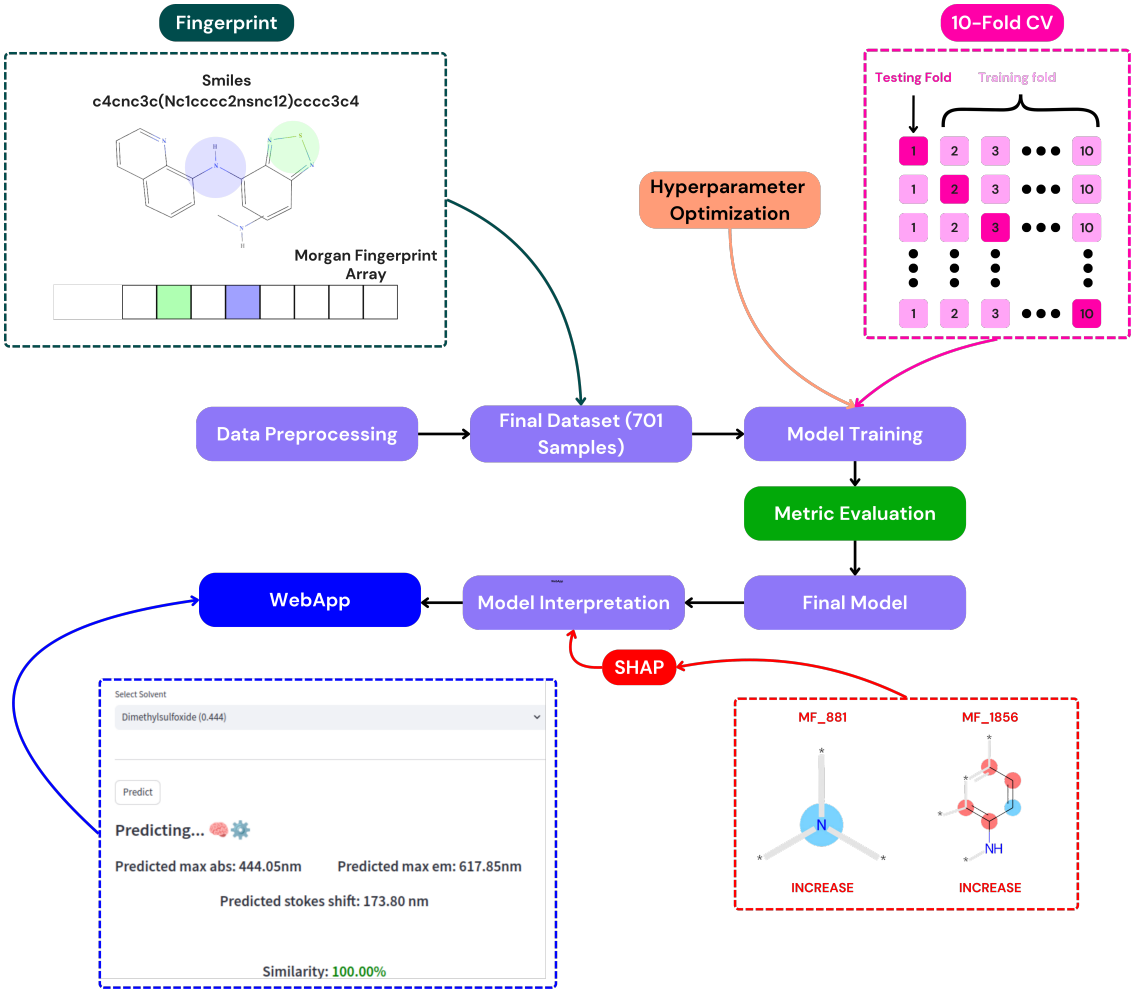


Figure 4.1: Representation of the model development workflow, illustrating key steps including data collection, preprocessing, fingerprint generation, hyperparameter optimization using Optuna, 10-fold cross-validation, model training with three algorithms (eXtreme Gradient Boosting, Random Forest, Light Gradient Boosting Machine), metric evaluation (R^2, Q^2, MAE, RMSE), SHAP-based model interpretation, and deployment as a web application for prediction and visualization of results.

Chapter 5

Results and Discussion

5.1 Validations

To define the optimal molecular descriptors, a systematic scan of the Morgan fingerprint length and radius was performed. Based on the model performance across the tested parameters, as shown in Figure 5.1, an optimal configuration of a 2048-bit vector and a radius of 4 was identified. Altough not the most performant, this configuration was selected for its balance between computational efficiency and predictive accuracy. The 2048-bit vector length provides a sufficiently detailed representation of the molecular structure, while the radius of 4 captures relevant local features without introducing excessive noise. These parameters were subsequently used for all fingerprint generation in the training and validation stages.

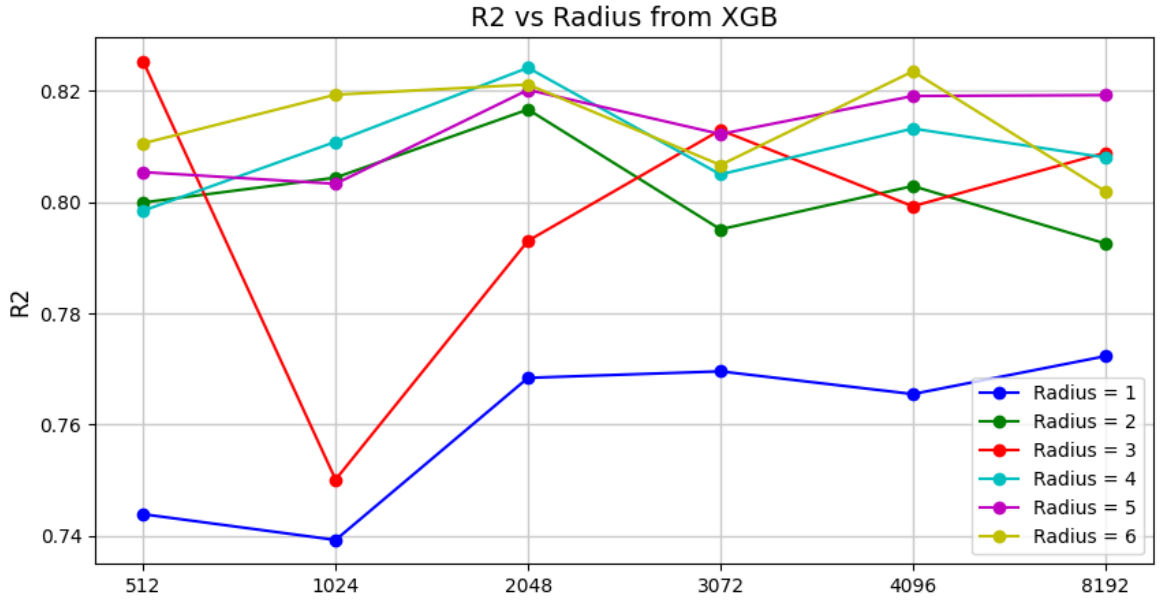


Figure 5.1: Scan of the lenght and radii of the MF to determine the optimal values for the descriptors. The optimal values were found to be 2048 for the lenght and 4 for the radii

Following descriptor selection, hyperparameter optimization was conducted for the XGBoost, RF, and LightGBM algorithms using the Optuna framework. This process determined

that the RF and LightGBM models yielded optimal performance with their default parameters. In contrast, the performance of the XGBoost model was significantly improved with a set of fine-tuned hyperparameters, which are detailed in Table 5.1.

Table 5.1: Fine-tuned hyperparameters obtained through Optuna optimization for XGBoost in predicting $\lambda_{\max}^{abs}$ and $\lambda_{\max}^{em}$. The hyperparameters are listed along with their optimal values for both properties.

Hyperparameter	$\lambda_{\max}^{abs}$	$\lambda_{\max}^{em}$
n_estimators	522	502
max_depth	8	7
learning_rate	0.0286	0.0331
subsample	0.8384	0.7929
colsample_bytree	0.7201	0.7203
gamma	0.0018	2.7×10^{-6}
reg_alpha	7.71×10^{-5}	2.15×10^{-6}
reg_lambda	0.0995	1.08×10^{-4}
min_child_weight	4	5

The internal performance of the models was evaluated using a 10-fold cross-validation procedure. Four statistical metrics were employed: the coefficient of determination (R^2), the cross-validated coefficient of determination (Q^2), the Mean Absolute Error (MAE), and the Root Mean Squared Error (RMSE). The average values for each metric across the ten folds are summarized in Table 5.2.

Photophysical Property	Algorithms	R^2	Q^2	MAE (nm)	RMSE (nm)
Max Absorption	XGB	0.8976±0.1075	0.8999±0.1039	7.2155±1.2992	14.9539±8.4224
	RF	0.9252±0.0400	0.9259±0.0401	6.8863±0.8979	12.0208±2.8983
	LGBM	0.8981±0.0377	0.8993±0.0373	8.0827±1.3499	14.3505±3.6646
Max Emission	XGB	0.8909±0.1054	0.8937±0.1023	13.3673±2.6616	21.2640±9.9791
	RF	0.8884±0.0300	0.8923±0.0280	14.8029±1.1918	21.4308±1.7766
	LGBM	0.8740±0.0508	0.8763±0.0488	14.5597±1.8141	21.7825±3.2040

Table 5.2: Performance comparison of three machine learning algorithms-eXtreme Gradient Boosting, RF and Light Gradient Boosting Machine-in predicting photophysical properties ($\lambda_{\max}^{abs}$ and $\lambda_{\max}^{em}$) using a 10 fold cross-validation. For each property the metrics, R^2, Q^2, MAE, and RMSE are reported along with their standard deviations. Each metrics is reported along with their standard deviation.

An analysis of the validation metrics reveals that the RF model provided the most reliable predictions for λ_{max}^{abs}, achieving the highest R^2 of 0.9252 and Q^2 of 0.9259, along with a comparatively low MAE and smaller standard deviations. For predicting λ_{max}^{em}, the XGBoost model showed slightly higher mean R^2 and Q^2 values. However, its performance was less consistent, as indicated by its larger standard deviations compared to the RF model. Given the close performance in accuracy and the superior predictive stability of the RF model, it was selected for all subsequent analyses and discussions.

The close agreement between the R^2 and Q^2 values for all models, with observed differences generally below 0.2, indicates a strong goodness-of-fit and a low risk of overfitting.[70] To further validate the models against spurious correlations, a y-randomization test was performed. This procedure involved retraining the models on datasets where the target property values ($\lambda_{\max}^{abs}$ and $\lambda_{\max}^{em}$) were randomly shuffled. As shown in Figures 5.2 and 5.3, this resulted in a significant drop in performance, confirming that the original models learned meaningful structure-property relationships rather than capitalizing on chance correlations in the data.

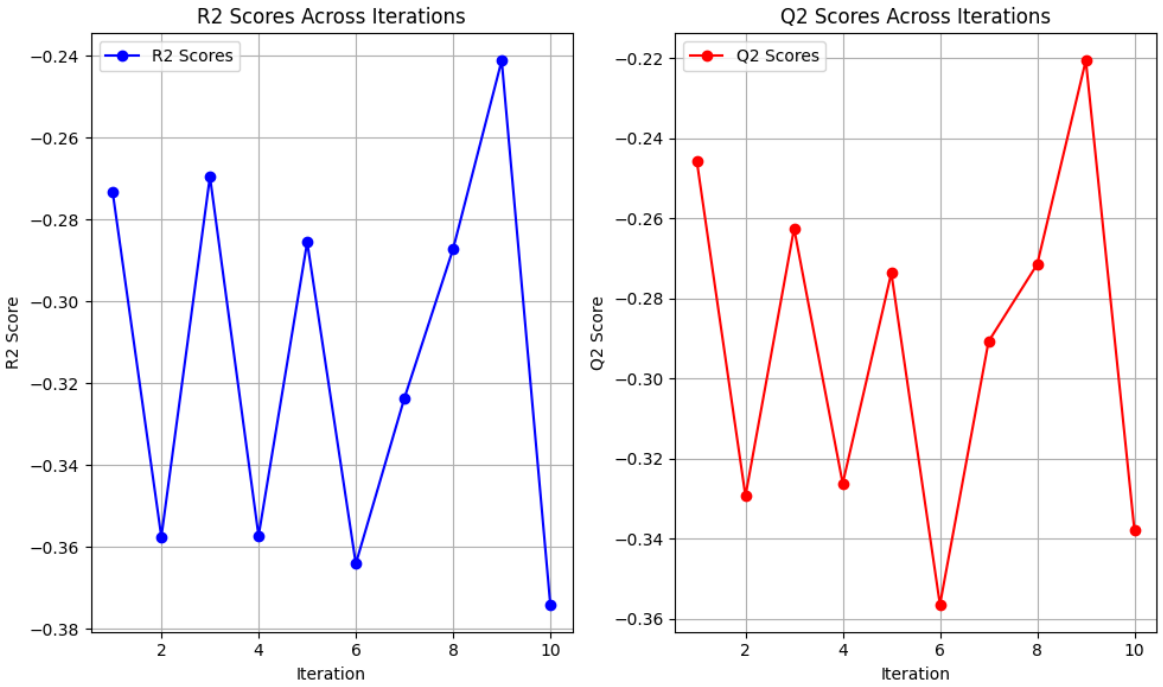


Figure 5.2: R^2 and Q^2 scores across 10 iterations for the y-randomization test for the RF model predicting λ_{max}^{abs}, demonstrating significantly worse performance metrics compared to the original model, thus validating its robustness and generalization capability.

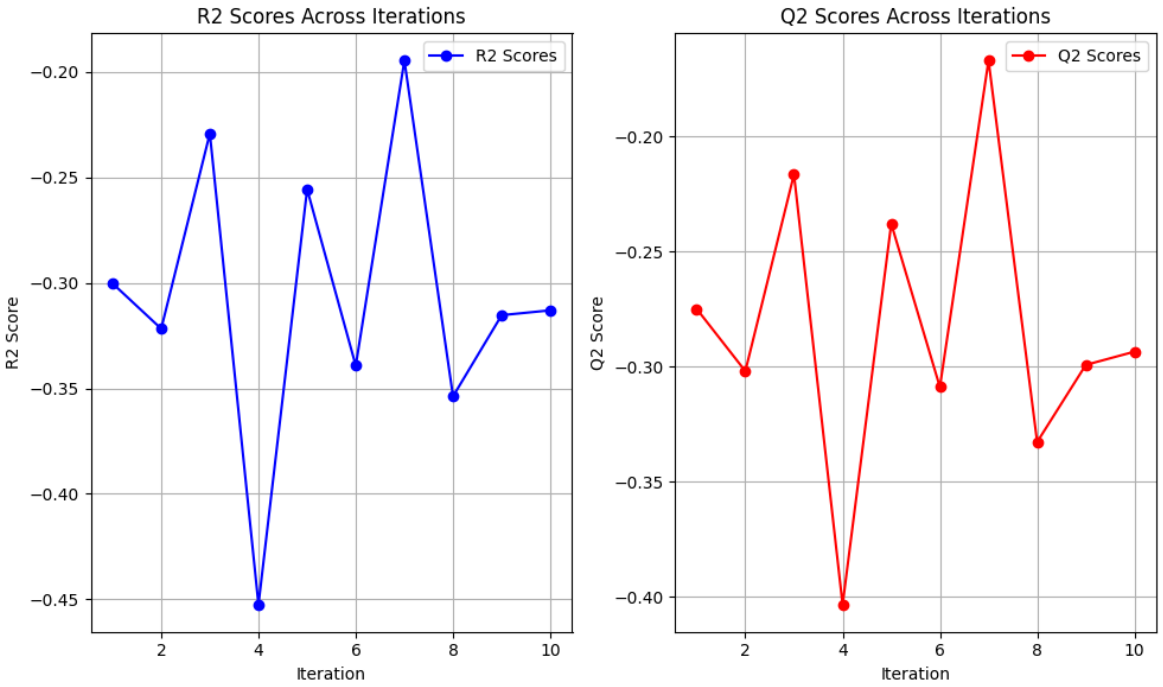
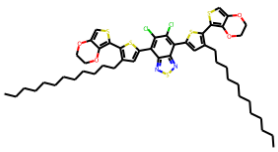
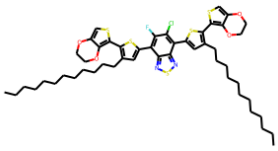


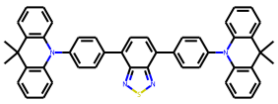
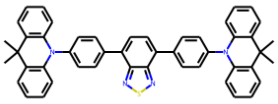
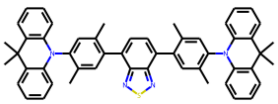
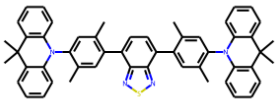
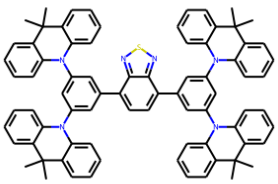
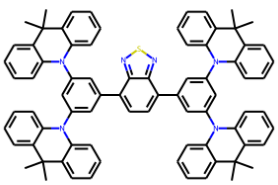
Figure 5.3: R^2 and Q^2 scores across 10 iterations for the y-randomization test for the RF model predicting λ_{max}^{em}, demonstrating significantly worse performance metrics compared to the original model, thus validating its robustness and generalization capability.

To assess the model's ability to generalize to new chemical matter, an external validation was performed using recently published experimental data for compounds not included in the original training set. Although this external set comprises less than 3% of the total database, its structural diversity provides a valuable and independent assessment of the model's predictive capabilities. Table A.1 presents the results of this test, comparing the experimental and predicted values for $\lambda_{\max}^{abs}$ and $\lambda_{\max}^{em}$. It also includes the prediction error and a similarity percentage, which indicates the structural similarity of each test compound to its nearest neighbor in the training set, helping to contextualize the model's performance within its applicability domain.

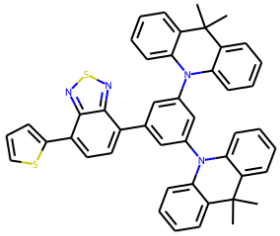
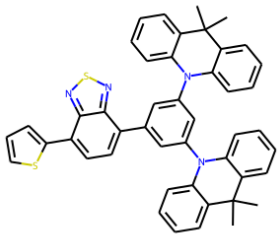
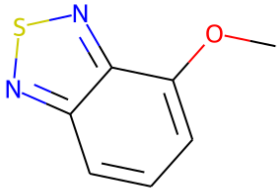
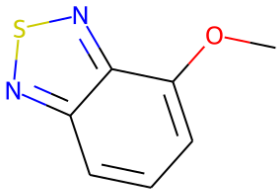
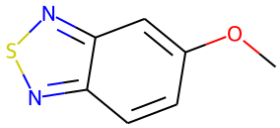
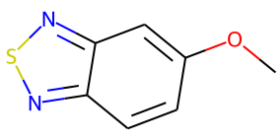
Table 5.3: External testing of the RF model in predicting maximum absorption and emission energies (eV). Experimental values (Exp) were measured in different solvents: [A] Tetrahydrofuran (THF), [B] Toluene (TOL), [C] Methanol (MeOH), and [D] Dichloromethane (DCM). Predicted values (Pred) represent the model outputs.

Molecules	Max Absorption (eV)			Max Emission (eV)			Similarity (%)
	Exp	Pred	Error	Exp	Pred	Error	
	3.10[A]	3.29	6.13%	2.41	2.50	3.73%	99.73
	3.40[A]	3.19	6.18%	2.61	2.44	6.51%	97.46

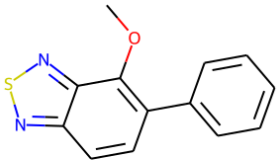
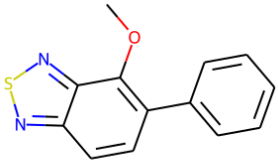
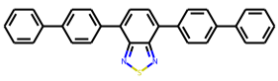
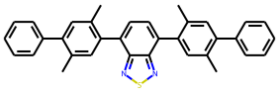
Continuation of Table A.1

Molecules	Max Absorption (eV)			Max Emission (eV)			Similarity (%)
	Exp	Pred	Error	Exp	Pred	Error	
	3.84[B]	3.24	15.6%	2.87	2.56	10.8%	74.43
	3.82[C]	3.23	15.4%	3.10	2.71	12.6%	74.43
	2.53[D]	2.82	11.5%	1.96	2.05	4.6%	72.51
	2.63[D]	2.72	3.42%	1.96	2.05	4.6%	69.80
	3.26[B]	3.21	1.53%	1.68	2.29	36.3%	66.90
	3.25[D]	3.20	1.54%	2.16	2.33	7.87%	66.90

Continuation of external validation

Molecules	Max Absorption (eV)			Max Emission (eV)			Similarity (%)
	Exp	Pred	Error	Exp	Pred	Error	
	2.97[B]	3.04	2.36%	2.03	2.27	11.8%	63.85
	2.98[D]	3.04	2.01%	1.62	2.18	34.6%	63.85
	3.47[B]	3.20	7.78%	1.70	2.28	34.1%	63.80
	3.48[D]	3.20	8.05%	2.30	2.28	0.87%	63.80
	3.51[B]	3.15	10.3%	2.56	2.19	14.5%	63.55
	3.51[C]	3.13	10.8%	2.33	2.00	14.2%	63.55

Continuation of external validation

Molecules	Max Absorption (eV)			Max Emission (eV)			Similarity (%)
	Exp	Pred	Error	Exp	Pred	Error	
	3.67[B]	3.11	15.3%	3.17	2.36	25.6%	56.94
	3.68[C]	3.11	15.5%	2.95	2.25	23.7%	56.94
	3.44[B]	3.02	12.2%	2.47	2.21	10.5%	50.17
	3.47[C]	3.04	12.4%	2.79	2.36	15.4%	50.17

The external validation results confirm a clear trend between structural similarity and predictive accuracy. Molecules with high similarity to the training set compounds generally yielded low prediction errors, while prediction errors tended to increase as structural similarity decreased. This demonstrates the importance of the model's applicability domain; molecules that are very dissimilar from the training data fall outside this reliable prediction range and produce larger errors.

However, similarity alone does not perfectly predict model performance. Some compounds with low similarity scores were still predicted with minimal error, suggesting the model successfully identified key photophysical features that transcend superficial structural differences. The analysis also indicates that emission properties are more sensitive to structural variations than absorption properties. Therefore, while the model shows strong promise for generalization, caution should always be exercised when interpreting predictions for molecules that are structurally distant from the training data.

The correlation between experimental and predicted values for both $\lambda_{\text{max}}^{abs}$ and $\lambda_{\text{max}}^{em}$ is visualized in Figure 5.4. On these plots, the x-axis represents experimental values, while the y-axis represents predicted values. The dashed line indicates a perfect prediction scenario, where experimental and predicted values are equal. The closer the points are to this line, the more accurate the predictions are. Ideally, the points should be distributed around the dashed line, indicating that the model is effectively capturing the relationship between the features and the target properties. As shown in Figure 5.4, RF predictions cluster closely around the diagonal, reinforcing the conclusion that the model offers a balanced combination of accuracy and consistency for forecasting the desired photophysical properties.

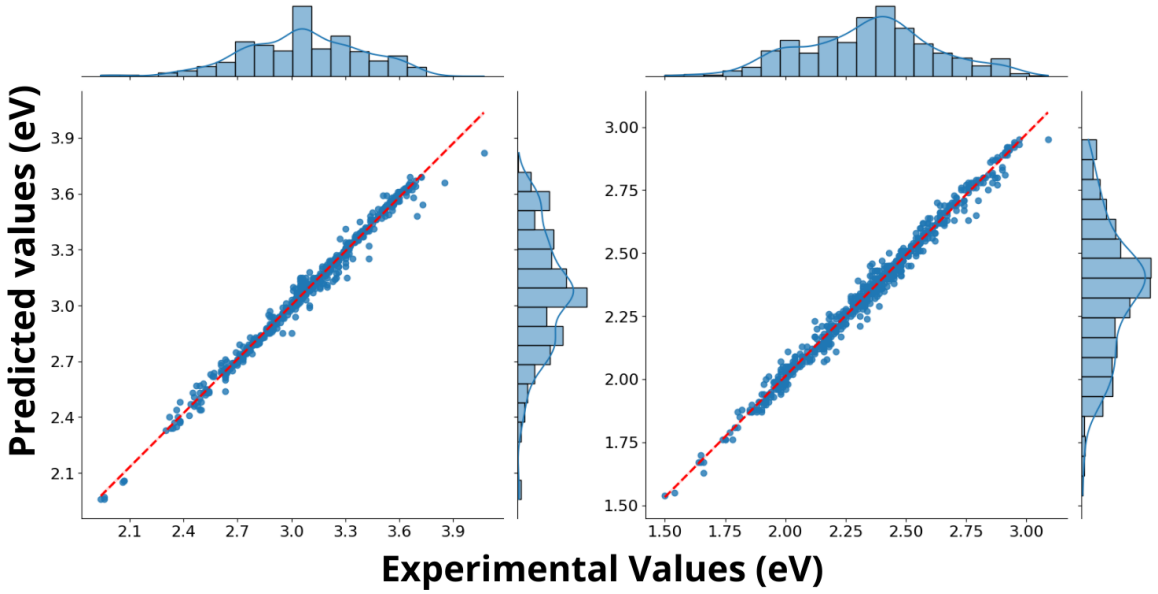


Figure 5.4: Predicted values versus experimental values for a) max absorption and b) max emission for RF, red line represents the ideal prediction and blue points the actual prediction.

Residual plots were analyzed to assess the model's performance and validate key statistical assumptions. Figures 5.5 and 5.6 display the residuals for the RF model predicting $\lambda_{\text{max}}^{abs}$

and $\lambda_{\max}^{em}$, respectively. For both properties, the residuals form a horizontal band centered around zero, with no strong "fan" or "cone" shape. This lack of a pattern suggests that the errors are homoscedastic, meaning their variance is constant across the range of predicted values.[71]

The random scatter also indicates that the model does not systematically over or under predict the target properties. While a few points stand out as potential outliers (notably near 470 nm for absorption and 600 nm for emission), the overall distribution suggests that the model is performing well, with no major violations of the assumptions required for regression analysis[71].

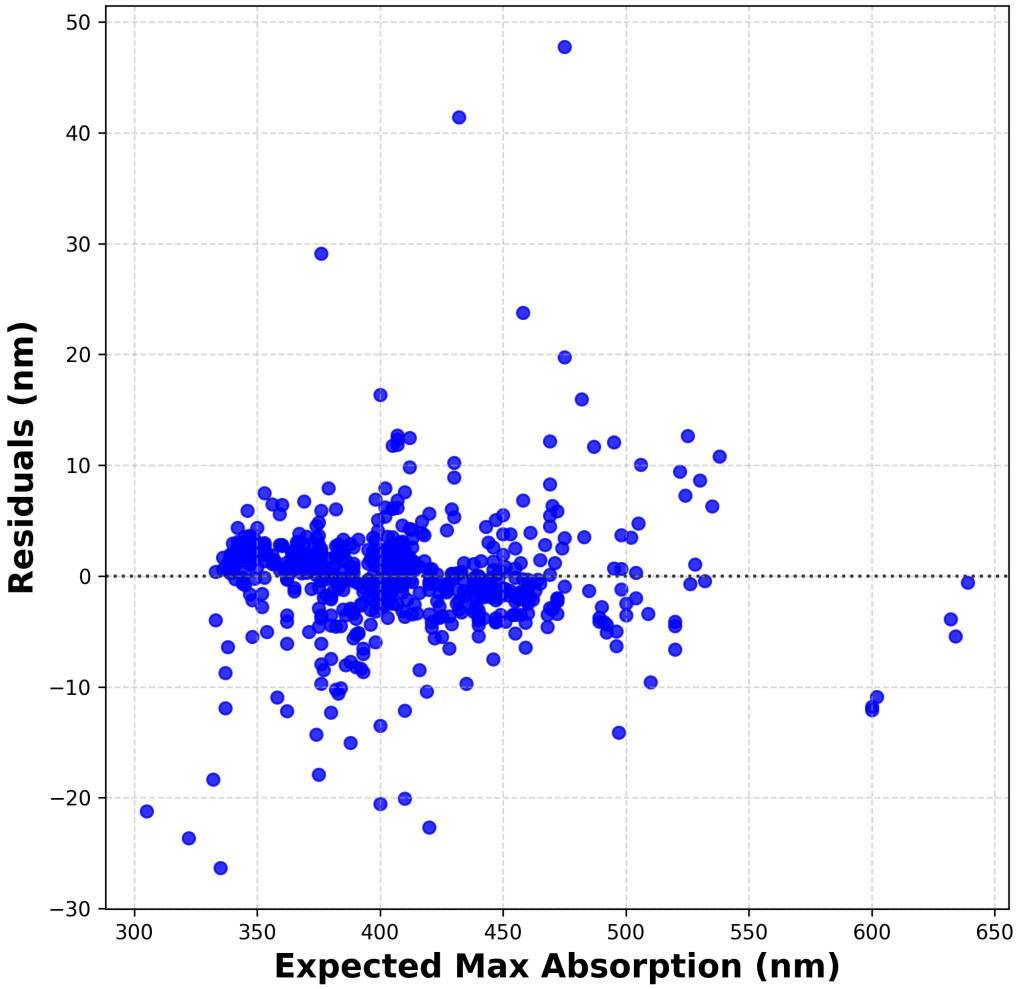


Figure 5.5: Residual plot showing the difference between predicted and expected values for the maximum absorption wavelengths (nm). The model captures the overall trends, with deviations primarily concentrated around the expected values.

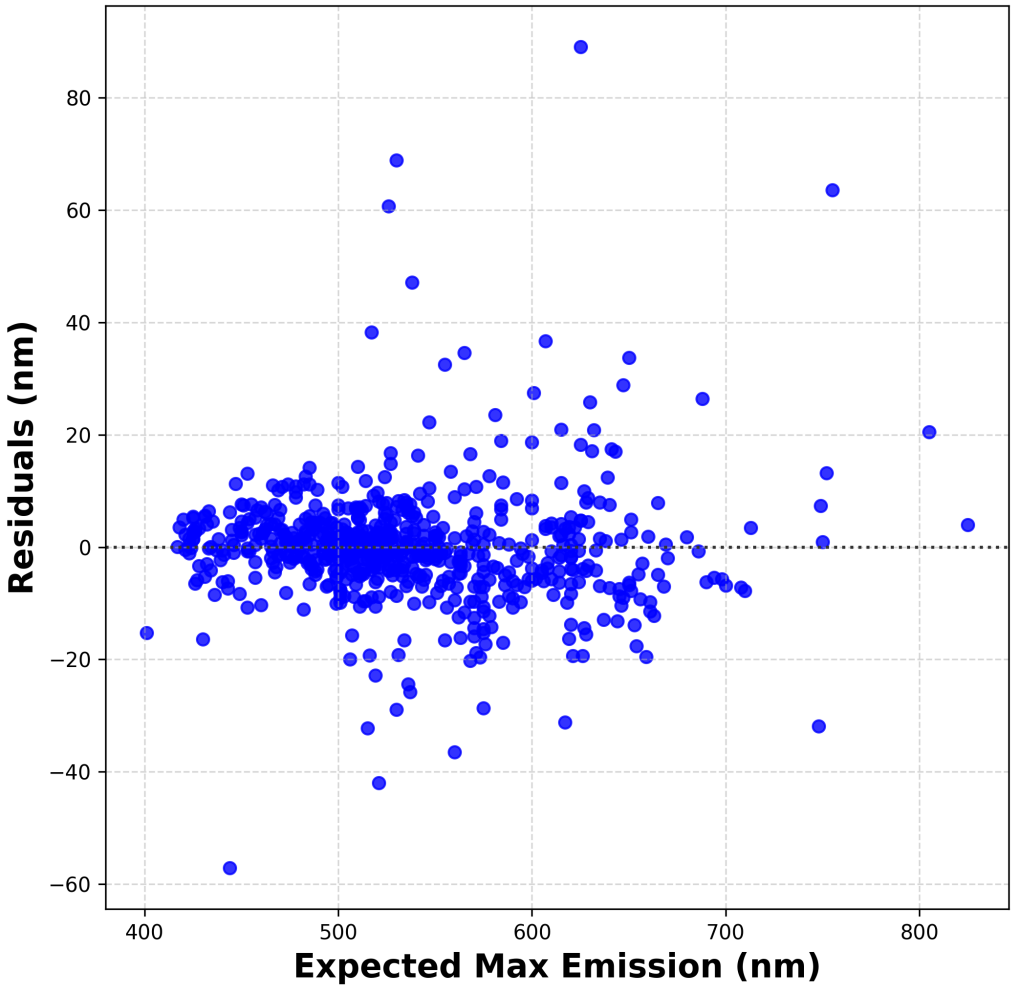


Figure 5.6: Residual plot showing the difference between predicted and expected values for the maximum emission wavelengths (nm). The model demonstrates good generalization, with residuals scattered closely around the expected values.

5.2 SHapley Additive exPlanations

To move beyond statistical performance metrics and interpret the model's decision-making process, SHAP was employed. This method quantifies the positive or negative contribution of each feature to the final predicted value for a given molecule.

The SHAP summary plots in Figures 5.7a and 5.8a illustrate the impact of the most influential features on the λ_{max}^{abs} and λ_{max}^{em} wavelengths, respectively. For λ_{max}^{abs}, the model's predictions arise from the combination of many small contributions, with no single feature showing a dominant effect. In stark contrast, the prediction of λ_{max}^{em} is overwhelmingly influenced by two specific features: the solvent polarity, represented by the E_T^N parameter, and the presence of a tertiary amine substructure (feature 881).

The SHAP analysis reveals that the two most influential features driving a bathochromic (red) shift in the emission wavelength are the presence of a tertiary amine and high solvent polarity (E_T^N). The SHAP dependence plots show that both features have a greater impact than variations in the BTD core itself (Figures 5.7b, 5.7c, 5.8b, 5.8c). This observation aligns with previous studies showing that tertiary amines act as strong electron-donating groups, enhancing the ICT character of the excited state[72,73]. This creates an efficient push-pull system that stabilizes the more polar excited state, thereby reducing the $S_1 \rightarrow S_0$ energy gap and red-shifting the emission spectrum[74].

The SHAP summary plot for λ_{max}^{abs} (5.7b) shows that the SHAP values for absent substructures correctly cluster at zero, confirming they do not influence the model's predictions. In contrast, the analysis highlights specific features with significant impact. For example, feature 1947, which represents a tertiary amine attached to both a benzene and a thiophene ring, exerts a strong positive influence on the predicted λ_{max}^{abs}.

This particular substructure functions as a potent electron-donating group (EDG), creating a classic Donor-π-Acceptor (D-π-A) system with the electron-deficient benzothiadiazole core. The amine's strong donating capacity enhances the ICT character of the primary $S_0 \rightarrow S_1$ electronic transition. This increased ICT lowers the energy of the excited state, thereby reducing the transition energy gap and causing a significant bathochromic (red) shift in the absorption spectrum.

A higher E_T^N value, indicative of a more polar solvent, leads to a bathochromic (red) shift in the emission spectrum. This finding is consistent with the well-established principles of solvatochromism, particularly for push-pull systems like BTD derivatives[45]. Polar solvents stabilize the more polar excited state to a greater extent than the ground state, which reduces the energy gap between the first excited state (S_1) and the ground state (S_0), thus causing the emission to be red-shifted.

This process begins with a vertical transition to a Franck-Condon excited state, which initially retains the ground-state's solvation shell[30]. Subsequently, the surrounding solvent molecules reorient and rearrange around the excited state's larger dipole moment, a relaxation process that further lowers its energy before emission occurs. The Lippert-Mataga[75,76]

equation, mathematically describes this phenomenon, relating the magnitude of the Stokes shift directly to the change in the molecule's dipole moment between the ground and excited states ($\Delta\mu = \mu_e - \mu_g$).

The model's ability to generalize this photophysical principle is confirmed by the SHAP analysis. The solvent polarity feature, $E_{\mathrm{T}}^{\mathrm{N}}$, shows a consistently positive SHAP value for all molecules in the dataset, indicating that the model has correctly learned that increasing solvent polarity leads to a bathochromic shift in the λ_{max}^{em}. This demonstrates that the model is not merely fitting to data but is capturing a fundamental chemical relationship.

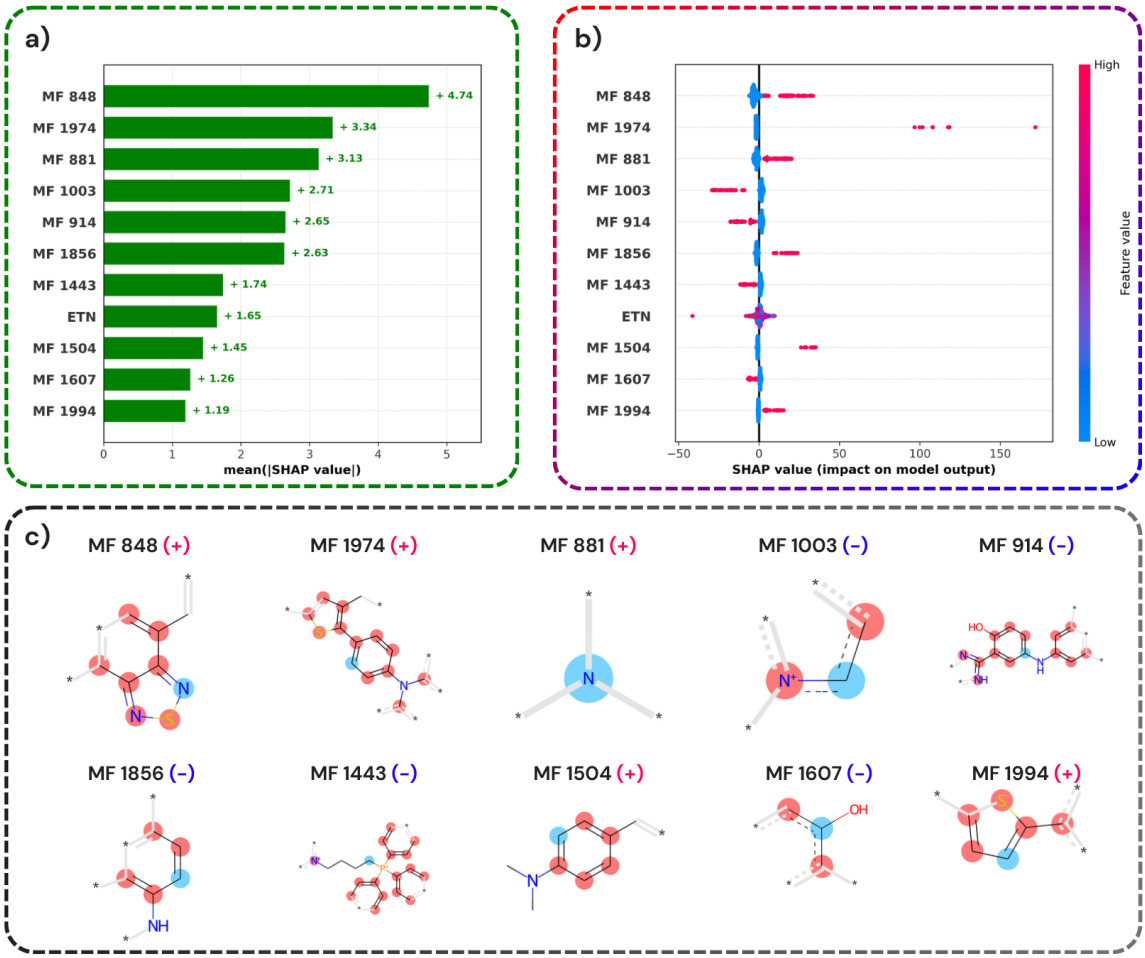


Figure 5.7: Visualization of the SHAP analysis for the maximum absorption wavelength (λ_{max}^{abs}) RF model in nm. (a) Feature importance plot showing the mean absolute SHAP value for the most impactful molecular features. (b) SHAP summary plot, where each point's position on the x-axis indicates its impact on the model output for a specific molecule. The color corresponds to the feature's value (red: high/present, blue: low/absent). (c) Chemical structures of selected substructures, illustrating examples that contribute positively (+) or negatively (-) to the predicted λ_{max}^{abs}.

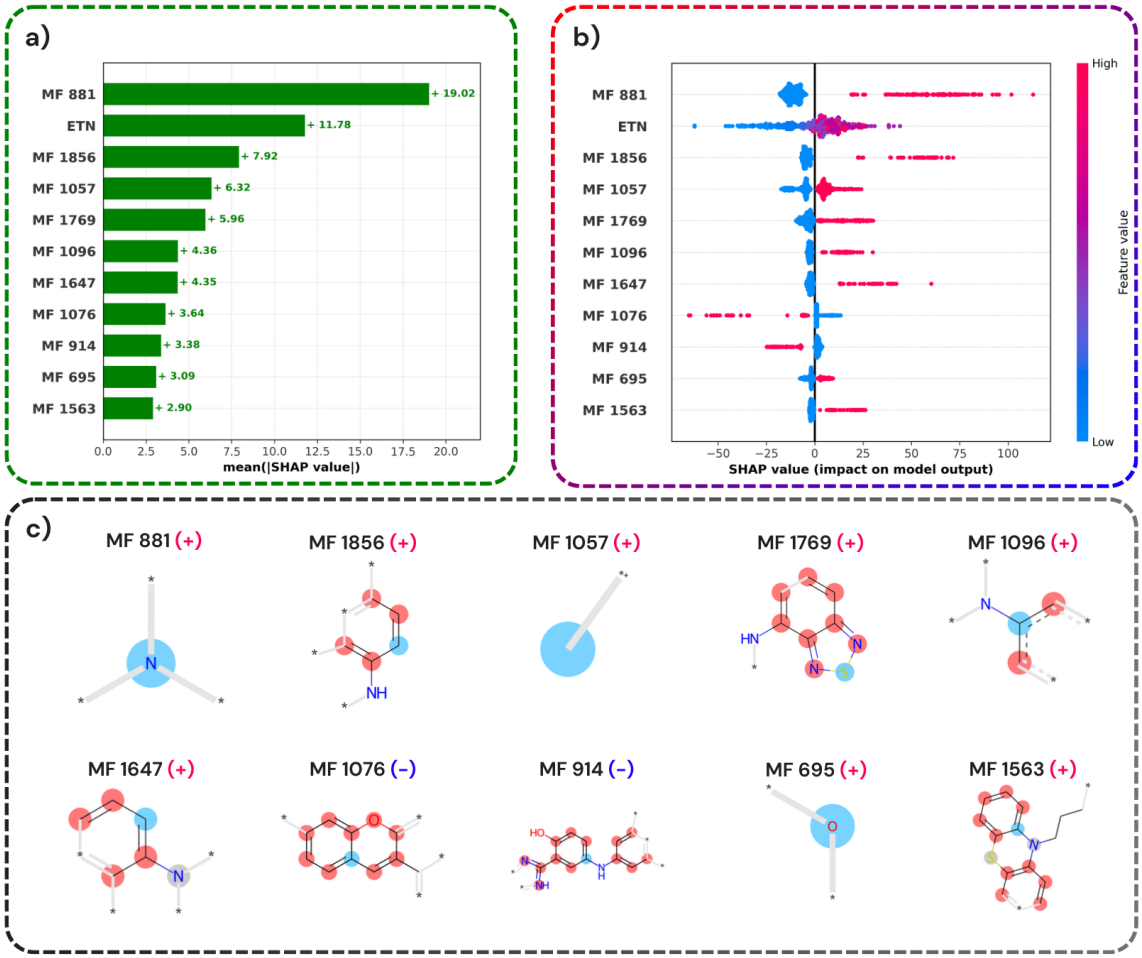


Figure 5.8: Visualization of the SHAP analysis for the maximum emission wavelength (λ_{max}^{em}) RF model in nm. (a) Feature importance plot showing the mean absolute SHAP value for the most impactful molecular features. (b) SHAP summary plot, where each point's position on the x-axis indicates its impact on the model output for a specific molecule. The color corresponds to the feature's value (red: high/present, blue: low/absent). (c) Chemical structures of selected substructures, illustrating examples that contribute positively (+) or negatively (-) to the predicted λ_{max}^{em}.

5.3 Web Application

A significant challenge in computational chemistry is translating predictive models into practical tools for the broader scientific community. The use of machine learning models often requires specialized expertise in programming and data science, creating an accessibility barrier for many experimental researchers. To address this gap, we developed a user-friendly web application to facilitate the use of the models presented in this work (Figure 5.9a).

The application provides two convenient methods for users to input the molecular structure of a BTD derivative. A user can either paste the molecule's SMILES string directly into the text field (Figure 5.9b) or draw the molecule using the integrated chemical editor, which automatically generates the corresponding SMILES string (Figure 5.9c). This flexibility ensures that the tool is accessible to all researchers, including those without prior knowledge of SMILES notation.

After providing the molecular structure, the user selects a solvent from a dropdown menu containing the options present in the original dataset. Once the query is submitted, the application returns the predicted λ_{max}^{abs} and λ_{max}^{em} wavelengths based on the selected machine learning model (Figure 5.9). To provide context for the prediction's reliability, the application also displays a similarity score comparing the input molecule to its nearest neighbor in the training set. This gives the user insight into the prediction's expected accuracy based on proximity to the model's learned chemical space. For record-keeping or further analysis, all prediction results can be downloaded in CSV format (Figure 5.9e) and as of recent all SHAP interpretations can be visualized alongside with its MF contribution making it easier to understand how the BTD substructures influence the predicted properties. To ensure scientific transparency, the web application provides full access to the underlying dataset used for model development. This allows users to explore the training data and better understand the chemical space on which the predictions are based.

By making these predictive models accessible through an intuitive interface, this work aims to lower the barrier to entry for applying machine learning in photochemistry. It enables a broader community of chemists and materials scientists to leverage advanced computational tools, potentially accelerating the design and discovery of novel fluorophores. Ultimately, this application serves as a bridge between specialized computational modeling and practical experimental research, fostering innovation and collaboration across disciplines.[20]

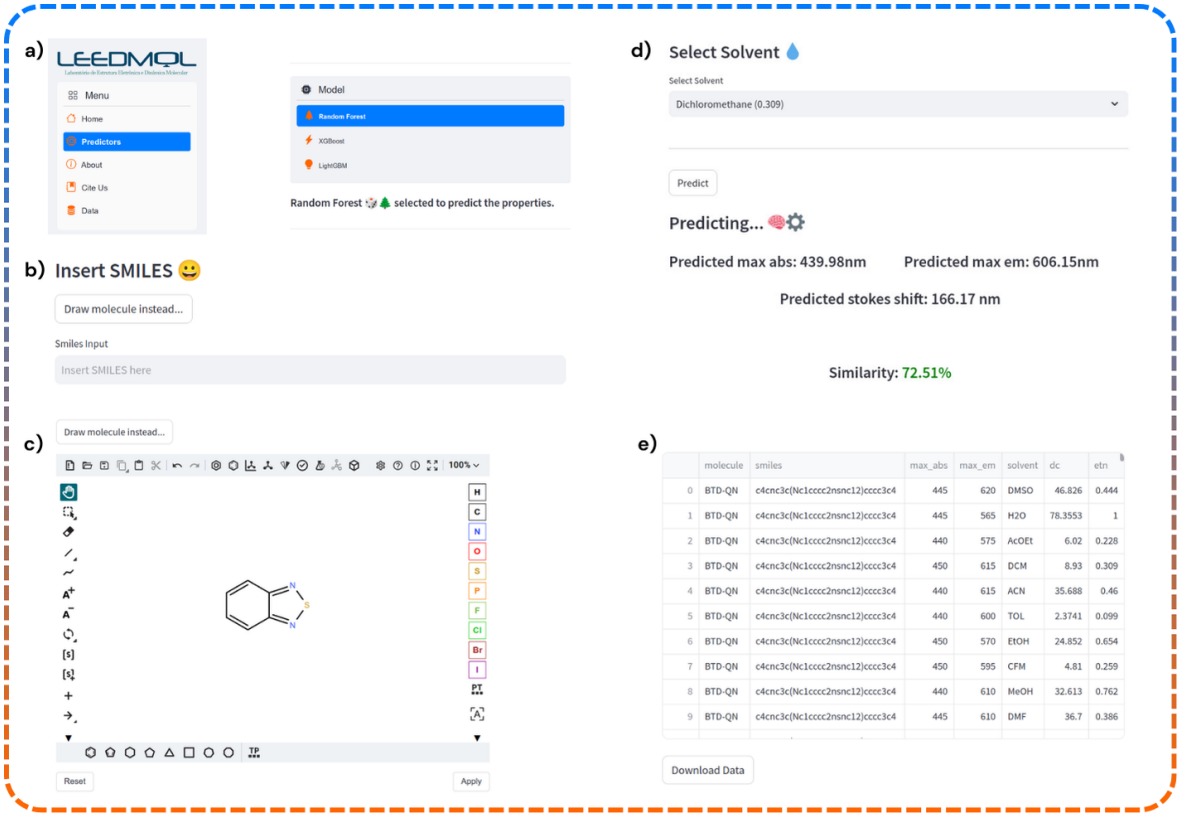


Figure 5.9: Overview of the web application workflow for predicting the photophysical properties of benzothiadiazole derivatives. (a) The main interface, where users can select the desired predictive model. The application offers two methods for structure input: (b) direct entry of a SMILES string or (c) use of an integrated molecular editor. (d) The output panel displays the selected solvent, the predicted maximum absorption and emissionwavelengths, the calculated Stokes shift, and the structural similarity to the nearest neighbor in the training data. (e) The interface also provides access to the underlying training dataset for transparency and further exploration.

5.4 Limitations

This study has several limitations stemming from the nature of its dataset. As the data was curated from existing literature, it reflects the biases present in published research, leading to an imbalance in the representation of certain solvents and structural motifs. This may limit the model's generalizability to less-explored chemical space. Furthermore, the model is specifically focused on derivatives of the 2,1,3-benzothiadiazole scaffold. Its predictive accuracy may be reduced when applied to other BTD isomers or derivatives whose electronic properties are not well-represented in the current training set.

To ensure transparency and promote further development, the source code for the web application is publicly available at https://github.com/LEEDMOL/BTD-JCIM-PAPER. The application is designed to protect user privacy by not storing any input data or prediction results on the server.

Chapter 6

Final Remarks

This work demonstrates the effectiveness of integrating machine learning (ML) models with interpretability analysis to predict and understand the photophysical properties of benzothiadiazole (BTD) derivatives. Using algorithms such as Random Forest (RF), LightGBM, and XGBoost, models were developed with high predictive accuracy, as confirmed by robust statistical metrics including R^2, Q^2, MAE, and RMSE. Beyond mere prediction, the application of SHAP analysis provides a clear mechanistic bridge to traditional photophysical theory.

The model successfully captures the fundamental relationship between the fluorophore and its environment, specifically regarding the phenomenon of solvatochromism. SHAP analysis reveals that the prediction of maximum emission (λ_{max}^{em}) is overwhelmingly governed by two factors: solvent polarity, represented by the E_T^N parameter, and the presence of tertiary amine substructures. Consistent with the Lippert-Mataga framework, high E_T^N values lead to significant bathochromic (red) shifts. This occurs because polar solvents stabilize the more polar excited state (S_1) to a greater extent than the ground state (S_0), thereby reducing the transition energy gap. This process initiates with a vertical transition to a Franck-Condon excited state, followed by solvent reorientation to reach a relaxed excited state from which emission occurs.

The study identifies tertiary amines as the primary structural drivers for emission red-shifts. These substructures function as potent electron-donating groups (EDGs), creating a classic Donor–π–Acceptor (D–π–A) system with the electron-deficient BTD core. This "push-pull" architecture enhances the Intramolecular Charge Transfer (ICT) character of the electronic transition, lowering the excited-state energy and causing a significant red-shift in both absorption and emission spectra.

External validation confirmed a clear trend between structural similarity and predictive accuracy. Molecules with high similarity to the training set (typically >70%) generally yielded low prediction errors. However, the model demonstrated an ability to generalize beyond superficial topology, accurately predicting some compounds with low similarity scores by identifying key underlying photophysical features. These findings establish a data-driven

strategy for the rational design of BTD-based fluorophores, allowing for the targeted modification of structures and solvent selection to achieve desired optical properties efficiently.

6.1 Perspectives

Future research will transition toward a multi-scale modeling approach by incorporating quantum-chemical descriptors derived from Density Functional Theory and Time-Dependent Density Functional Theory calculations. Specifically, features extracted from electron density distributions and molecular orbital symmetries will be integrated to complement existing topological Morgan fingerprints.

This hybrid framework is anticipated to provide a more physically rigorous representation of how substituent-induced electronic perturbations modulate photophysical behavior at the atomic level. By merging the predictive power of machine learning algorithms with the mechanistic transparency of SHAP analysis and the first-principles precision of quantum chemistry, subsequent studies will seek to establish a fully integrated discovery pipeline. Such an approach will further accelerate the rational design of BTD derivatives with finely tuned optoelectronic properties, facilitating their application in high-performance photonic devices and advanced bioimaging technologies.

Bibliography

[1] Justin Thomas, K.; Lin, J. T.; Velusamy, M.; Tao, Y.-T.; Chuen, C.-H. Color tuning in benzo [1, 2, 5] thiadiazole-based small molecules by amino conjugation/deconjugation: bright red-light-emitting diodes. *Advanced Functional Materials* **2004**, *14*, 83–90.

[2] Kitamura, C.; Tanaka, S.; Yamashita, Y. Design of narrow-bandgap polymers. Syntheses and properties of monomers and polymers containing aromatic-donor and o-quinoid-acceptor units. *Chemistry of Materials* **1996**, *8*, 570–578.

[3] Suzuki, T.; Fujii, H.; Yamashita, Y.; Kabuto, C.; Tanaka, S.; Harasawa, M.; Mukai, T.; Miyashi, T. Clathrate formation and molecular recognition by novel chalcogen-cyano interactions in tetracyanoquinodimethanes fused with thiadiazole and selenadiazole rings. *Journal of the American Chemical Society* **1992**, *114*, 3034–3043.

[4] Karikomi, M.; Kitamura, C.; Tanaka, S.; Yamashita, Y. New Narrow-Bandgap Polymer Composed of Benzobis(1,2,5-thiadiazole) and Thiophenes. *Journal of the American Chemical Society* **1995**, *117*, 6791–6792.

[5] Wu, J.; Lai, G.; Li, Z.; Lu, Y.; Leng, T.; Shen, Y.; Wang, C. Novel 2,1,3-benzothiadiazole derivatives used as selective fluorescent and colorimetric sensors for fluoride ion. *Dyes and Pigments* **2016**, *124*, 268–276.

[6] Farré, Y.; Raissi, M.; Fihey, A.; Pellegrin, Y.; Blart, E.; Jacquemin, D.; Odobel, F. Synthesis and properties of new benzothiadiazole-based push-pull dyes for p-type dye sensitized solar cells. *Dyes and Pigments* **2018**, *148*, 154–166.

[7] Carvalho, P. H.; Correa, J. R.; Guido, B. C.; Gatto, C. C.; De Oliveira, H. C.; Soares, T. A.; Neto, B. A. Designed benzothiadiazole fluorophores for selective mitochondrial imaging and dynamics. *Chemistry–A European Journal* **2014**, *20*, 15360–15374.

[8] Baranov, M. S.; Solntsev, K. M.; Baleeva, N. S.; Mishin, A. S.; Lukyanov, S. A.; Lukyanov, K. A.; Yampolsky, I. V. Red-Shifted Fluorescent Aminated Derivatives of a Conformationally Locked GFP Chromophore. *Chemistry–A European Journal* **2014**, *20*, 13234–13241.

[9] Yadav, K. P.; Rahman, M. A.; Nishad, S.; Maurya, S. K.; Anas, M.; Mujahid, M. Synthesis and biological activities of benzothiazole derivatives: A review. *Intelligent Pharmacy* **2023**, *1*, 122–132.

[10] Schleder, G. R.; Padilha, A. C.; Acosta, C. M.; Costa, M.; Fazzio, A. From DFT to machine learning: recent approaches to materials science–a review. *Journal of Physics: Materials* **2019**, *2*, 032001.

[11] Young, D. *Computational chemistry: a practical guide for applying techniques to real world problems*; John Wiley & Sons, 2004.

[12] Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Muller, K.-R.; Tkatchenko, A. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chemical reviews* **2021**, *121*, 9816–9872.

[13] Ceballos-Avila, D.; Vazquez-Sandoval, I.; Ferrusca-Martinez, F.; Jimenez-Sanchez, A. Conceptually innovative fluorophores for functional bioimaging. *Biosensors and Bioelectronics* **2024**, *264*, 116638.

[14] Joung, J. F.; Han, M.; Hwang, J.; Jeong, M.; Choi, D. H.; Park, S. Deep learning optical spectroscopy based on experimental database: potential applications to molecular design. *JACS Au* **2021**, *1*, 427–438.

[15] Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58–63.

[16] Sanches-Neto, F. O.; Dias-Silva, J. R.; Keng Queiroz Junior, L. H.; Carvalho-Silva, V. H. "py SiRC": machine learning combined with molecular fingerprints to predict the reaction rate constant of the radical-based oxidation processes of aqueous organic contaminants. *Environmental Science & Technology* **2021**, *55*, 12437–12448.

[17] Mahato, K. D.; Kumar, U. Optimized Machine learning techniques Enable prediction of organic dyes photophysical Properties: Absorption Wavelengths, emission Wavelengths, and quantum yields. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2024**, *308*, 123768.

[18] Ju, C.-W.; Bai, H.; Li, B.; Liu, R. Machine learning enables highly accurate predictions of photophysical properties of organic fluorescent materials: Emission wavelengths and quantum yields. *Journal of Chemical Information and Modeling* **2021**, *61*, 1053–1065.

[19] Mai, J.; Lu, T.; Xu, P.; Lian, Z.; Li, M.; Lu, W. Predicting the maximum absorption wavelength of azo dyes using an interpretable machine learning strategy. *Dyes and Pigments* **2022**, *206*, 110647.

[20] Veríssimo, R. F.; Matias, P. H. F.; Barbosa, M. R.; Neto, F. O. S.; Neto, B. A. D.; de Oliveira, H. C. B. Integrating Machine Learning and SHAP Analysis to Advance the Rational Design of Benzothiadiazole Derivatives with Tailored Photophysical Properties. *Journal of Chemical Information and Modeling* **2025**, *65*, 7874–7886, PMID: 40300554.

[21] Rohatgi-Mukherjee, K. *Fundamentals of photochemistry*; New Age International, 1978.

[22] Klán, P.; Wirz, J. *Photochemistry of organic compounds: from concepts to practice*; John Wiley & Sons, 2009.

[23] Lakowicz, J. R. *Principles of fluorescence spectroscopy*; Springer, 2006.

[24] Reichardt, C.; Welton, T. *Solvents and solvent effects in organic chemistry*; John Wiley & Sons, 2010.

[25] Jablonski, A. Efficiency of anti-Stokes fluorescence in dyes. *Nature* **1933**, *131*, 839–840.

[26] Condon, E. U. Nuclear Motions Associated with Electron Transitions in Diatomic Molecules. *Phys. Rev.* **1928**, *32*, 858–872.

[27] Hantzsch, A. Über die Halochromie und» Solvatochromie «des Dibenzal-acetons und einfacherer Ketone, sowie ihrer Ketochloride. *Berichte der deutschen chemischen Gesellschaft (A and B Series)* **1922**, *55*, 953–979.

[28] Ishchenko, A. A. Structure and spectral-luminescent properties of polymethine dyes. *Russian Chemical Reviews* **1991**, *60*, 865.

[29] Allen, N. S. Polymethine dyes: structure and properties : Edited by N. TyutTyulkov, J. Fabian,A. Melhorn, F. Dietz and A. Tadjer, published by St. Kliment Ohridski University Press Bulgaria, 250 pp., 1991. *Journal of Photochemistry and Photobiology A-chemistry* **1993**, *73*, 2–4.

[30] Reichardt, C. Solvatochromic Dyes as Solvent Polarity Indicators. *Chemical Reviews* **1994**, *94*, 2319–2358.

[31] Influence of the solvent on the colour of dyes (solvatochromism). *Russian Chemical Reviews* **1960**, *29*, 618–626.

[32] Dahne, S.; Moldenhauer, F. Structural principles of unsaturated organic compounds: evidence by quantum chemical calculations. *Progress in physical organic chemistry* **1985**, *15*, 1–130.

[33] Bakhshiev, N. Solvatochromism: problems and methods. *Izd. Leningr. Gos. Univ., Leningrad* **1989**,

[34] Haberfield, P. Carbonyl n. far.. pi.* solvent blue shift. Excited state solvation vs. ground state solvation. *Journal of the American Chemical Society* **1974**, *96*, 6526–6527.

[35] Taylor, P. R. On the origins of the blue shift of the carbonyl n-. pi.* transition in hydrogen-bonding solvents. *Journal of the American Chemical Society* **1982**, *104*, 5248–5249.

[36] DeBolt, S. E.; Kollman, P. A. A theoretical examination of solvatochromism and solute-solvent structuring in simple alkyl carbonyl compounds. Simulations using statistical mechanical free energy perturbation methods. *Journal of the American Chemical Society* **1990**, *112*, 7515–7524.

[37] Karelson, M.; Zerner, M. C. On the n-. pi.* blue shift accompanying solvation. *Journal of the American Chemical Society* **1990**, *112*, 9405–9406.

[38] Fox, T.; Rösch, N. The calculation of solvatochromic shifts: the n-π* transition of acetone. *Chemical physics letters* **1992**, *191*, 33–37.

[39] Botrel, A.; le Beuze, A.; Jacques, P.; Strub, H. Solvatochromism of a typical merocyanine dye. A theoretical investigation through the CNDO/SCI method including solvation. *J. Chem. Soc., Faraday Trans. 2* **1984**, *80*, 1235–1252.

[40] Jacques, P. On the relative contributions of nonspecific and specific interactions to the unusual solvtochromism of a typical merocyanine dye. *The Journal of Physical Chemistry* **1986**, *90*, 5535–5539.

[41] Kosower, E. M. The Effect of Solvent on Spectra. III. The Use of Z-Values in Connection with Kinetic Data. *Journal of the American Chemical Society* **1958**, *80*, 3267–3270.

[42] Dimroth, K.; Reichardt, C.; Siepmann, T.; Bohlmann, F. Über Pyridinium-N-phenol-betaine und ihre Verwendung zur Charakterisierung der Polarität von Lösungsmitteln. *Justus Liebigs Annalen der Chemie* **1963**, *661*, 1–37.

[43] Reichardt, C.; Harbusch-Görnert, E. Über Pyridinium-N-phenolat-Betaine und ihre Verwendung zur Charakterisierung der Polarität von Lösungsmitteln, X. Erweiterung, Korrektur und Neudefinition der ET-Lösungsmittelpolaritätsskala mit Hilfe eines lipophilen penta-tert-butyl-substituierten Pyridinium-N-phenolat-Betainfarbstoffes. *Liebigs Annalen der Chemie* **1983**, *1983*, 721–743.

[44] Neto, B. A. D.; Lapis, A. A. M.; da Silva Júnior, E. N.; Dupont, J. 2,1,3-Benzothiadiazole and Derivatives: Synthesis, Properties, Reactions, and Applications in Light Technology of Small Molecules. *European Journal of Organic Chemistry* **2013**, *2013*, 228–255.

[45] Souza, V. S.; Corrêa, J. R.; Carvalho, P. H.; Zanotto, G. M.; Matiello, G. I.; Guido, B. C.; Gatto, C. C.; Ebeling, G.; Gonçalves, P. F.; Dupont, J.; Neto, B. A. Appending ionic liquids to fluorescent benzothiadiazole derivatives: Light up and selective lysosome staining. *Sensors and Actuators B: Chemical* **2020**, *321*, 128530.

[46] Shi, Y.-F.; Yang, Z.-X.; Ma, S.; Kang, P.-L.; Shang, C.; Hu, P.; Liu, Z.-P. Machine learning for chemistry: basics and applications. *Engineering* **2023**, *27*, 70–83.

[47] Kass, G. V. An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **1980**, *29*, 119–127.

[48] Breiman, L. Some properties of splitting criteria. *Machine learning* **1996**, *24*, 41–47.

[49] Hunt, K. W. Recent measures in syntactic development. *Elementary English* **1966**, *43*, 732–739.

[50] Quinlan, J. R. *Machine learning*; Elsevier, 1983; pp 463–482.

[51] Murthy, S. K. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data mining and knowledge discovery* **1998**, *2*, 345–389.

[52] Sammut, C.; Webb, G. I. *Encyclopedia of machine learning and data mining*; Springer Publishing Company, Incorporated, 2017.

[53] Blanc, G.; Lange, J.; Tan, L.-Y. Top-down induction of decision trees: rigorous guarantees and inherent limitations. *arXiv preprint arXiv:1911.07375* **2019**,

[54] Breiman, L. Bagging predictors. *Machine learning* **1996**, *24*, 123–140.

[55] Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.

[56] Biau, G.; Scornet, E. A random forest guided tour. *Test* **2016**, *25*, 197–227.

[57] Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016; pp 785–794.

[58] Friedman, J.; Hastie, T.; Tibshirani, R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics* **2000**, *28*, 337–407.

[59] Guolin, K.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* **2017**, *30*.

[60] Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Computer Sciences* **2002**, *42*, 1273–1280, PMID: 12444722.

[61] Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. In *Chapter 12 - PubChem: Integrated Platform of Small Molecules and Biological Activities*; Wheeler, R. A., Spellmeyer, D. C., Eds.; Annual Reports in Computational Chemistry; Elsevier, 2008; Vol. 4; pp 217–241.

[62] Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation* **1965**, *5*, 107–113.

[63] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754, PMID: 20426451.

[64] Landrum, G. et al. rdkit/rdkit: 2024_09_6 (Q3 2024) Release. 2025; `https://doi.org/10.5281/zenodo.14943932`.

[65] Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **2017**, *30*.

[66] Lundberg, S. M.; Erion, G. G.; Lee, S.-I. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888* **2018**,

[67] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988**, *28*, 31–36.

[68] Martinez-Trevino, S. H.; Uc-Cetina, V.; Fernández-Herrera, M. A.; Merino, G. Prediction of natural product classes using machine learning and 13C NMR spectroscopic data. *Journal of Chemical Information and Modeling* **2020**, *60*, 3376–3386.

[69] Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019; pp 2623–2631.

[70] Kiralj, R.; Ferreira, M. Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. *Journal of the Brazilian Chemical Society* **2009**, *20*, 770–787.

[71] Larsen, W. A.; McCleary, S. J. The use of partial residual plots in regression analysis. *Technometrics* **1972**, *14*, 781–790.

[72] Colas, K.; Doloczki, S.; Kesidou, A.; Sainero-Alcolado, L.; Rodriguez-Garcia, A.; Arsenian-Henriksson, M.; Dyrager, C. Photophysical Characteristics of Polarity-Sensitive and Lipid Droplet-Specific Phenylbenzothiadiazoles. *ChemPhotoChem* **2021**, *5*, 632–643.

[73] Dyrager, C.; Vieira, R. P.; Nystrom, S.; Nilsson, K. P. R.; Storr, T. Synthesis and evaluation of benzothiazole-triazole and benzothiadiazole-triazole scaffolds as potential molecular probes for amyloid-B aggregation. *New J. Chem.* **2017**, *41*, 1566–1573.

[74] Fonseca, T.; de Oliveira, H.; Castro, M. Theoretical study of the lowest electronic transitions of sulfur-bearing mesoionic compounds in gas-phase and in dimethyl sulfoxide. *Chemical Physics Letters* **2008**, *457*, 119–123.

[75] Lippert, E. Spektroskopische Bestimmung des Dipolmomentes aromatischer Verbindungen im ersten angeregten Singulettzustand. *Zeitschrift für Elektrochemie, Berichte der Bunsengesellschaft für physikalische Chemie* **1957**, *61*, 962–975.

[76] Mataga, N.; Kaifu, Y.; Koizumi, M. Solvent Effects upon Fluorescence Spectra and the Dipolemoments of Excited Molecules. *Bulletin of the Chemical Society of Japan* **1956**, *29*, 465–470.

Appendix A

Appendix

A.1 Scatter Plots

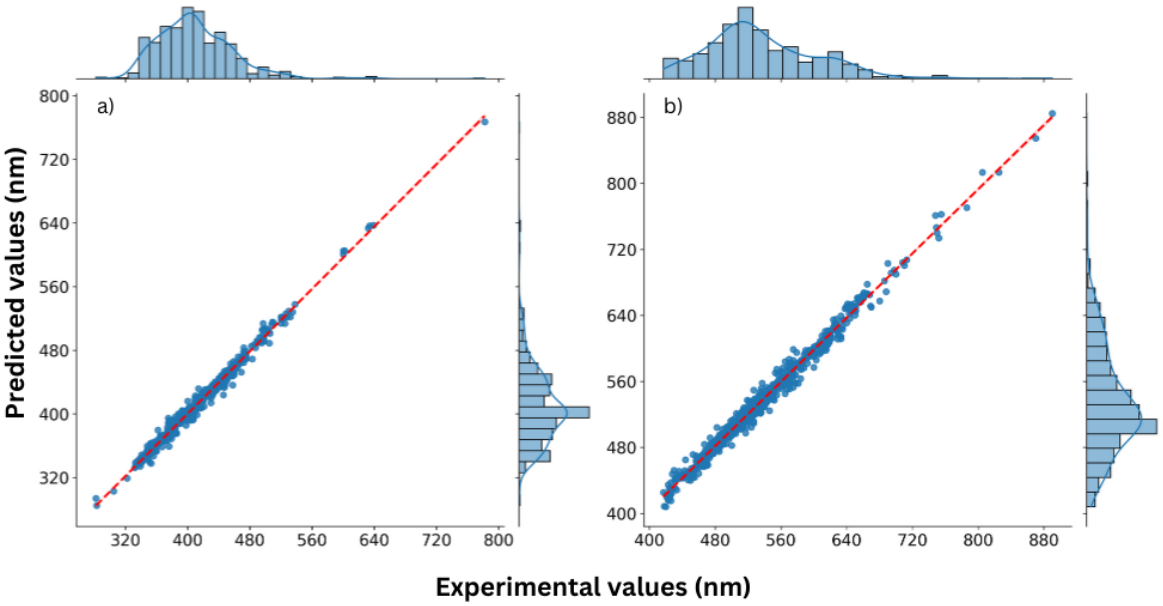


Figure A.1: Predicted values versus experimental values for a) max absorption and b) max emission for XGBoost, where the red line represents the ideal prediction and blue points depict the actual prediction.

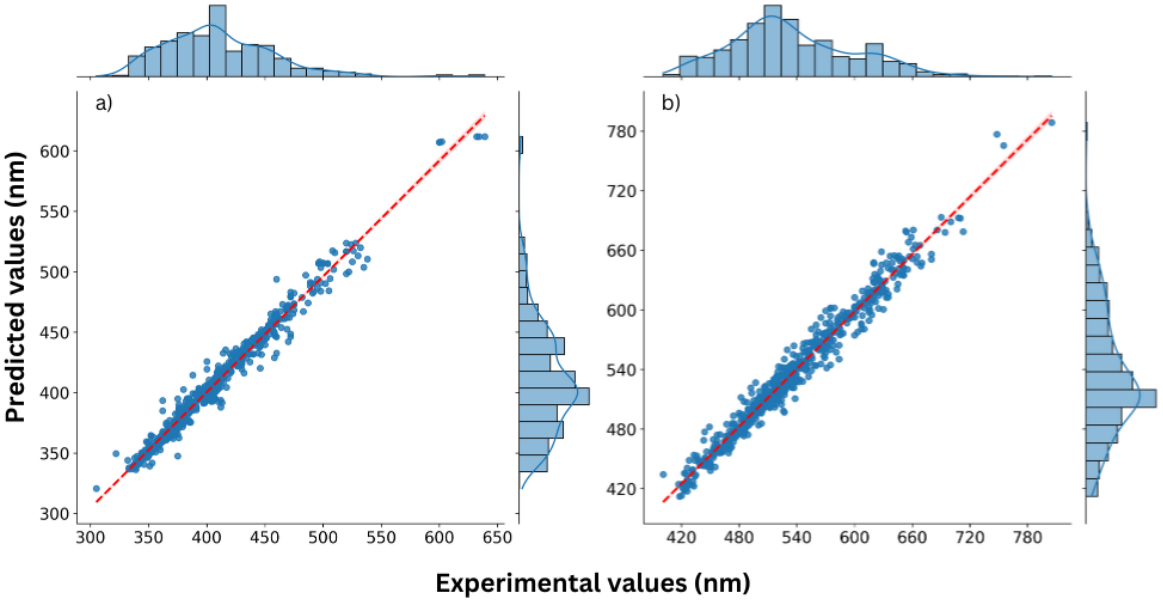
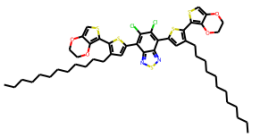
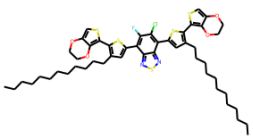


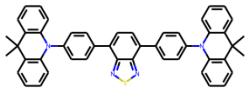
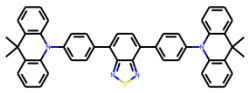
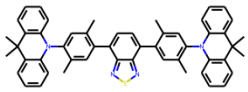
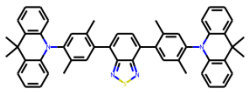
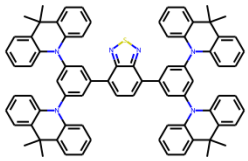
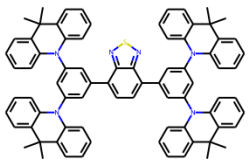
Figure A.2: Predicted values versus experimental values for a) max absorption and b) max emission for LightGBM, where the red line represents the ideal prediction and blue points depict the actual prediction.

A.2 Data in nanometers

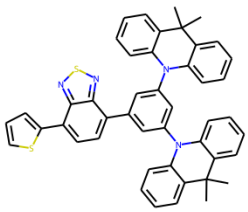
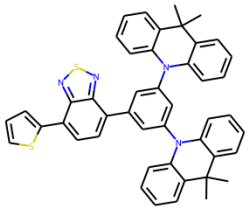
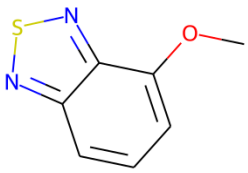
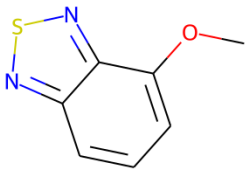
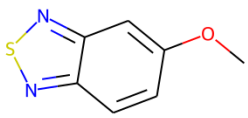
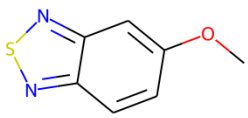
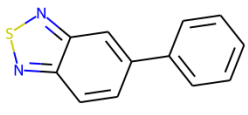
Table A.1: External testing of the RF model in predicting max absorption and max emission with experimental data from the literature in nm. Experimental values (Exp) were measured in different solvents: [A] Tetrahydrofuran (THF), [B] Toluene (TOL), [C] Methanol (MeOH), and [D] Dichloromethane (DCM). Predicted values (Pred) represent the model's outputs.

Molecules	Max Absorption			Max Emission			Similarity (%)
	Exp	Pred	Error	Exp	Pred	Error	
	399.00[A]	376.15	5.73%	514.00	495.07	3.68%	99.73
	364.00[A]	388.43	6.71%	475.00	508.82	7.12%	97.46

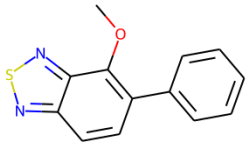
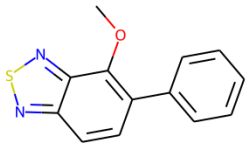
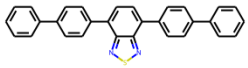
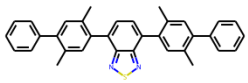
Continuation of Table A.1

Molecules	Max Absorption			Max Emission			Similarity (%)
	Exp	Pred	Error	Exp	Pred	Error	
	323.00[B]	383.25	18.65%	432.00	484.61	12.18%	74.43
	325.00[C]	384.39	18.27%	400.00	457.42	14.36%	74.43
	490.00[D]	439.98	10.21%	634.00	606.15	4.39%	72.51
	471.00[D]	455.84	3.22%	632.00	604.76	4.31%	69.80
	381.00[B]	386.52	1.45%	738.00	541.49	26.63%	66.90
	382.00[D]	387.14	1.35%	574.00	532.52	7.23%	66.90

Continuation of external validation

Molecules	Max Absorption			Max Emission			Similarity (%)
	Exp	Pred	Error	Exp	Pred	Error	
	418.00[B]	408.39	2.30%	610.00	546.44	10.42%	63.85
	416.00[D]	407.82	1.97%	766.00	569.19	25.69%	63.85
	357.00[B]	387.28	8.48%	728.00	544.48	25.21%	63.80
	356.00[D]	387.23	8.77%	540.00	544.87	0.90%	63.80
	353.00[B]	394.15	11.66%	484.00	566.0	16.94%	63.55
	353.00[C]	395.92	12.16%	533.00	619.83	16.29%	63.55
	379.00[B]	385.12	1.61%	622.00	535.34	13.93%	63.19

Continuation of external validation

Molecules	Max Absorption			Max Emission			Similarity (%)
	Exp	Pred	Error	Exp	Pred	Error	
	338.00[B]	399.01	18.05%	391.00	526.53	34.66%	56.94
	337.00[C]	398.47	18.24%	421.00	551.54	31.01%	56.94
	360.00[B]	410.61	14.06%	501.00	560.89	11.95%	50.17
	357.00[C]	407.43	14.13%	445.00	525.46	18.08%	50.17

A.3 SHAP

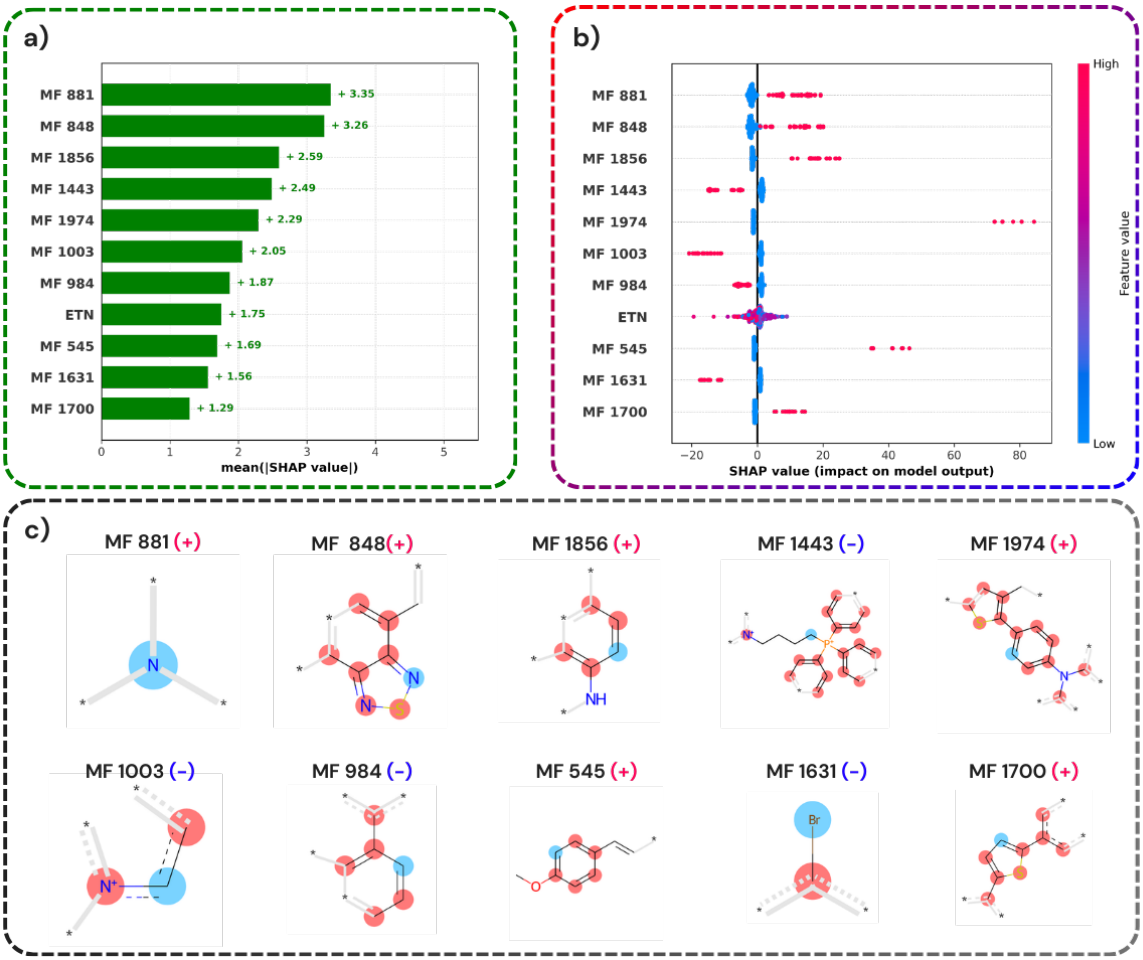


Figure A.3: Visualization of the SHAP analysis for the maximum absorption wavelength (λ_{max}^{abs}) XGBoost model. (a) Feature importance plot showing the mean absolute SHAP value for the most impactful molecular features. (b) SHAP summary plot, where each point's position on the x-axis indicates its impact on the model output for a specific molecule. The color corresponds to the feature's value (red: high/present, blue: low/absent). (c) Chemical structures of selected substructures, illustrating examples that contribute positively (+) or negatively (-) to the predicted λ_{max}^{abs}.

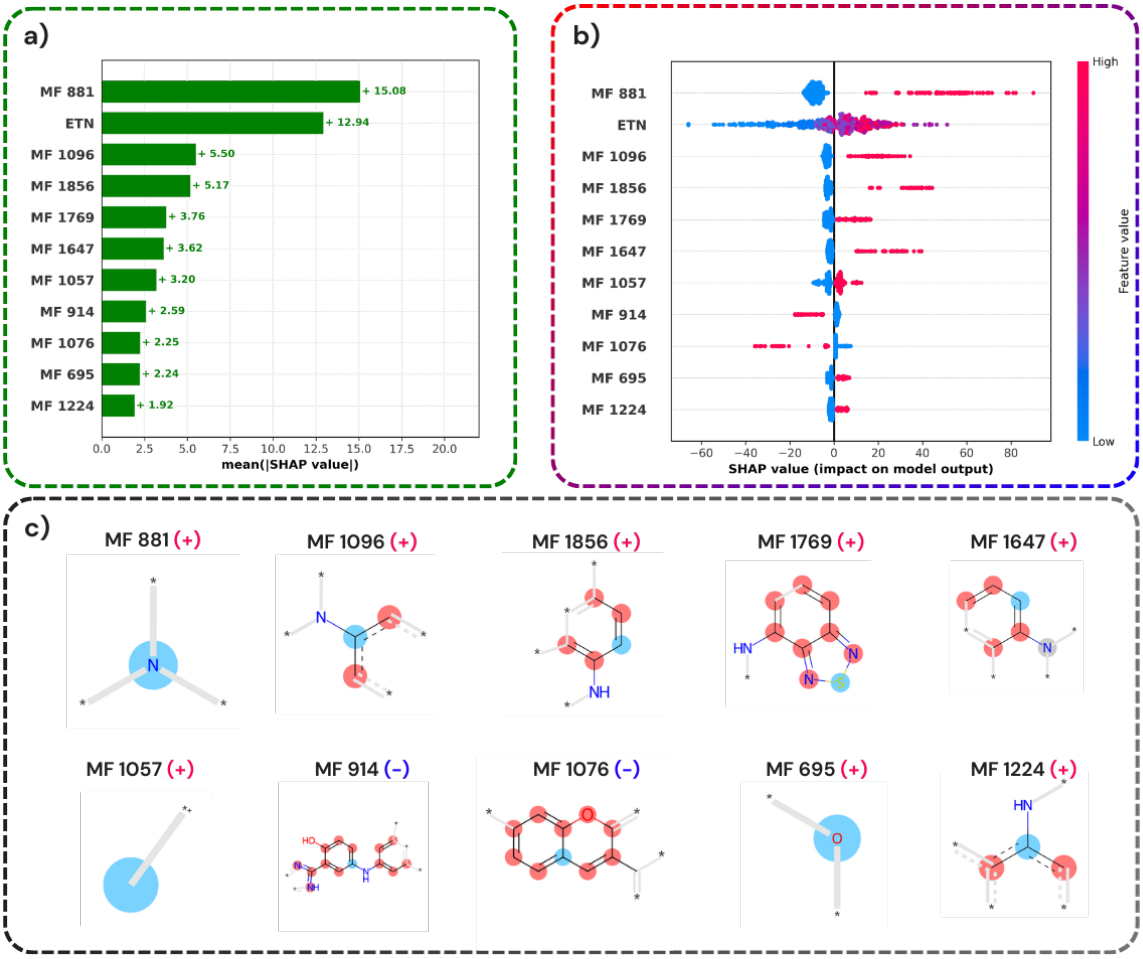


Figure A.4: Visualization of the SHAP analysis for the maximum emission wavelength (λ_{max}^{em}) XGBoost model. (a) Feature importance plot showing the mean absolute SHAP value for the most impactful molecular features. (b) SHAP summary plot, where each point's position on the x-axis indicates its impact on the model output for a specific molecule. The color corresponds to the feature's value (red: high/present, blue: low/absent). (c) Chemical structures of selected substructures, illustrating examples that contribute positively (+) or negatively (-) to the predicted λ_{max}^{em}.

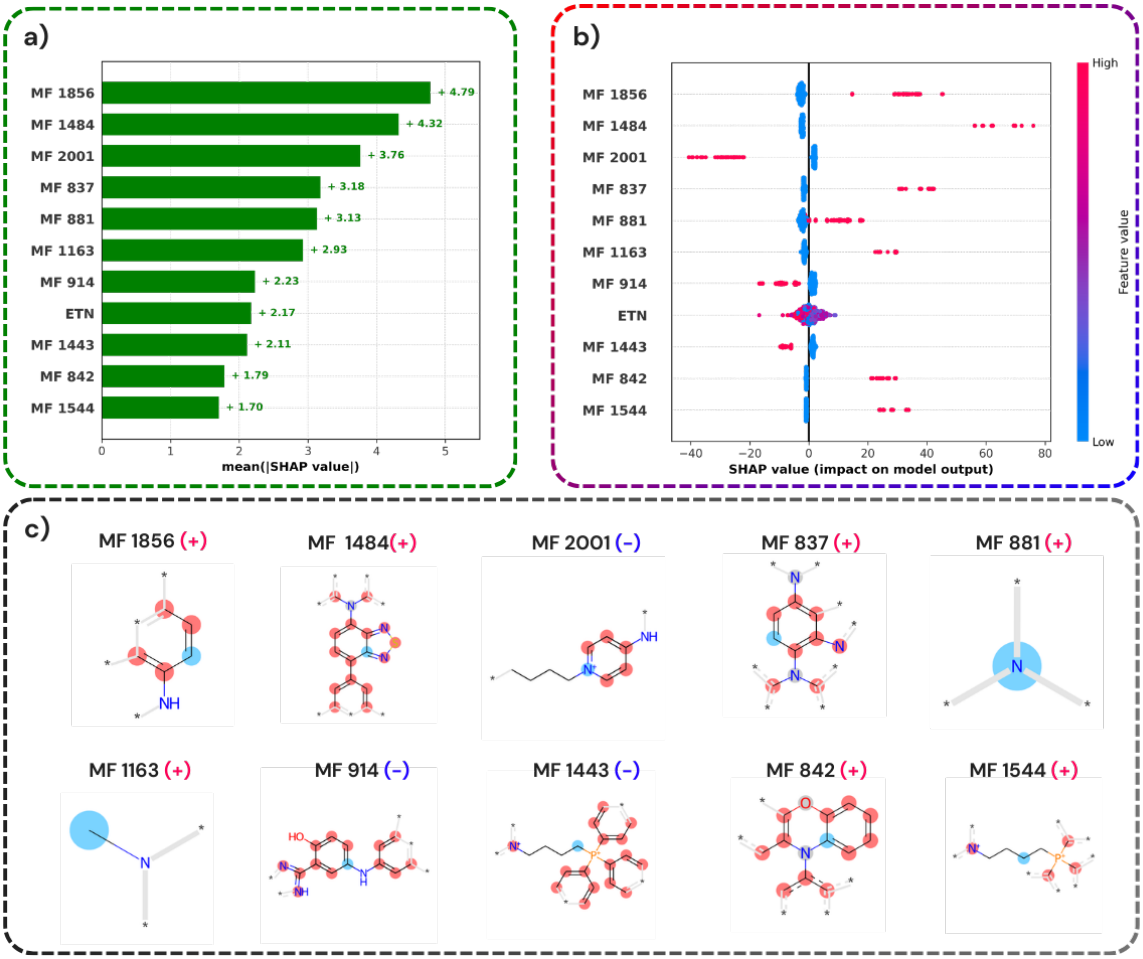


Figure A.5: Visualization of the SHAP analysis for the maximum absorption wavelength (λ_{max}^{em}) LightGBM model. (a) Feature importance plot showing the mean absolute SHAP value for the most impactful molecular features. (b) SHAP summary plot, where each point's position on the x-axis indicates its impact on the model output for a specific molecule. The color corresponds to the feature's value (red: high/present, blue: low/absent). (c) Chemical structures of selected substructures, illustrating examples that contribute positively (+) or negatively (-) to the predicted λ_{max}^{em}.

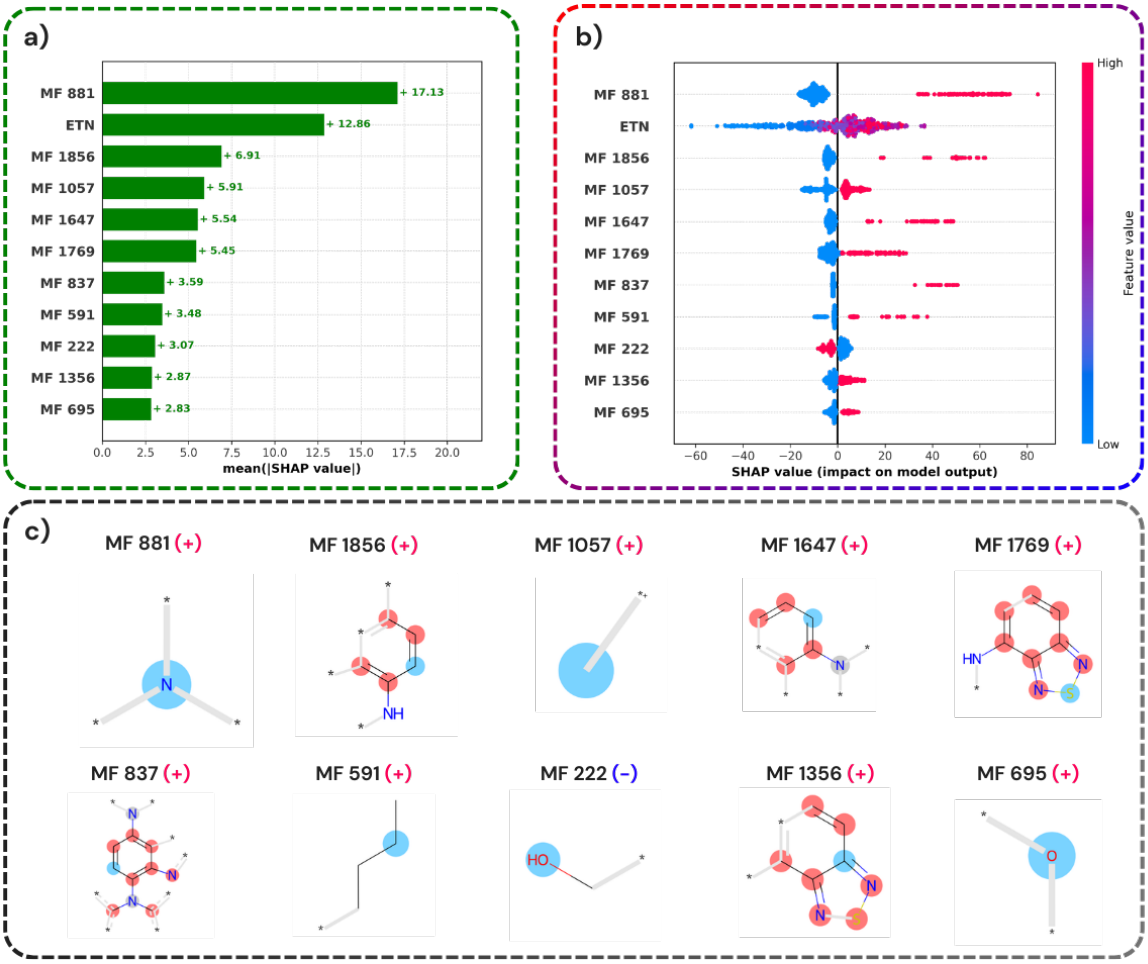


Figure A.6: Visualization of the SHAP analysis for the maximum emission wavelength (λ_{max}^{em}) LightGBM model. (a) Feature importance plot showing the mean absolute SHAP value for the most impactful molecular features. (b) SHAP summary plot, where each point's position on the x-axis indicates its impact on the model output for a specific molecule. The color corresponds to the feature's value (red: high/present, blue: low/absent). (c) Chemical structures of selected substructures, illustrating examples that contribute positively (+) or negatively (-) to the predicted λ_{max}^{em}.

A.4 Princial Contributions of the Dissertation

This section summarizes the main contributions of the dissertation. There is one publication 20 as the first author, which was published in an international journal of relevance in the field of computational chemistry and machine learning. The first page of this article is included in this section. Additionally, this article was selected as the front cover of the 15th issue of the journal and it too is included in this section. The article is available online at `https://pubs.acs.org/doi/10.1021/acs.jcim.4c02414`.

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Open Access
This article is licensed under CC-BY 4.0

pubs.acs.org/jcim
Article

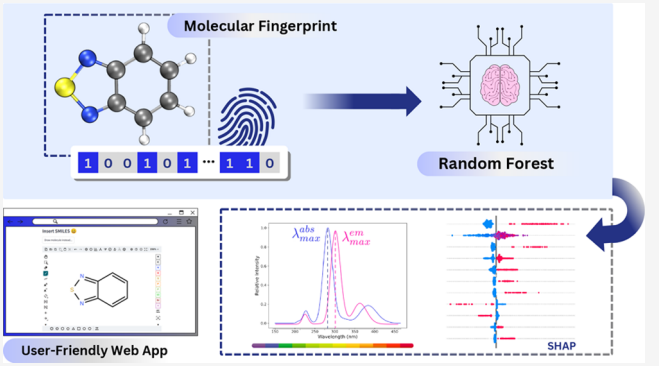
Integrating Machine Learning and SHAP Analysis to Advance the Rational Design of Benzothiadiazole Derivatives with Tailored Photophysical Properties

Rafael F. Veríssimo, Pedro H. F. Matias, Mateus R. Barbosa, Flávio O. S. Neto,* Brenno A. D. Neto, and Heibbe C. B. de Oliveira*

Cite This: https://doi.org/10.1021/acs.jcim.4c02414
Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: 2,1,3-Benzothiadiazole (BTD) derivatives show promise in advanced photophysical applications, but designing molecules with optimal desired properties remains challenging due to complex structure−property relationships. Existing computational methods have a high cost when predicting precise photophysical characteristics. Machine learning with Morgan fingerprints was employed to forecast BTD derivative maximum absorption and emission wavelengths. Three flavors of machine learning models were applied, namely, Random Forest, LigthGBM, and XGBoost. Random forest achieved R^2 values of 0.92 for absorption and 0.89 for emission, validated internally with 10-fold cross-validations and externally with recent experimental data. SHapley Additive exPlanations (SHAP) analysis revealed critical design insights, highlighting the tertiary amine presence and solvent polarity as key drivers of red-shifted emissions. By the development of a web-based predictive tool, the potential of machine learning to accelerate molecular design is demonstrated, providing researchers a powerful approach to engineer BTD derivatives with enhanced photophysical properties.



INTRODUCTION

2,1,3-Benzothiadiazole (BTD) derivatives have attracted significant attention due to their versatile photophysical properties and potential applications across various scientific and technological domains. These heterocyclic systems, featuring a fused thiadiazole ring and strong electron-withdrawing capacity,[1−3] have been widely explored in organic photovoltaics,[4] optoelectronics,[5] and as fluorescent probes in bioimaging,[6,7] among other applications (Figure 1). BTD-based fluorophores, known for their brightness,[8] large Stokes shifts,[9,10] and strong intramolecular charge transfer (ICT) interactions,[11−13] have demonstrated remarkable utility in fields where precise photophysical behavior is paramount. These derivatives often form ordered crystalline packings and can be tuned structurally to achieve the desired optical outputs. Such adaptability has led to their incorporation in devices requiring finely controlled emission spectra,[14] as well as their application in advanced fluorescence imaging, including the selective staining of mitochondria in cancer cell lines.[15−17]

Despite the promise of BTD analogues, identifying optimal candidates with target photophysical characteristics frequently relies on iterative synthetic efforts, extensive experimental screenings, or high-level computational chemistry. Traditional quantum chemical protocols, although insightful, can be expensive and time-consuming,[14,15] limiting their routine use for rapid compound discovery. This challenge is compounded by the subtle interplay of substituents, solvent polarity, and molecular conformation on the electronic transitions governing absorption and emission.[18] Consequently, there is a need for more accessible, efficient, and systematic strategies to predict photophysical properties without relying exclusively on expensive or specialized computational tools.

In recent years, machine learning (ML) methodologies have emerged as powerful alternatives for predicting molecular properties with reduced computational demands.[19] By leveraging large data sets and pattern recognition capabilities, ML models can correlate structural features with observed properties, bypassing the necessity for explicit quantum chemical descriptor calculations.[20,21] A particularly successful approach involves representing molecules through molecular fingerprints (e.g., Morgan or MACCS), which encode the presence or absence of specific substructures as binary

Received: January 14, 2025
Revised: March 27, 2025
Accepted: April 7, 2025

Downloaded via UNIV FED DE GOIAS on June 6, 2025 at 22:03:43 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

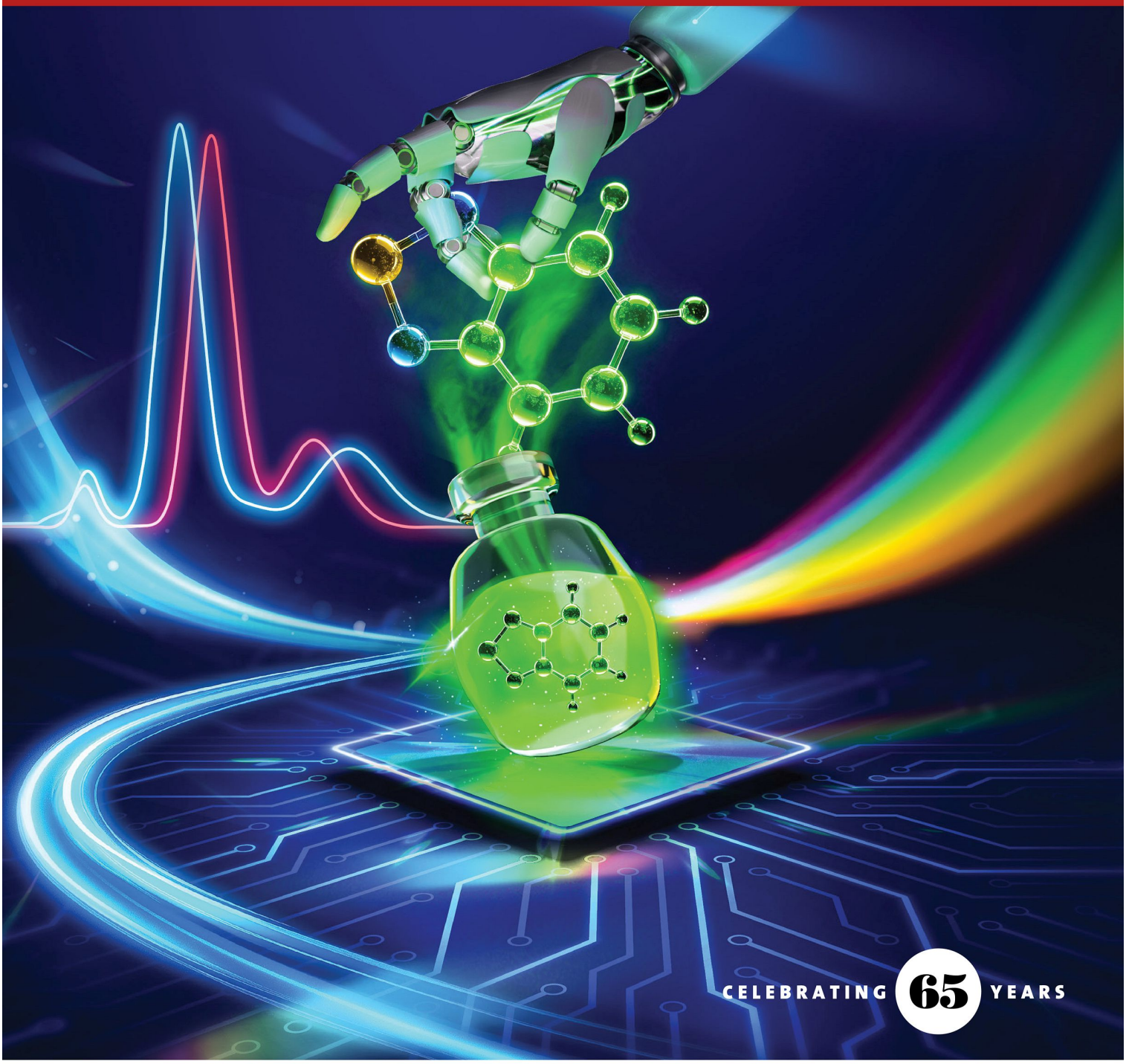
© XXXX The Authors. Published by American Chemical Society

August 11, 2025 Volume 65, Issue 15

pubs.acs.org/jcim

JCIM

JOURNAL OF CHEMICAL INFORMATION AND MODELING



ACS Publications
Most Trusted. Most Cited. Most Read.

www.acs.org