Journal of Metals, Materials and Minerals, 37(1), e2581, 2027

Machine learning prediction of energy band gaps in curcumin derivatives

Chamod Hasaranga PEIRIS[1,2], Elena ZABOROVA[3], Olivier MARGEAT[3], Rungroj CHANAJAREE[4], and Stephan Thierry DUBAS[1,2,5,*]

[1] *The Petroleum and Petrochemical College, Chulalongkorn University. Bangkok 10330, Thailand*
[2] *Center of Excellence on Petrochemical and Materials Technology, Bangkok 10330, Thailand*
[3] *Aix-Marseille Univ, CNRS, CINaM, Marseille 13288, France*
[4] *Metallurgy and Materials Science Research Institute, Chulalongkorn University, Phayathai Road, Wangmai, Patumwan, Bangkok 10330, Thailand*
[5] *Machine Learning for Polymers and Materials Discovery Research Unit, The Petroleum and Petrochemical College, Chulalongkorn University, Bangkok 10330, Thailand*

***Corresponding author e-mail**: stephan.d@chula.ac.th

Received date:
25 November 2025
Revised date:
23 June 2026
Accepted date:
18 August 2026

Keywords:
Machine learning;
Curcumin derivatives;
Molecular fingerprints;
Organic photovoltaics

Abstract

Machine learning (ML) regression models using SMILES-encoded molecular fingerprints were developed to predict the band gaps of 19 synthesized curcumin derivatives. The goal was to develop a ML-based framework capable of estimating the band-gap values prior to synthesis, thereby guiding chemists in their experimental effort. Training data for the model were obtained from the Harvard Clean Energy Project (HOPV-15), restricted to 52,000 non-fullerene acceptor (NFA) organic molecules with density functional theory (DFT)-calculated band gaps. Multiple fingerprinting schemes, including Morgan fingerprints, MACCS keys, atom-pairs, and FCFP, were used to train a pool of regression models. These models were then applied to the 19 curcumin derivatives for band-gap prediction and compared to DFT simulation calculations. The accuracy of the predictions was enhanced by using the Most Common Substructure (MCS) approach, narrowing it to molecules sharing key curcumin motifs. This refinement reduced the average predictive error by 30% compared with using the full HOPV-15 dataset, highlighting the benefit of targeted feature selection. Finally, the model was applied to virtually generate and evaluate 1,225 unsynthesized curcumin derivatives by permuting side groups on the BF_2-curcuminoid backbone, identifying ten candidates with predicted band gaps above 3.2 eV, which were selected as a proof-of-concept case study to demonstrate the capability of the proposed machine learning framework for targeted virtual screening. This workflow demonstrates how ML-guided screening can uncover promising curcumin-based chromophores prior to synthesis, reducing experimental costs while providing chemical insights into structure–property relationships.

1. Introduction

The development of organic solar cells (OSCs) represents a promising frontier in renewable energy, offering the potential for lightweight, flexible, and cost-effective alternatives to traditional silicon-based photovoltaics. Central to this advancement is the quest for efficient, stable, and low-cost donor-acceptor molecules which form the heterojunction that absorbs sunlight and separates charges in the cell. Zhang *et al.* reported the role of molecular engineering in optimizing donor-acceptor interactions for OSC for improved power conversion efficiency [1]. Despite their critical importance, the development of new chromophores is slow, tedious and expensive. Curcumin has been proposed as a good candidate, as it has an extended π-conjugated system (with conjugated double bonds and keto–enol tautomerism) that enables strong visible light absorption and facilitates charge transfer. As a natural dye with a high molar absorptivity, Curcumin also stands as a green chemistry eco-friendly material for OSCs [2]. However, pristine curcumin is limited by low photostability and poor solubility in organic solvents. Early studies of curcumin-based devices (e.g. in dye-sensitized cells) have only achieved modest efficiencies, in the order of a few percent [2,3], which, although lower than the 20% typically achieved nowadays by organic photovoltaics, underscores curcumin's potential in early-stage research [4].

To overcome these limitations, researchers are designing curcumin derivatives by attaching suitable donor or acceptor groups at various positions. Although chemical modifications improve solubility and stability while retaining curcumin's advantageous light-harvesting properties, scientists are limited by time and resources to synthesize thousands of curcumin derivatives with promising side groups. In recent years, particular attention has turned toward borondifluoride-curcuminoid (BF_2-curcuminoid) derivatives as functional chromophores. These dyes are best known for their application in organic light-emitting diodes (OLEDs), where the incorporation of a BF_2 unit enhances structural rigidity, enforces planarity, and increases photoluminescence quantum yields, resulting in bright and stable emission [5]. Beyond their emissive properties, BF_2-curcuminoids also offer strong and

tunable visible-light absorption, high molar extinction coefficients, and improved chemical stability compared with pristine curcumin. Such attributes address two of the major limitations of natural curcumin in optoelectronics (poor photostability and low solubility) thereby opening opportunities for energy-harvesting applications. Although their implementation in organic photovoltaics (OPVs) remains comparatively less explored, the optoelectronic profile of BF_2-curcuminoids makes them promising candidates for donor–acceptor systems. Substituent engineering at the R1and R2 positions provides a straightforward strategy to tailor HOMO–LUMO alignment, optimize molecular packing, and facilitate charge transfer across the heterojunction. These same design principles, which have driven their success in OLED devices, can thus be repurposed toward photovoltaic optimization. In this context, combining curcuminoid chemistry with machine learning–guided screening represents a powerful approach to accelerate the discovery of efficient, stable, and eco-friendly chromophores for OPVs.

Machine learning (ML) is transforming materials discovery by predicting molecular and device properties from data, thereby reducing costly trial-and-error experiments [6,7]. For example, physics-informed ML methods that incorporate quantum-mechanical descriptors can accurately forecast OSC performance metrics (e.g. short-circuit current, open-circuit voltage, fill factor, and PCE) from molecular input [8]. ML has also been effectively utilized to predict redox potentials in organic redox flow batteries, enabling the identification of molecules with optimized electrochemical properties, thereby reducing experimental burdens [9,10]. Utilizing ML models on organic photovoltaics machine learning can predict the exciton binding energy, a critical parameter for organic photovoltaics, achieving high accuracy and enabling accelerated material discovery [11–14]. Similarly, ML-guided frameworks have been applied to improve the design of direct methanol fuel cells by addressing critical challenges like methanol crossover, showcasing its versatility in diverse energy applications [15]. This approach mitigates the need for resource-intensive experiments or simulations, such as Density Functional Theory (DFT), by providing accurate predictions based on existing data [16]. A notable example is physics-informed ML approach integrated quantum mechanical descriptors to predict photovoltaic performance metrics like short-circuit current and power conversion efficiency, demonstrating the potential of ML for sustainable energy technology by Khatua *et al.*, 2024 [17]. A critical requirement for all these approaches is high-quality data. Aspuru-Guzik and co-workers built a database of ~51,000 non-fullerene acceptor molecules (containing 106 common molecular fragments) from high-throughput virtual screening using the Harvard Clean Energy Project [18]. Such large datasets enable training of ML regressors to screen candidates and predict properties.

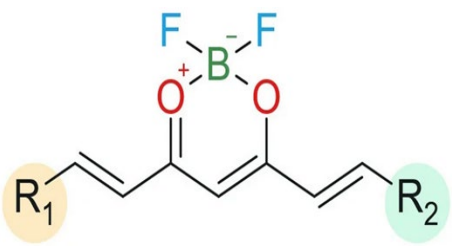


Figure 1. BF_2-curcuminoid bichromophore motif showing variable R1 and R2 end groups for derivative design.

The first step in ML involving organic molecules is to convert their structures into numerical features. A common strategy is to represent each molecule by a fingerprint: a binary or integer vector encoding the presence of chemical substructures or patterns. For example, toolkits like RDKit can convert a SMILES string into a bit-vector fingerprint (such as MACCS keys or Morgan circular fingerprints) [9,19–22]. Tanimoto, Dice and Cosine similarities are known to be among the most effective metrics for fingerprint-based comparisons. By selecting molecules that are most similar to a target (e.g. a known curcumin-based fragment), one can narrow the dataset to structurally relevant compounds. This helps the model train on data that share key functional features with the molecules of interest, improving prediction accuracy [21,22].

In this work we apply these concepts to curcumin-derived OSC materials. We focus on the subset of the non-fullerene acceptor (NFA) database relevant to small-molecule OSCs. First, we identify the six most common molecular fragments present in a set of 19 known curcumin derivatives. Using each of the six fragment as a query, we perform similarity searches in the NFA dataset (51,000+ molecules) to retrieve topologically similar compounds. This yields several fragment-specific subsets datasets of candidate molecules. We then encode all molecules with multiple fingerprint schemes (e.g. MACCS, Morgan, etc.) and train regression models (e.g. random forests, gradient boosting, etc.) to predict the fundamental band gap. To further refine the training data, we intersect the fragment-derived subsets: only molecules containing all of the key curcumin fragments are retained. This targeted selection concentrates on compounds that share curcumin's structural motifs. Our results show that this intersection-based dataset, together with diverse fingerprint features, produces the best predictive performance for band gap. Overall, this approach combines chemical insight (curcumin's fragments) with data-driven ML methods.

2. Experimental and methodology

A BF_2-curcuminoid bichromophore was used as the base scaffold for designing new curcumin derivatives. The general motif, illustrated in Figure 1, was modified at the R1 and R2 positions to generate a library of 19 derivatives.

Density functional theory (DFT) was utilized to optimize the molecular structures of curcumin derivatives and to calculate their lowest unoccupied molecular orbital (LUMO) and highest occupied molecular orbital (HOMO). All calculations were performed using the ORCA package [23] with the B3LYP/def2-SVP level of theory [24–26]. The energy gaps (LUMO-HOMO) of the 19 derivatives are summarized in the table below. The calculated energy gaps of curcumin derivatives are then used in the HOPV dataset. These values were used to evaluate model performance via the cumulative error, defined as the total absolute difference between ML-predicted and DFT-calculated band gaps across the 19 derivatives.

3. Molecular encoding approach and model training

Phase 1 – Full dataset model training:

The first phase involved training regression models—including Light Gradient Boosting Machine, Bayesian Ridge, and Ridge

Regression—on a complete Non-Fullerene Acceptor (NFA) dataset of 52,000 molecules extracted from the Harvard Organic Photovoltaic Project. Multiple molecular fingerprints, such as Morgan, MACCS, atom-pair, and FCFP, were employed to encode molecular structures. The trained models were then used to predict the band gaps of the 19 curcumin derivatives.

Phase 2 – Subset optimization via structural similarity:

In the second phase, the NFA dataset was filtered to include molecules structurally similar to the curcumin derivatives. The Most Common Substructure (MCS) algorithm from the RDKit library was used to identify six prevalent structural motifs across the derivatives. Molecular similarity searches were conducted using Tanimoto, Dice, Cosine, and Kulczynski metrics, combined with the three fingerprint types. Top-performing models were further optimized through recursive feature elimination to enhance predictive accuracy on the filtered dataset.

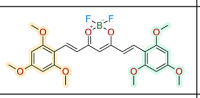
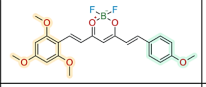
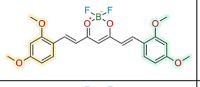
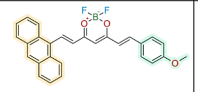
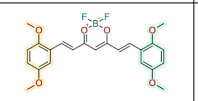
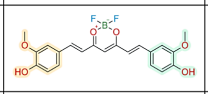
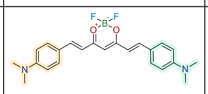
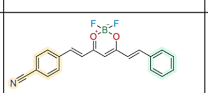
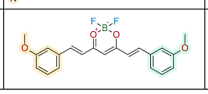
Phase 3 – Virtual molecule generation and prediction:

The final phase involved combinatorial generation of 1,225 virtual molecules by permuting all available R1 and R2 side groups from the initial dataset. The best-trained model from Phase 2 was applied to predict the band gap values of these virtual molecules. This approach identified promising candidates for experimental synthesis and subsequent evaluation of their optical properties.

4. Results and discussions

The primary objective of this study was to accurately predict the energy band gaps of all 19 curcumin derivatives. To achieve optimal predictive performance, a pool of regression models was employed. Using Scikit-learn and LightGBM libraries, a pipeline of 11 regression models were applied to the full Non-Fullerene Acceptor (NFA) dataset, with molecular structures encoded using four fingerprint types: Morgan, MACCS Keys, Atom-Pair, and FCFP. These fingerprints convert SMILES representations into binary vectors, with lengths varying depending on the fingerprint type. The predictive performance of each model-fingerprint combination, quantified by R^2 values, is summarized in Figure 3(a). Morgan and FCFP fingerprints consistently outperformed the others. Morgan Fingerprints utilize the Extended-Connectivity Fingerprint (ECFP) algorithm to encode local atomic environments iteratively, while FCFP (Functional Class Fingerprints) captures functional group information.

The superior performance of Morgan and FCFP fingerprints can be attributed to their ability to represent both detailed atomic environments (Morgan) and functional group interactions (FCFP) in a flexible, data-driven manner. In contrast, MACCS Keys and Atom-Pair fingerprints rely on predefined structural features or pairwise atom relationships, often missing critical molecular details. Most models achieved R^2 values around 0.85, with LightGBM consistently outperforming others (Figure 3). Given the balance between computational efficiency and structural representation, Morgan Fingerprints were selected to encode the curcumin derivatives in subsequent analyses. Each molecule was represented by a 2048-bit vector, effectively capturing detailed structural information relevant to band gap prediction.

Side Group 01	Side Group 02	Curcumin Derivative	Homo-Lumo Gap
			3.0770 eV
			3.0772 eV
			3.2593 eV
			2.6254 eV
			2.9377 eV
			2.9457 eV
			2.8115 eV
			3.2316 eV
			3.1367 eV

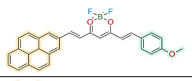
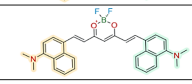
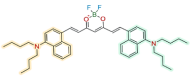
Side Group 01	Side Group 02	Curcumin Derivative	Homo-Lumo Gap
			2.7351 eV
			2.8777 eV
			2.6858 eV
			2.7700 eV
			3.0008 eV
			2.7022 eV
			2.9698 eV
			2.8902 eV
			2.8594 eV
			2.5251 eV

Figure 2. BF_2 curcuminoid template molecules with derivatives representing 19 possible configurations with HOMO-LUMO band gap values.

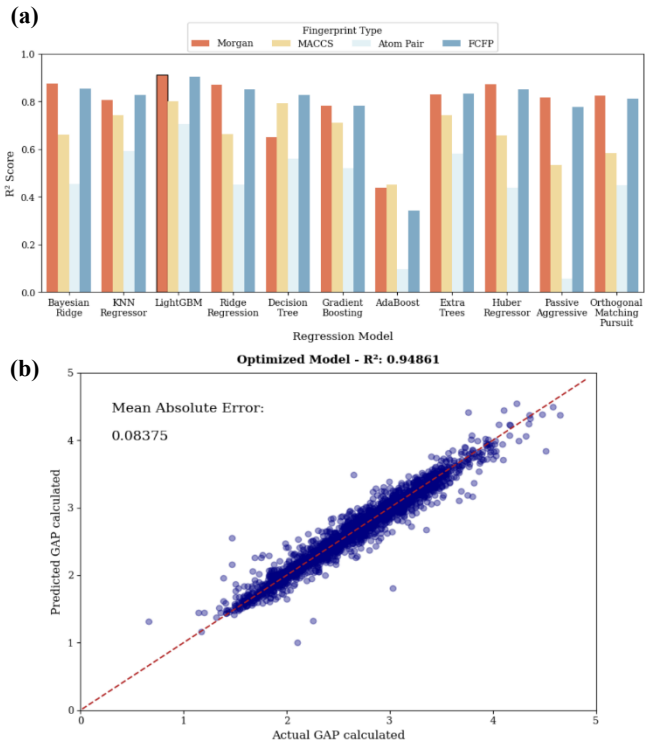


Figure 3. (a) Regression model R2 score comparison of each fingerprint method, and (b) Best model (Light GBM) optimized model.

To prevent overfitting of the model, the dataset was randomly partitioned into an 80% training set and a 20% independent testing before the evaluation. During the model optimization phase, a 5-fold cross-validation strategy (cv=5) was employed on the training data to tune hyperparameters. The final predictive performance metrics were calculated exclusively on the unseen 20% test set. This demonstrates the model's ability to generalize to new molecular structures.

The Light GBM model was used to predict the Band gaps of the initial set of 19 curcumin derivatives and compared with DFT-calculated values. Figure 4 illustrates the predictions, where the sum of the error between ML predicted and DFT calculated gap are equal to the total prediction error. Validation against the 19 synthesized curcumin molecules demonstrated good accuracy, with a cumulative error of 0.15265 eV relative to the DFT-calculated band gaps. This validation confirmed that the optimized model could reliably predict the HOMO–LUMO gaps of new curcumin derivatives and established a baseline for further model refinement.

5. Identification of the most common structures

Although training the LightGBM model with Morgan fingerprints achieved an R^2 value of 0.93 on the full dataset, prediction accuracy for the curcumin derivatives was slightly inconsistent, with notable deviations observed for molecules 8, 9, and 16 (Figure 4). This variability is likely due to the structural diversity of molecules in the NFA dataset when compared to the curcumin motif. To address this, the training dataset was restricted to molecules structurally similar to curcumin or containing substructures present in curcumin derivatives. The Most Common Substructure (MCS) approach in RDKit was used to identify six frequently occurring segments within the curcumin derivatives. Each segment was then subjected to similarity searches and fingerprint analysis using four similarity metrics—Dice, Tanimoto, Cosine, and Kulczynski—combined with three fingerprint types: Morgan, Atom-Pair, and FCFP.

This strategy generated a preliminary set of 72 models (Figure 5) for further evaluation, improving the likelihood of accurate predictions for the curcumin derivative series.

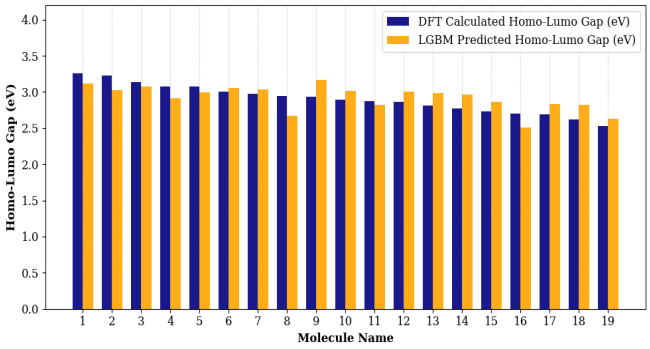


Figure 4. Band gap prediction for 19 curcumin derivative (yellow) and DFT calculated values (blue).

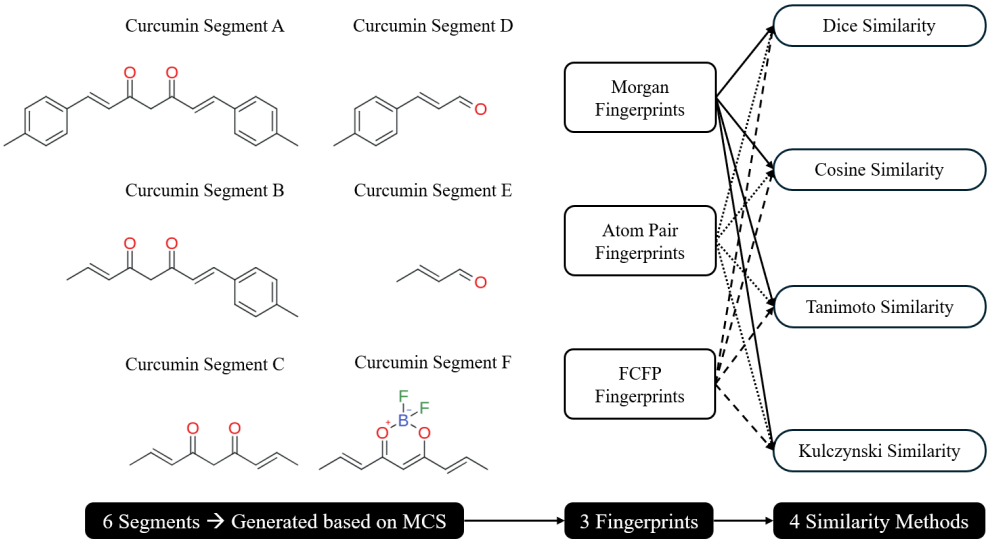


Figure 5. Segmented analysis - 72 models building principle using 6 MCS segmented structures, 3 fingerprint methods and 4 similarity algorithms.

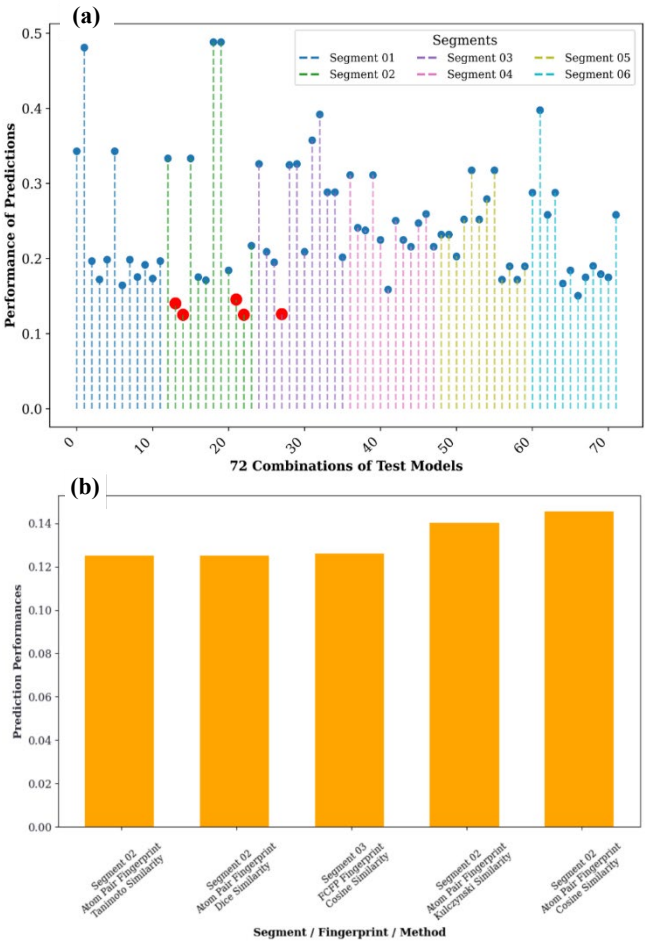


Figure 6. (a) Prediction cumulative error for the 72 combinations of test models, and (b) Prediction cumulative error for the best 5 model (seen as red dots in (a)).

The RDKit Most Common Substructure (MCS) tool was employed to identify key substructures shared across the curcumin derivatives. The analysis highlighted three prevalent motifs: the conjugated backbone (π–π*), end-group benzene rings (dense π-electron clouds), and carbonyl functionalities (π-electron-rich double bonds) (Figure 5). These identified segments were then submitted to our predictive algorithm. The alignment between the MCS results and the model-based learning approach is notable. By focusing on π-electron-rich segments during similarity filtering, the models effectively captured structural features that are directly relevant to the energy gap. This segmented approach reduces interference from less relevant molecular details, enhancing the learning framework and improving predictive accuracy for organic photovoltaic compounds.

From the initial set of 72 models, five top-performing models were identified based on their R^2 scores and cumulative error, which reflect predictive accuracy (Figure 6). Figure 6(a) shows the model performance on the training and test sets, while Figure 6(b) presents validation results against the 19 curcumin derivatives. Notably, models incorporating Atom-Pair fingerprints combined with Tanimoto or Dice similarity consistently exhibited superior performance (Figure 7(c-d). For example, Segment 02, trained with Atom-Pair fingerprints and Dice similarity, achieved a cumulative error of 0.12522 eV, demonstrating excellent prediction accuracy (Figure 6(d)). Atom-Pair fingerprints capture the frequency and presence of atom pairs, including functional groups or ring systems, making them sensitive to local structural features that critically influence HOMO–LUMO gaps in curcumin derivatives. During validation, all five top models maintained strong predictive performance, despite minor variations in error metrics. Segment 06 (Atom-Pair with Cosine similarity) achieved a cumulative error of 0.15603 eV (Figure 7(b)), indicating accurate representation of the electronic environment of the derivatives. Among all segments, Segment 02—which incorporates the conjugated backbone, aromatic rings, and carbonyl groups—emerged as the most effective for encoding the dataset. When trained with the optimized LightGBM model using Morgan fingerprints, this segment produced the highest predictive accuracy. Its combination with Atom-Pair fingerprints allows for extensive capture of functional features and π–π* electron interactions, explaining why segment-based filtering consistently outperforms models trained on the full dataset.

The segmentation studies provide insight into how molecular structure governs electronic characteristics, particularly the HOMO–LUMO gap. They underscore how localized structural motifs strongly influence electronic properties. They highlight the theoretical foundations underlying the impact of molecular structure on electronic properties. The success of the segmented models underscores the importance of localized electronic interactions within specific molecular segments. Features such as electron density around regions of the molecule significantly affect the electrical distribution in curcumin compounds. Since the HOMO-LUMO energy gap is influenced by the complex interplay of electronic effects throughout the entire molecule, elimination of extra features leads the model to capture this information more effectively.

Figure 9 emphasizes the upgrades studied throughout the project, where the final model improved to predict specifically derivatives of curcumin with minimum error. By concentrating on the common structural features while eliminating the unwanted atomic information of the molecules, this method successfully reduces noise and improves model precision. It aligns with contemporary advances in cheminformatics and machine learning, which advocate considering multiple structural perspectives to achieve more accurate predictions of molecular properties.

6. Prediction of virtual molecules

Once model trained with the best circumstances, the potential of curcumin derivatives as promising candidates for energy applications is observed by systematically permutating 1,225 unique molecules using RDKit. Each derivative was created by substituting two side groups on the curcumin-BF_2 backbone (Figure 1), with 35 carefully selected side groups for each substitution site. The generated dataset represents (Figure 10) a diverse chemical space of curcumin derivatives, enabling the evaluation of their electronic properties.

Utilizing the machine learning model developed and comprehended, we predicted the HOMO–LUMO gaps for all 1,225 derivatives. From these predictions, we identified the top 5 molecules with the highest band gaps, showcasing their potential for enhanced electronic performance. This workflow demonstrates the effectiveness of combining computational chemistry with machine learning to accelerate the discovery of novel functional materials.

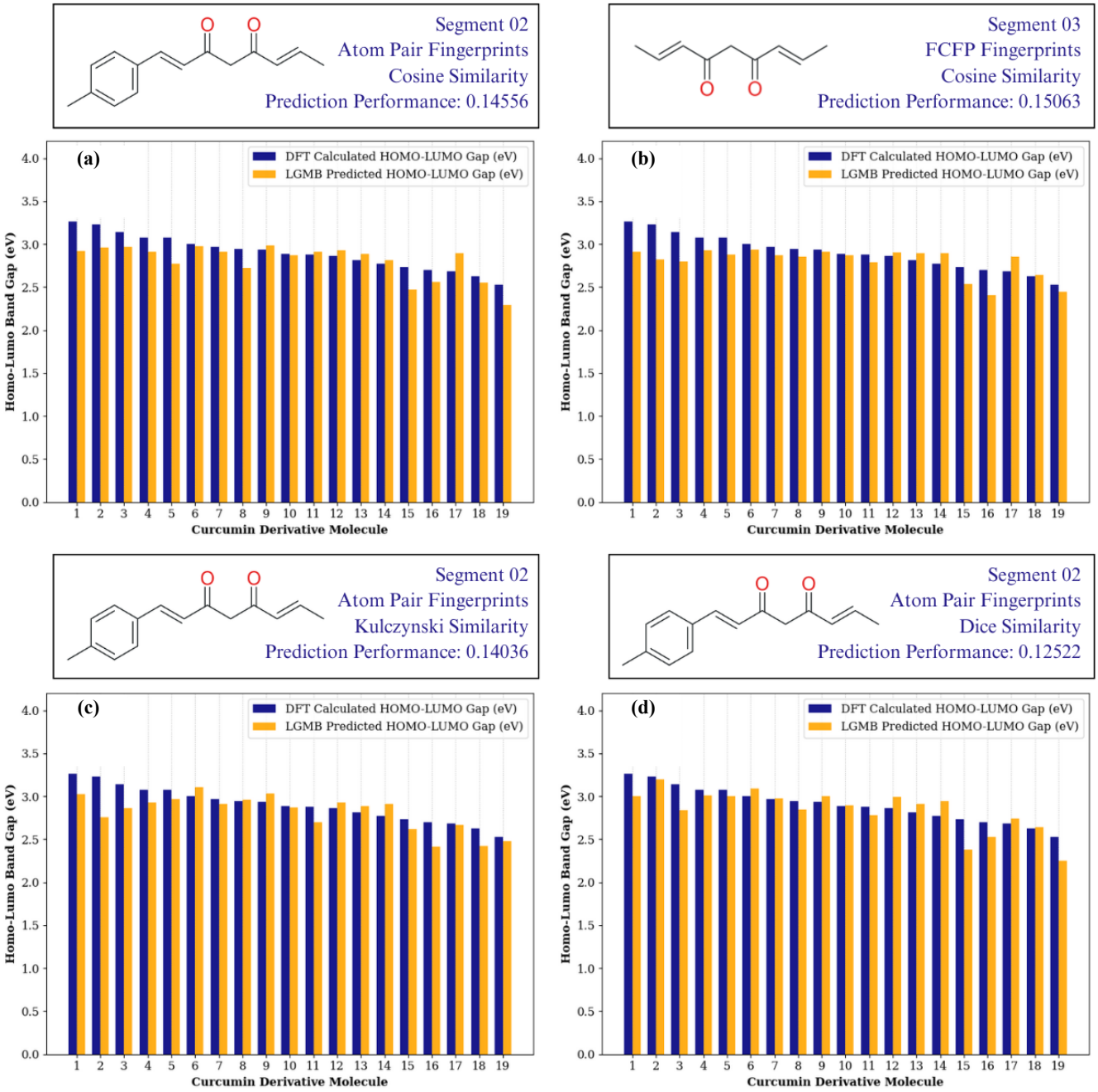


Figure 7. Data Validation Performed using best models from Similarity Analysis on 19 Curcumin Derivatives Band Gap.

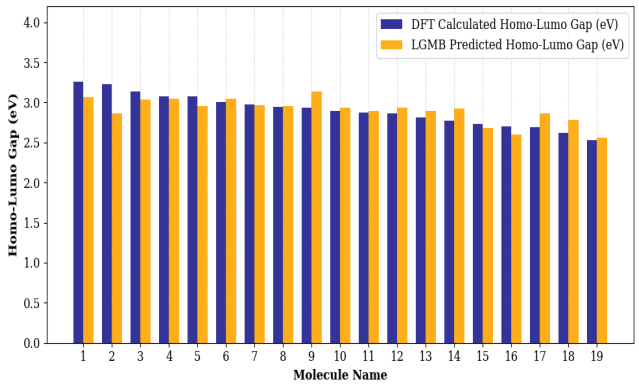


Figure 8. Validation test on curcumin derivatives with the band gap prediction of 810 selected features trained model.

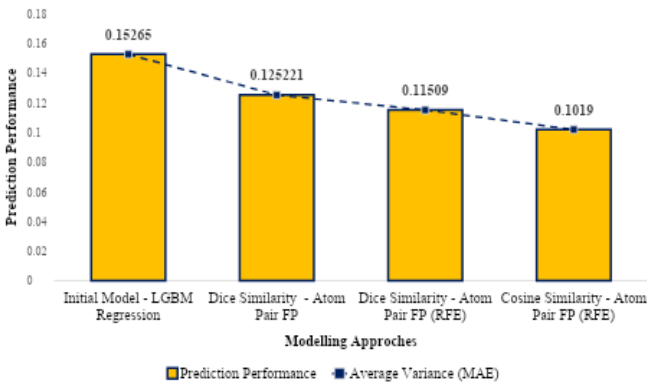


Figure 9. Comparison of Average Variance with cumulative Error for Different Approaches taken throughout the project to predict curcumin band gap.

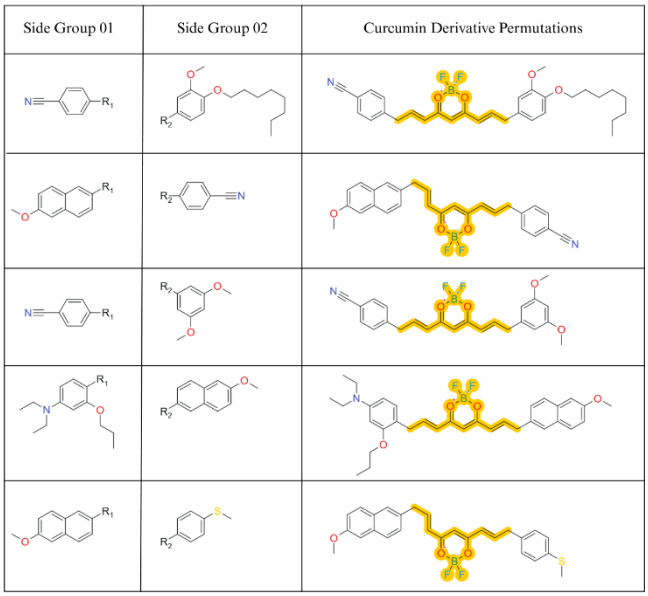


Figure 10. Top Five permutated molecules that could be used as synthesis candidates from the 1225 permutated molecules database

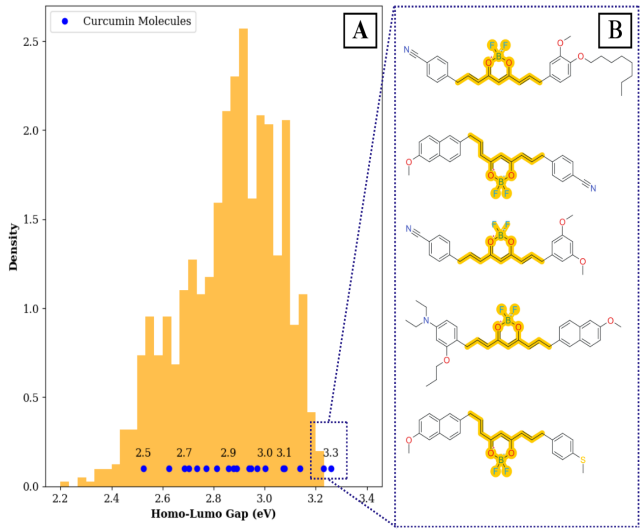


Figure 11. (a) Distribution of permutated curcumin derivatives with the 19 synthesized curcumin molecules band gap range, and (b) Top 5 suggested permutated curcumin derivatives with band gap value higher than 3.20 eV.

We successfully demonstrated the predictive capability of the proposed machine learning model by identifying curcumin derivatives with targeted HOMO–LUMO band gap values. The analysis of the permutated curcumin molecules revealed the top five candidates, with predicted band gaps consistent with the trends observed for the experimentally synthesized derivatives. As a proof-of-concept demonstration of the virtual screening framework, we selected molecules with predicted band gaps exceeding 3.20 eV. This threshold was chosen solely to illustrate the model's ability to identify compounds satisfying a predefined electronic-property criterion and was not intended to represent the optimal band-gap range for organic photovoltaic absorbers.

These virtual derivatives demonstrate the capability of the proposed machine learning framework to efficiently screen large chemical spaces and prioritize candidates for future experimental investigation. Although the present work used a threshold of 3.20 eV as an illustrative case study, the workflow is readily adaptable to any desired target property, including the lower band-gap range typically required for high-performance organic photovoltaic materials. This approach highlights the potential of integrating machine learning with molecular engineering to accelerate the rational design and experimental selection of new curcumin-based functional materials.

7. Conclusion

This study successfully established a machine learning pipeline for accurately predicting the HOMO–LUMO band gaps of curcumin derivatives. The optimized model, refined through molecular segmentation and recursive feature elimination, achieved a cumulative error of just 0.1019 eV representing a substantial improvement over the broad-dataset baseline. This underscores the importance of targeted data curation and feature selection in improving predictive precision. The optimized framework was then applied to virtually screen 1,225 curcumin derivatives, leading to the identification of several candidate molecules with predicted band gaps exceeding 3.20 eV as a proof-of-concept demonstration of the proposed virtual screening methodology. These findings demonstrate the utility of data-driven approaches in accelerating the discovery and design of high-performance organic materials. Beyond curcumin, the methodology can be extended to other natural or bio-inspired chromophores, opening pathways toward sustainable molecular design. Integrating additional descriptors such as exciton binding energies or charge-transport parameters into the ML pipeline could further enhance predictive power.

Acknowledgements

This research was financially supported by the 90th anniversary of chulalongkorn university scholarship and the petroleum and petrochemical college, chulalongkorn university.

References

[1] B. Zhang, M. Jiang, P. Mao, S. Wang, R. Gui, Y. Wang, H. Y. Woo, H. Yin, J.-L. Wang, and Q. An, "Manipulating alkyl inner side chain of acceptor for efficient as-cast organic solar cells," *Advanced Materials*, vol. 36, no. 36, p. 2405718, 2024.

[2] G. Sharma, S. K. Saini, K. Mulchandani, A. Bheemaraju, and C. Lal, "Investigation of ultrafast carrier dynamics in curcumin dye for environment friendly dye-sensitized solar cell," *Environmental Science and Pollution Research*, vol. 30, pp. 121175–121181, 2023.

[3] Y. Liu, H. Yang, Y. Wu, H. Fan, X. Li, K. Hu, C. Cui, and Y. Li, "Simultaneous improvement in efficiency and photostability of organic solar cells by modifying the ZnO electron-transport layer with curcumin," *Journal of Materials Chemistry C*, vol. 11, no. 38, pp. 13010–13017, 2023.

[4] X. Zarate, I. González P, S. Caramori, E. Benazzi, T. Barra, L. Arrue, Y.-n. Wu, C. Diaz-Uribe, W. Vallejo, and E. Schott, "Experimental and DFT study of natural curcumin derived dyes as n-type sensitizers," *Solar Energy*, vol. 225, pp. 305–315. 2021.

[5] C. Tonnelé, M. Catherin, M. Giorgi, G. Canard, D. Casanova, F. Castet, E. Zaborova, and F. Fages, "Optoelectronic properties of a self-assembling rigidly-linked BF_2-curcuminoid bi-chromophore," *Dyes and Pigments*, vol. 207, p. 110677, 2022.

[6] S. Zhong, W. Hsu, H. Chen, T. Yang, J. Yi, C. Zhu, S. Yin, Z. Li, L. Gao, J. Lin, L. Ying, and N. Li, "A simple and promising prediction model to analyze the optical properties of organic photovoltaic materials," *Solar RRL*, vol. 8, no. 12, p. 2400288, 2024.

[7] N. Meftahi, M. Klymenko, A. J. Christofferson, U. Bach, D. A. Winkler, and S. P. Russo, "Machine learning property prediction for organic photovoltaic devices," *NPJ Computation Materials*, vol. 6, no. 166, 2020.

[8] G. H. Kim, K. Yoon, C. Lee, M. Nam, and D.-H. Ko, "Molecular structural descriptor-assisted machine learning for organic photovoltaics with perylenediimide acceptors," Bulletin of the Korean Chemical Society, vol. 45, no. 2, pp. 125–130, 2023.

[9] P. Gao, D. Kochan, Y.-H. Tang, X. Yang, and E. G. Saldanha "Machine learning for the redox potential prediction of molecules in organic redox flow battery," *Journal of Power Sources*, vol. 629, p. 236035, 2025.

[10] W. Chen, Y. Fu, and P. Stinis, "Physics-informed machine learning of redox flow battery based on a two-dimensional unit cell model, *Journal of Power Sources*, vol. 584, p. 233548, 2023.

[11] L. Zhu, M. Huang, G. Han, Z. Wei, and Y. Yi, "The key descriptors for predicting the exciton binding energy of organic photovoltaic materials," *Angewandte Chemie International Edition*, vol. 64, no. 1, p. e202413913, 2024.

[12] K. Khoussa, P. Leveque, and L. Boubchir, "On the use of machine learning to discover novel donor-acceptor pairs for organic photovoltaic devices," in 2024 IEEE International Conference on Big Data, Institute of Electrical and Electronics Engineers Inc, 2024, p. 4659–4662.

[13] W. Sun, Y. Zheng, K. Yang, Q. Zhang, A. A. Shah, Z. Wu, Y. Sun, L. Feng, D. Chen, Z. Xiao, S. Lu, Y. Li, and K. Sun, "Machine learning-assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials," *Science Advances*, vol. 5, no. 11, p. eaay4275, 2019.

[14] L. Zhu, M. Huang, G. Han, Z. Wei, and Y. Yi, "The Key descriptors for predicting the exciton binding energy of organic photovoltaic materials," *Angewandte Chemie International Edition*, vol. 64, no. 1, p. e202413913, 2024.

[15] W. J. M. Kort-Kamp, C. L. Vecchio, D. Sebastián, X. Lyu, P. Atanassov, P. Zelenay, V. Baglio, and A. Serov, "Machine learning-guided design of direct methanol fuel cells with a platinum group metal-free cathode," *Journal of Power Sources*, vol. 626, p. 235758, 2025.

[16] B. Das, and A. Mondal, "Predictive modeling and design of organic solar cells: A data-driven approach for material innovation," *ACS Applied Energy Materials*, vol. 7, no. 20, pp. 9349–9363, 2024.

[17] R. Khatua, B. Das, and A. Mondal, "Physics-informed machine learning with data-driven equations for predicting organic solar cell performance, *ACS Applied Materials & Interfaces*, vol. 16, no. 42, pp. 57467–57480, 2024.

[18] S. A. Lopez, B. Sanchez-Lengeling, J. de G. Soares, and A. Aspuru-Guzik, "Design principles and top non-fullerene acceptor candidates for organic photovoltaics," *Joule*, vol. 1, no. 4, pp. 857–870, 2017.

[19] S. Zhong, and Guan X. "Count-based morgan fingerprint: A more efficient and interpretable molecular representation in developing machine learning-based predictive regression models for water contaminants' activities and properties," *Environmental Science & Technology*, vol. 57, no. 46, pp. 18193–18202, 2023.

[20] M. C. Hutter, "Differential multimolecule fingerprint for similarity search—making use of active and inactive compound sets in virtual screening," *Journal of Chemical Information and Modeling*, vol. 62, no. 11, pp. 2726–2736, 2022.

[21] U. V. Ucak, I. Ashyrmamatov, and J. Lee, "Reconstruction of lossless molecular representations from fingerprints," *Journal of Cheminformatics*, vol. 15, p. 26, 2023.

[22] D. Bajusz, A. Rácz, and K. Héberger, "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?," *Journal of Cheminformatics*, vol. 7, p. 20, 2015.

[23] F. Neese, "The ORCA program system," *WIREs Computational Molecular Science*, vol. 2, no. 1, pp. 73–78, 2012.

[24] F. Weigend, and R. Ahlrichs, "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy," *Physical Chemistry Chemical Physics*, vol. 7, no. 18, pp. 3297–3305, 2005.

[25] C. Lee, W. Yang, R. G. Parr, "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density," *Physical Review B*, vol. 37, pp. 785–789, 1988.

[26] A. D. Becke, "Density-functional thermochemistry. III. The role of exact exchange." *The Journal of Chemical Physics*, vol. 98, pp. 5648–5652, 1993.